CONSTRUCTING A POLYSEMOUS ACADEMIC VOCABULARY EXTENT TEST VIA POLYTOMOUS RASCH MODEL MEASUREMENT ANALYSES

A Dissertation
Submitted
to the Temple University Graduate Board

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

By
Phillip B. Rowles
December, 2015

Examining Committee Members

David Beglar, Advisory Chair, Teaching and Learning
Paul Nation, External Member, University of Victoria
Steven Ross, External Member, University of Maryland
James Elwood, External Member, Meiji University
Tomoko Nemoto, Teaching and Learning

ABSTRACT

Educational measurement research faces an unresolved dilemma: competently meeting the longstanding demand for improved vocabulary strength (depth) aspect assessments. My original contribution to knowledge in the written receptive vocabulary knowledge construct research domain is twofold. My first contribution is proposing an a priori metasynonymy awareness hypothesis based on a vocabulary strength aspect extension of O'Connor's (1940) written receptive vocabulary acquisition developmental stage theory. My second contribution is designing and constructing a vocabulary extent (the nexus between vocabulary size (breadth) and strength aspects) test. The test, called the Polysemous Academic Vocabulary Extent Test, utilizes ordered triple rank (OTR) responses and a complementary six-tier incremental scoring guide rubric. An example test item includes a sentence stem with a bold keyword and three options, such as:

All the reviews of the movie were **positive**.

positive: a) sure b) good c) environmental

Test-takers are instructed to rank all three options in synonymy order (null, decontextualized and synonym) from left to right: c) environmental (null), a) sure (decontextualized, that is, possible in another imagined sentence) and b) good (synonym). The test was administered to a sample consisting of Japanese female and male tertiary institution English learners (N = 379) from first-year undergraduate to graduate levels who were instructed to arrange ordered triple rank (OTR) responses by sequencing all three multiple-choice options for each test item according to an ordered synonymy rule. The test-taker responses were empirically evaluated by applying polytomous Rasch model analyses. After a posteriori category rescoring, all three

adjacent Rasch-Andrich thresholds indicated empirical order. In conclusion, there was preliminary empirical evidence supporting both the metasynonymy awareness hypothesis and the Polysemous Academic Vocabulary Extent Test's psychometric soundness.

ACKNOWLEDGMENTS

I would like to thank my dissertation committee of Dr. David Beglar, Prof. Paul Nation, Dr. Steven Ross, Dr. James Elwood and Dr. Tomoko Nemoto for their feedback, suggestions and time to assist me during the dissertation writing process.

My main advisor, Dr. Beglar, motivated me to realize the potential research connections between vocabulary acquisition and testing. His extensive knowledge of both these important domain areas was invaluable and inspirational. Prof. Nation helped me to clarify my conceptualization of the vocabulary theories used in this study. Dr. Ross's contributions, including his wide expertise in measurement and statistical methodologies, were very crucial. Dr. Elwood's sharp eye for detail about measurement and statistics was very much appreciated. Dr. Nemoto's insights into English to Japanese loanwords and statistics were very useful and precise.

The cohort mates, family members, friends, colleagues and faculty members who helped make this study possible were (in surname alphabetical order) Paul, Joff, Paul, Matt, Paul, Peter, Gary, David, Hiroshi, Yoshinobu, Ruby, Joe, Amy, Richard, Doug, Michael, Marvin, Kana, Robert, Paul and Jim. I also would like to thank the hundreds of anonymous Japanese tertiary education level students who participated in this study. Dr. Kenneth Schaefer, previous Graduate TESOL director, was a great source of inspiration and encouragement. To all those persons involved in this study, thank you very much for your participation, patience and time. The study would not have been completed without your generous contributions.

DEDICATION

This is dedicated to the lives and many loving memories of my parents, Neville "Hangman" and Norma "Rose" Rowles, who both passed away during the writing process. I definitely would not be here today without the both of you. Thanks so much for the precious memories and see you when the boat comes in.

TABLE OF CONTENTS

PAGE

ABSTRACT iii

ACKNOWLEDGMENTS v

DEDICATION vi

LIST OF TABLES xi

LIST OF FIGURES xv

CHAPTER

1. INTRODUCTION 1

The Background of the Issue 1

Statement of the Problem 5

Purposes and Significances of the Study 14

Theoretical Perspective 17

The Audience for the Study 19

Delimitations 20

The Organization of the Study 22

2. REVIEW OF THE LITERATURE 23

Vocabulary Testing 23

Dichotomous Vocabulary Size 24

Vocabulary Levels Test 25

Vocabulary Size Test 37

Polytomous Strength 38

Rasch Measurement Theory 39

Multiple-Choice Testing 41
Gaps in the Literature 45
Purposes of the Study 47
Research Questions 48
3. METHODS 50
Participants 50
Instrumentation 51
Procedures 54
Analyses 55
4. ITEM SPECIFICATION AND CONSTRUCTION PROCESS 72
Conceptual Framework Rationale 84
Construct Maps 86
Items Design/Outcome Space 95
Measurement Model 175
Quality Control of Baseline Data Fit to Model 180
Four Building Blocks Summary 185
Quality Control of Collapsed Data Fit to Model 188
Quality Control of Additional Data Fit to Model 189
5. RESULTS 191
Research Question 1: Rasch Analysis Requirements 191
Research Question 2: Hypothesized Difficulty of Loanwords and Non-loanwords 201
Research Question 3: Palmer's (1917) Word Frequency Acquisition Theory 206
Research Question 4: Metasynonymy Awareness Hypothesis 212

Research Question 5: Vocabulary Knowledge Estimates and the Type of Analysis....215

Research Question 6: Guessing and Carelessness....225

6. DISCUSSION....237

Research Question 1: Rasch Analysis Requirements....237

Research Question 2: The Difficulty of Loanwords and Non-loanwords....238

Research Question 3: Palmer's (1917) Word Frequency Acquisition Theory....245

Research Question 4: Metasynonymy Awareness Hypothesis....247

Research Question 5: Vocabulary Knowledge Estimates and the Type of Analysis....257

Research Question 6: Guessing and Carelessness....262

Implications for Vocabulary Assessment....265

7. CONCLUSION....267

Summary of the Findings....267

Limitations....271

Suggestions for Future Research....271

Final Conclusions....274

REFERENCES....277

APPENDICES

A) KEYWORDS BY JACET 8000 WORD FREQUENCY BAND AND KATAKANA DICTIONARY PAGE REPRESENTATION....288

B) ANSWERS TO ITEM OPTIONS....289

C) POLYSEMOUS ACADEMIC VOCABULARY EXTENT TEST (PAVET)....290

D) OPTICAL MARK RECOGNITION (OMR) ANSWER SHEET....294

E) LOANWORD RECOGNITION YES/NO CHECKLIST SURVEY AND ANSWER SHEET.....296

F) WRITTEN PERMISSION REQUESTED VIA EMAIL FOR ADAPTION OF A COPYRIGHTED TABLE INTO TABLE 2 (ON PAGE 40) 298

G) WRITTEN PERMISSION GRANTED VIA EMAIL FOR ADAPTION OF A COPYRIGHTED TABLE INTO TABLE 2 (ON PAGE 40) 299

LIST OF TABLES

Table Page

1. Scoring Approaches of Vocabulary Tests Influencing this Study 24
2. Three Main Statistical and Measurement Theories 40
3. Typical Multiple-Choice Item Single Best Answer Response Scoring Rubric.....46
4. Atypical Multiple-Choice Item Ordered Triple Rank Response Scoring Rubric 46
5. Item Specification and Construction Process Conceptual Framework 72
6. Reduced Item Specification and Construction Process Conceptual Framework 73
7. Construct Map for the Matter Strand of the Living by Chemistry Project.....77
8. A Priori Inter-Item Difficulty Parameter Vertical Construct Map 86
9. A Priori Intra-Item Difficulty Parameter Horizontal Construct Map.....94
10. Academic Vocabulary Extent (Size and Strength) Construct Framework.....94
11. Items Design/Outcome Space Building Blocks Protocol.....96
12. Test Blueprint for the 36 Bolded Keywords in the Sentence Stems 97
13. Rank-Ordered Options Permutations 111
14. A Priori Six Permutations as a Correct Answer Key 112
15. Correct Answer Key for 36 Items 113
16. Cartesian Product Rule Rubric 115
17. Underlying Mathematical Relations Via the Cartesian Product Rule Lists Resulting in Six Permissible Responses.....115
18. Six Permutations of Permissible Correct Answer Patterns Based on Cartesian Product Rule Subset Patterns.....116
19. Reverse Engineering the Permutational Word Sense Stages of the Metasynonymy Awareness Hypothesis 125

20. Metasynonymy Awareness Hypothesis Represented as Permutational Word Sense Stages with Scoring Guide 126

21. A Priori Six Category Responses-Focused Hierarchical Scoring Guide 129

22. A Priori Six Category Scores-Focused Hierarchical Scoring Guide....130

23. A Priori Six Sextuples-Focused Hierarchical Scoring Guide 131

24. Ordered Triple Rank Responses Scoring Guide for Item 5 with Correct Answer (c, b, a)....132

25. Proxy Single Best Answer Responses Scoring Guide for Item 5....132

26. Ordered Triple Rank Responses Scoring Guide for Item 1 with Correct Answer (b, c, a)....133

27. Proxy Single Best Answer Responses Scoring Guide for Item 1....133

28. Ordered Triple Rank Responses Scoring Guide for Item 3 with Correct Answer (c, a, b)....134

29. Proxy Single Best Answer Responses Scoring Guide for Item 3....135

30. Ordered Triple Rank Responses Scoring Guide for Item 6 with Correct Answer (a, c, b)....135

31. Proxy Single Best Answer Responses Scoring Guide for Item 6....136

32. Ordered Triple Rank Responses Scoring Guide for Item 4 with Correct Answer (b, a, c)....136

33. Proxy Single Best Answer Responses Scoring Guide for Item 4....137

34. Ordered Triple Rank Responses Scoring Guide for Item 2 with Correct Answer (a, b, c)....138

35. Proxy Single Best Answer Responses Scoring Guide for Item 2....138

36. Ordered Triple Rank (OTR) Responses from First Six Persons for Items 1, 2, 16, 17, 35, and 36....171

37. Ordered Triple Rank (OTR) Responses from First Six Persons via Lossless Data Compression Senary Scoring 171

38. Ordered Triple Rank (OTR) Responses from First Six Persons via Lossless Data Compression Binary Scoring....172

39. Single Best Answer (SBA) Proxy Response from First Six Persons via Lossy Data Compression Ternary Scoring 172

40. Single Best Answer (SBA) Proxy Response from First Six Persons via Lossy Data Compression Binary Scoring 173

41. Ordered Triple Rank (OTR) Responses and Respective Data Compression Scores from First Six Persons for Items 1, 2, 16, 17, 35, and 36 174

42. Pilot Study 1 Summary of Category Structure 176

43. Pilot Study 2 Summary of Category Structure 177

44. Pilot Study 3 Summary of Category Structure 178

45. Initial Main Study Summary of Category Structure 183

46. Secondary Main Study Summary of Category Structure 184

47. Pilot and Main Studies Summary 185

48. Construct Modeling Approach to the Item Specification and Construction Process 186

49. Four Types of Analysis 186

50. Four Types of Analysis in a Matrix Design 187

51. Person and Item Separation Statistics for the Four Types of Analysis 192

52. Item Difficulty Calibration Hierarchy 197

53. Summary of Category Structure 198

54. Standardized Item Residual Variance 200

55. Item Difficulty Calibration Hierarchy by Loanword Status 204

56. Item Difficulty Calibration Hierarchy by Word Frequency Band 209

57. Summary of Category Structure for RSM 011223 Scale Ordered Triple Rank 213

58. Description of Four Rasch Model Scale Analyses 215

59. Lower Asymptote Index Initial Approximations Indicating Loanword Keywords Eliciting Item Guessability 226

60. Lower Asymptote Index Initial Approximations Indicating Non-Loanword Keywords Eliciting Item Guessability 227

61. Lower Asymptote Index Initial Approximations Indicating Loanword and Non-Loanword Keywords Eliciting Item Guessability 228

62. Upper Asymptote Index Initial Approximations Indicating Loanword Keywords Eliciting Item Carelessness 231

63. Upper Asymptote Index Initial Approximations Indicating Non-Loanword Keywords Eliciting Item Carelessness 232

64. Upper Asymptote Index Initial Approximations Indicating Loanword and Non-Loanword Keywords Eliciting Item Carelessness 233

65. Loanword Status Recognition Yes/No Checklist Survey by Item Difficulty Calibration Hierarchy 244

66. Metasynonymy Awareness Hypothesis: Comparing A Priori Prediction and A Posteriori Realization Scoring Guide 248

67. Success Probabilities at the Lowest, Mean and Highest Difficulty Items for the Lowest, Mean and Highest Ability Persons at Thresholds 1, 2, and 3 253

68. Success Probabilities of the Lowest, Mean and Highest Ability Persons for the Easiest, Mean and Most Difficult Items at Thresholds 1, 2, and 3 254

69. Tentative Estimates of Vocabulary Extent 255

70. Guessing Extent 263

71. Carelessness Extent 264

LIST OF FIGURES

Figure Page

1. Wright map example 80
2. Category characteristic curve map example....85
3. Three listed construct options with single best answer (SBA) instruction and responses 101
4. Three listed construct options with ordered triple rank (OTR) instruction and responses 104
5. Ordered 3-element arrangements of the list (n, d, s) resulting in sextuples 105
6. Decision tree for ordered triple responses to three options 107
7. Permutations decision process without repetition. 110
8. Optical mark recognition 3-by-3 matrix item response design example....117
9. Outcome space design for typical single best answer (SBA) item response....118
10. Outcome space design for atypical ordered triple rank (OTR) item responses....119
11. Typical single best answer (SBA) item response: "c" 119
12. Atypical ordered triple rank (OTR) item responses: Bifurcation of SBA "c" 120
13. Typical single best answer (SBA) item response: "b" 120
14. Atypical ordered triple rank (OTR) item responses: Bifurcation of SBA "b" 120
15. Typical single best answer (SBA) item response: "a" 121
16. Atypical ordered triple rank (OTR) item responses: Bifurcation of SBA "a" 121
17. Ordered triple rank (OTR) response decision process without repetition....122
18. Proxy single best answer (SBA) response....122
19. Ordered triple rank (OTR) response decision process without repetition....123
20. Proxy single best answer (SBA) response....123
21. Ordered triple rank (OTR) response decision process without repetition....123

22. Proxy single best answer (SBA) response 124

23. (a) A bijective function (bijection) between the 3 elements of the 2 sets; (b) and (c) are non-bijections 139

24. (a) A permissible ordered triple rank (OTR) response; (b) and (c) are impermissible ordered triple rank (OTR) responses 141

25. Ordered triple rank (OTR) responses for item 2 where (a, b, c) is correct 142

26. Ordered triple rank (OTR) responses for item 4 where (b, a, c) is correct 142

27. Ordered triple rank (OTR) responses for item 6 where (a, c, b) is correct 142

28. Ordered triple rank (OTR) responses for item 3 where (c, a, b) is correct 143

29. Ordered triple rank (OTR) responses for item 1 where (b, c, a) is correct 143

30. Ordered triple rank (OTR) responses for item 5 where (c, b, a) is correct 144

31. Scoring guide tree for item 5 with correct answer (c, b, a) 145

32. Scoring guide tree for item 1 with correct answer (b, c, a) 146

33. Scoring guide tree for item 3 with correct answer (c, a, b) 147

34. Scoring guide tree for item 6 with correct answer (a, c, b) 148

35. Scoring guide tree for item 4 with correct answer (b, a, c) 150

36. Scoring guide tree for item 2 with correct answer (a, b, c) 151

37. (a) Function composition representing lossless and (b) lossy data compression for item 5 with correct answer (c, b, a) 153

38. (a) Function composition representing lossless and (b) lossy data compression for item 1 with correct answer (b, c, a) 154

39. (a) Function composition representing lossless and (b) lossy data compression for item 3 with correct answer (c, a, b) 155

40. (a) Function composition representing lossless and b) lossy data compression for item 6 with correct answer (a, c, b) 156

41. (a) Function composition representing lossless and (b) lossy data compression for item 4 with correct answer (b, a, c) 157

42. (a) Function composition representing lossless and (b) lossy data compression for item 2 with correct answer (a, b, c) 158

43. Three function composition representing ordered triple rank (OTR) responses for item 5 with correct answer (c, b, a) 159

44. Three function composition representing proxy single best answer (SBA) responses for item 5 with correct answer (c, b, a) 160

45. Three function composition representing ordered triple rank (OTR) responses for item 1 with correct answer (b, c, a) 161

46. Three function composition representing proxy single best answer (SBA) responses for item 1 with correct answer (b, c, a) 162

47. Three function composition representing ordered triple rank (OTR) responses for item 3 with correct answer (c, a, b) 163

48. Three function composition representing proxy Single Best Answer (SBA) responses for item 3 with correct answer (c, a, b) 164

49. Three function composition representing ordered triple rank (OTR) responses for item 6 with correct answer (a, c, b) 165

50. Three function composition representing proxy single best answer (SBA) responses for item 6 with correct answer (a, c, b) 166

51. Three function composition representing ordered triple rank (OTR) responses for item 4 with correct answer (b, a, c) 167

52. Three function composition representing proxy single best answer (SBA) responses for item 4 with correct answer (b, a, c) 168

53. Three function composition representing ordered triple rank (OTR) responses for item 2 with correct answer (a, b, c) 169

54. Three function composition representing proxy single best answer (SBA) responses for item 2 with correct answer (a, b, c) 170

55. Categorization of available measurement information 188

56. Wright map for RSM 011223 baseline scale.....193

57. Polysemous Wright map for baseline quaternary RSM 011223 baseline scale 195

58. Three Rasch-Andrich thresholds of RSM 011223 scale by category score 199

59. Wright map according to loanword status (loanwords and non-loanwords).....203

60. Boxplot of keyword loanword status means according to item difficulty calibrations 205

61. Rank-ordered option response 4-category RSM 011223 Wright map according to keyword high, medium and low word frequency band levels of the JACET 8000 academic word list.....207

62. Boxplot of keyword frequency band means according to item difficulty calibrations 210

63. Category probability curves for 3 categories of the RSM 011223 scale.....214

64. Difference between scoring RSM 011223 and DRM 000001 in person ability measures 219

65. Difference between scoring RSM 011223 and RSM 001122 in person ability measures 221

66. Difference between scoring RSM 011223 and DRM 000011 in person ability measures 222

67. Item stem keyword difficulty calibrations classified according to loanword status and word frequency band 240

68. Andrich thresholds in relation to person ability and item difficulty 251

69. Lexical space extent displaying Rasch-Andrich thresholds in relation to person vocabulary size and strength at item's third threshold 256

CHAPTER 1

INTRODUCTION

The Background of the Issue

Synonyms have been of interest to academics since at least the Classical Greek era, as shown in Aristotle's (1924/2004) *Rhetoric* written in the fourth century B.C.. In the two millennia since then, testing of keyword synonyms and definitions in second or foreign language contexts has become prevalent. These tests typically attempt to measure and assess receptive vocabulary ability.

Most receptive vocabulary tests aim to evaluate written receptive vocabulary knowledge, in other words, vocabulary knowledge accessed through the reading comprehension macro skill. Assessing this type of knowledge gained through reading is especially important in foreign language contexts, such as learning English in Japan, where access to English native speakers is limited. Furthermore, modern technology developments have introduced the Internet, email and other social media resources, which rely heavily on English written words for intercultural communication.

Receptive vocabulary tests are used for many purposes in educational curricula. They can be used for placement, diagnostic and achievement purposes within educational contexts. Placement purposes include streaming various levels of learners into particular classes. Diagnostic purposes include determining individual learner strengths and weaknesses in order to identify knowledge gaps that need focus. Achievement purposes

include the measuring of learner accomplishments, most likely at the end of some instructional period.

The curriculum influencing the teaching of English vocabulary in Japan has been evolving through some changes recently. The Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT, 2009, 2010) has implemented curriculum changes over the last couple of years for junior and senior high schools in regards to English vocabulary learning. Therefore, the combined number of English vocabulary taught in junior and senior high schools has increased by 800 from 2,200 to 3,000 words. These curriculum changes came fully into place in April, 2013.

Vocabulary extent in this study is defined as a fourfold concept, consolidating two skills and two modalities. In this study, the two vocabulary skills include receptive skills in the form of reading typical closed-response multiple-choice item stems and trinary (three alternative) options. Unique to this study, productive-like skills are responses within atypical incrementally (sequentially ordered by correctness) ordered triple rank response outcome spaces. In contrast to the usual multiple-choice item requirement of *choosing an* option, respondents are required to *rank all* options. Therefore, instead of utilizing a traditional multiple-choice (MC) test, this study outlines the proposal of developing an incremental multiple-rank (MR) test. MC formats typically require respondents to choose only a unary (single) response in the correct outcome space for each item. In contrast, the MR response format in this study requires respondents to rank all trinary (triple) responses in an incrementally correct sequence. An example item, including the sentence stem and three options follows:

The **volume** of food almost covered the entire plate.

volume: a) loudness b) vent c) amount

The test-taker is instructed to rank all three options according to a synonymy rule ordered left to right of null, decontextualized (possible in another imagined sentence) and synonym. The rank order synonymy rule is from left (incorrect) through middle (partially correct) to right (correct), that is, in terms of these item options, b) vent, a) loudness and c) amount. The term, partially correct response, is defined here as referring to an intermediate response influenced by incomplete semantic knowledge that is to a limited extent and could be visualized on an inclined slope between an incorrect and a correct response.

The scoring rubric of polytomous trinary MRs requires bifurcation of each related unary MC scoring category. Thus, for example in a trinary partial knowledge item case, instead of having only three MC scoring categories per item, a parallel MR would have six scoring categories per item.

The two vocabulary modalities include first, the size (i.e., breadth) modality represented visually by inter-item difficulty height differences on a Wright Map. Theoretically, size modality is defined by loanword status macro-frequency as an amendment of Palmer's (1917) vocabulary frequency theory. English loanwords borrowed into the Japanese language are defined in this study as part of *gairaigo*, that is, words loaned from mainly European languages (Schmidt, 2009). Second, the strength (i.e., depth) modality is visually represented by advancing Andrich (1978) thresholds, which is represented by ordered inter-threshold (intra-, or within item) width differences in a category characteristic curve figure. Theoretically, strength modality is defined by semantic

micro-frequency as an amendment of O'Connor's (1940) first language (L1) vocabulary acquisition stage theory.

Measuring partial knowledge of receptive vocabulary is an area of research that has not been fully explored. Many receptive vocabulary tests focus on the purpose of measuring vocabulary size (breadth) and utilize the test design of multiple-choice items with typically four options, consisting of a correct definition or synonym answer and three incorrect distractors. As a single best answer response mandate is usually applied, test- takers can only indicate a full-knowledge response. In contrast, partial, or incremental knowledge is typically not part of the test purpose or design.

Single best answer responses of full knowledge have matched the test purpose and design of assessing vocabulary size (breadth). The objective of this present study is to develop a way to build on this previous knowledge, that is, to develop a receptive vocabulary extent (size/breadth and strength/depth) test. Therefore, this additionally requires the measurement of incremental vocabulary knowledge to assess learner's vocabulary strength (depth).

In this study I attempt a unification of the two most remarkable human achievements. "After language, our greatest invention is numbers. Numbers make measures and maps and so enable us to figure out where we are, what we have, and how much it's worth" (Wright, 1997, p. 33). L. L. Thurstone apparently thought similarly in 1924 when he developed the *Psychological Examination for High School Graduates and College Freshmen*. Thurstone's test had two scores, one linguistic, L, and another quantitative, Q (Jones & Thissen, 2007). The purpose of this study is to consolidate these two eminent

human innovations, language and numbers, by measuring foreign language vocabulary knowledge. Wright (1997) further asserted, “The purpose of measurement is inference” (p. 43). In order to meet these ends, this study documents the construction of a new multiple-choice test response format and scoring guide rubric that measures the incremental knowledge of academic semantic lexical extent and the communication of meaningful inferences to the intended audience.

Statement of the Problem

The development of new word meanings is a partial acquisition process. Language students might have no, partial or full understanding of particular word meanings. Over 70 years ago, Cronbach (1942) in a review of diagnostic vocabulary tests, emphasized the need for measuring vocabulary knowledge strength by assessing degrees of meaning. This gradient of understanding might start with some knowledge and culminate in total fullness of meaning knowledge. Cronbach points out that often researchers restrict their studies to a single meaning of individual words. This is in contrast to a meaning richness ideal whereby multiple meanings of individual words could potentially be tested. Instead of using the all or none, that is, right or wrong binary categorization, vocabulary tests would ideally use degrees of meaning richness. Cronbach (1942) called for diagnostic vocabulary test development, in which both meaning breadth (use in a particular context) and more specifically, meaning precision—vocabulary exactness distinguishing between appropriate and inappropriate use in different circumstances—are essential. A year later, Cronbach (1943) described the shortcomings of vocabulary strength aspect diagnostic tests to

measure adequately. Previous tests had been designed to distinguish words that were merely unknown or known. Therefore, the typical vocabulary tests were not measuring on a partial scale of degrees, that is, precise word meaning understanding. Cronbach (1943) called for future superior analyses whereby partial degrees can be measured instead of a simplistic unknown or known knowledge label for a single meaning of a word. More specifically, in reference to test-takers, Cronbach (1943) asserted that "it is important to determine how precise a concept he has acquired" (Cronbach, 1943, p. 528).

Cronbach was not alone in the 1940's era with his call to arms for improved tests of vocabulary strength. Other researchers, including Seashore and Eckerson (1940) heralded the need for future research into meaning richness, that is, depth of knowledge into alternative multiple meanings beyond the solitary most common meaning. Another researcher of that time was Lovell (1941) who was also an advocate of vocabulary richness, that is, multiple meanings knowledge. Another pair of researchers pushing for vocabulary strength was Dolch and Leeds (1953) who critiqued contemporary vocabulary tests, which "were not trying to measure scientifically a certain degree of meaning" (p. 188). Further, the vocabulary tests of those days "ignore all but the most common meaning of a word" (Dolch & Leeds, 1953, p. 188).

Moving beyond a single meaning for future vocabulary testing is important because, "words are the symbols with which we communicate with one another, it is important to know just what meaning or meanings words have for different people" (Dolch & Leeds, 1953, p. 188). Dolch and Leeds (1953) asserted in the future, vocabulary testing research would contemplate some fundamental issues including:

> the first consideration will be that a word does not have a "meaning". That naïve assumption is all too prevalent even among educators. Instead, a word frequently has many meanings, and each of these meanings may be known in varying degrees by different people, or by the same person in the course of his life. (p. 188)

Individual persons are in the process of vocabulary development, therefore, assessments should not just indicate that a person "'knows something about' a certain word, but also how much he knows about it. Only then do we know what more he should know and how we can help him develop that added meaning" (Dolch & Leeds, 1953, p. 189).

However, despite this longstanding advice, even in more modern times no empirical test research adequately supports incremental semantic lexical stage development acquisition (Bruton, 2007). This occurs because assessing incremental knowledge is complex and testing methods lack consistency. "Any assessment of vocabulary development from reading is complicated by the fact that vocabulary knowledge is subject to partial increments" (Bruton, 2007, p. 174). Furthermore, "there are numerous weaknesses and a lack of consistency in the instruments used to assess vocabulary acquisition across a significantly wide range of empirical research studies, and consequently a lack of consistent quantification" (Bruton, 2007, p. 175).

Most tests of vocabulary full-knowledge use the multiple-choice format with a single best answer instruction. Typically, they are used because they are straightforward to administer and score. However, there are problems associated with using the traditional multiple-choice format when assessing incremental knowledge. Bruton (2007) stated that it is obvious that conventional multiple-choice test responses "cannot cope with partial lexical

knowledge" (p. 166), the rationale being "the fundamental problem with M/C formats, when only *one* option can be chosen by the subjects, is that it is impossible to include both partial and correct options without the choice between them interfering with each other" (Bruton, 2007, p. 169).

Complicating the matter further, there is research by Ruhl (1989) on monosemy in linguistic semantics, which proposes a monosemic bias hypothesis whereby a word is initially assumed to have a solitary, highly abstract meaning. Any influences that might make it seem polysemic are due to contextual factors only. This approach proposed by meaning-minimalists such as Ruhl (1989) assumes minimal word meanings and unambiguous words. This is in contrast to the approach proposed by meaning-maximalists, which assumes richness of word meanings and ambiguous words. Despite Ruhl's (1989) meaning-minimalist monosemic approach the view proposed in this present study is for a meaning-maximalist polysemic approach to word meanings.

Perhaps part of the problem is the changing definition of vocabulary strength over time. As previously described in this section, in the 1940 and 1950's, researchers such as Cronbach, (1942, 1943), Seashore and Eckerson (1940), Lovell (1941), and Dolch and Leeds (1953), perceived vocabulary strength as dealing with vocabulary richness, that is, multiple meanings of a word. In more recent times, Read (2004) outlined the vocabulary strength aspect as developing along three main lines of inquiry. The first line was precision of meaning, which is how the concept of vocabulary richness of multiple meanings described previously was perceived back in the 1940 and 1950's. The next two lines of inquiry are new; they are comprehensive word knowledge and network knowledge. The

first line of inquiry, precision of meaning, is the focus of this study. The vocabulary strength aspect as precision of meaning was also a feature of Anderson and Freebody's (1981) modern definition of vocabulary depth focused on word meanings (Gyllstad, 2013). Meaning precision "is consistent with how Anderson and Freebody (1981) originally defined depth of word knowledge" (Gyllstad, 2013, p. 18). One critique of the modern conception of vocabulary strength is that it has become too broad and out of focus. In contrast to vocabulary size, vocabulary strength is very wide ranging and has lost a lot of the research focus that it had back in the 1940's and 1950's. Although the wide perspective has brought a lot of valuable ideas into the vocabulary strength research domain, there is a trade off in that the research diversification has led to a lack of cohesion. Therefore, in this study the focus is on the original intent of the vocabulary strength research domain; in other words, precision of meaning, by testing multiple word meanings.

Research Problems

The research problems addressed by this study center on the following five distinct key issues: test designers' persistence in perpetuating only traditional test formats, word frequency acquisition theory, loanword acquisition theory, synonymy acquisition theory, and test-taker guessing/carelessness. These key issues are expanded upon in the following paragraphs.

Measuring partial vocabulary knowledge is important. However, because of the perceived impossibility of using traditional multiple-choice test response formats to measure incremental knowledge, vocabulary test designers typically do not attempt to give

credit for incremental knowledge. In the second and foreign language assessment field, there has been research usually focused on testing receptive vocabulary knowledge. The great majority of vocabulary research into this field has been limited to the work of Meara (Meara & Buxton, 1987), particularly in the area of the yes-no test and Nation, particularly in the area of levels (Nation, 1987, 1990) and size (Nation & Beglar, 2007; Nation, 2012) tests. Both of these research areas have been restricted to testing one meaning per item in a single best answer to a multiple-choice (1 from 2 options, 3 from 6 options and 1 from 4 options) design format. This study addresses these gaps in the research literature.

As a result of the research separation of vocabulary researchers from second language acquisition and language testing researchers particularly in the 1980's decade, Read (2013) stated that at "the conceptual level, the result of this isolation from wider influences was a relative lack of theory to account for the nature of L2 vocabulary knowledge and the processes of lexical development in L2" (p. 41). Therefore, to address this lack of theory several theories are referred to in this study.

In his vocabulary frequency acquisition theory, Palmer (1917) stated that higher frequency words would be encountered more often, be easier to learn and therefore would be acquired first by beginner learners. Palmer also proposed that lower frequency words are encountered less often, are more difficult to learn and therefore are acquired later by veteran learners. Palmer's (1917) theory is empirically tested in this study.

An extension of Palmer's (1917) vocabulary frequency acquisition theory is proposed in this study. Since the end of World War II, the Japanese language has been inundated with loanwords borrowed from English in particular (Tomoda, 2009). For

Japanese learners of English, these English loanwords could be of higher frequency, more likely to be encountered, easier to learn and therefore could be acquired first by beginner learners. No second language researchers have investigated the issue of loanwords in regard to an extension of Palmer's (1917) theory. The ease of acquiring English loan words in the Japanese English as a Foreign Language (JEFL) context is empirically tested in this study.

Loanwords from English are assumed to be easy in this study because they are relatively frequent in the Japanese language. They might also be easy because they are similar phonologically. The meaning of loanwords is often similar in English and Japanese.

However, there are exceptions. Sometimes, the meaning of loanwords can change when they move from language to language, for example in this study's case, from English to Japanese. Two examples of loanwords borrowed from English and used in Japanese with shifted meanings are the words *service* and *claim*. In the Japanese context, the meaning of the word service refers to a small free object given to a customer from a business staff member as a sign of thanks. For example, at the end of eating lunch at a restaurant, the owner gives a free dessert to the customer as service. Additionally in the Japanese context, the meaning of the word claim refers to a complaint. For example, if after checking into a hotel room the guest discovered that the room had not been previously cleaned that day, then the customer would make a claim to the management.

In his synonymy acquisition theory, O'Connor (1940) proposed that English first language learners develop through five stages—sounds-like, looks-like, context, antonym and synonym—in their progression toward acquiring an item's keyword synonym. Smith (1987) investigated O'Connor's theory with five categories in an English native speaker

context with disordered and reversed thresholds as a result. The critical problem of O'Connor's (1940) vocabulary acquisition theory appears to be that five stages are too many for meaningful measurement. One solution to this problem with O'Connor's theory is an amendment where less than five of the proposed major levels could be empirically tested. However, O'Connor's theory has never been investigated in a second or foreign language context prior to pre-pilot studies for this main study. Pre-pilot results for this study indicated that an amended four categories resulted in disordered thresholds. An amendment to O'Connor's (1940) stage theory using only three stages in the form of a metasynonymy awareness hypothesis is thus empirically tested in this study.

Guessing, which is a correct response to a difficult item above a person's ability level, and carelessness, an incorrect response to an easy item below a person's ability level, in regards to multiple-choice items, have not been investigated frequently in the vocabulary research literature. It is hypothesized that the test design in this study reduces guessing and carelessness. The extent of guessing and carelessness in responses is empirically tested in this study.

In the fields of second and foreign language assessment, there has been a focus on testing written receptive vocabulary size. As previously noted, the great majority of research into this area has been limited to the work of Meara and of Nation. This study focusing on vocabulary strength addresses this gap in the research literature.

Written receptive vocabulary assessment has been limited somewhat to the majority of instruments attempting to measure vocabulary size for diagnostic (Vocabulary Levels Test) or proficiency (Vocabulary Size Test) purposes. These test formats typically use

traditional multiple-choice items with single best answer response mandates and dichotomous scoring procedures. Therefore, a gap exists in the research literature, which challenges taking written receptive vocabulary testing in a new direction that has not been previously investigated. This study addresses this gap in the research literature.

Furthermore, despite the recent efforts of the Japanese Ministry of Education, Culture, Sports, Science and Technology's curriculum changes (MEXT, 2009, 2010) there is no systematic method of selecting which academic words learners should study in Japanese secondary schools. The recent additional 800-word mandate of vocabulary to be studied depends on the particular textbook series assigned by individual junior and senior high schools. Therefore, depending on which secondary-level school students attended, tertiary-level learners have encountered different English word lists, which have been arbitrarily selected by textbook publishers. This arbitrary vocabulary demarcation plausibly leads to wide variation in English academic vocabulary acquisition.

On the other hand, there is an overall mandate by the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT, 2009, 2010) that Japanese students learn English as a lingua franca, study abroad, and participate in global society. Furthermore, foreign higher education students are being encouraged to study in Japan. If these are realistic goals, then the selection of vocabulary from an internationally-recognized English academic vocabulary list would be more beneficial than arbitrary selection by individual publishing companies. As a preliminary stage towards achieving these goals, this study proposes developing a test based on an internationally recognized academic word list.

Purposes and Significances of the Study

The purposes and significances of this study are presented in parallel organization with the previous section, Research Problems. To reiterate, the five distinct key issues addressed are: test designer's persistence in perpetuating only traditional test formats, word frequency acquisition theory, loan words acquisition theory, synonymy acquisition theory, and test-taker guessing/carelessness.

The first purpose is to address the problem of perceived impossibility of measuring incremental knowledge and test designer's persistence on perpetuating only traditional test formats. The majority of vocabulary test designers use multiple-choice item formats, which incorporate a single best answer response instruction and dichotomous scoring procedures. The significance of attempting to shift away from this established paradigm is that researchers might begin to explore alternative item formats and polytomous scoring procedures in order to measure incremental knowledge. A research anomaly can be the first indicator of a future paradigm shift (Kuhn, 1996 and Andrich, 2004a, 2004b).

The second purpose is to address the problem of Palmer's (1917) word frequency acquisition theory not being applied in incremental knowledge acquisition contexts. The significance of applying Palmer's (1917) theory is that it is applied in incremental knowledge research for the first time. As test-takers are required to rank all three multiple-choice options, an incremental vocabulary knowledge metasynonymy awareness scale based on the item's keywords word frequency can be measured. Further, the word frequency band levels are classified from a list, the JACET 8000 academic word list (Aizawa, Ishikawa & Murata, 2008), which was constructed specifically for the Japanese

English as a Foreign Language context. Using an academic word frequency list that was designed specifically for the Japanese English as a Foreign Language learning context assists in transferring Palmer's (1917) word frequency acquisition theory into the Japanese situation-specific setting.

The third purpose is to address the problem of the proposal of loanwords as an extension of Palmer's (1917) word frequency theory. That is, as English loanwords to the Japanese English as a Foreign Language context are more frequently encountered, loan words are easier to acquire. Increased exposure to English-Japanese loanwords is hypothesized to make them easier to acquire a form-meaning match when compared with non-loanwords. As this is the first time for the proposal of such an amendment, the significance is that it addresses a gap in the research literature. Previous researchers have proposed that English loanwords to the Japanese English as a Foreign Language context are easier to acquire (Daulton, 2008). However, this proposal has not been adequately tested at an empirical level.

The fourth purpose is to address the problem of O'Connor's (1940) synonymy acquisition theory, which has not been investigated by foreign language vocabulary researchers. O'Connor's (1940) five-stage synonymy acquisition theory was empirically tested by Smith (1987) in an English as a native language context; however, the findings indicated that the incremental vocabulary acquisition stage thresholds were disordered. The significance of this is that for the first time, foreign language vocabulary research utilizing an amendment to O'Connor's acquisition theory is empirically investigated. The amendment uses only three of O'Connor's (1940) stages and as a further amendment, there

is a bifurcation of those three stages to effectively construct a hypothesized six-stage incremental metasynonymy acquisition theory.

The fifth purpose is to address the problem of test-taker guessing and carelessness. Many studies have investigated vocabulary test-taker guessing and carelessness in English as native speaker contexts. However, there are relatively few studies investigating vocabulary test-taker guessing and carelessness in a Japanese English as a Foreign Language context. Thus, there is a gap in the research literature addressing guessing and carelessness. The significance of this is that the newly proposed test design and scoring guide procedures in this study attempt to reduce test-taker guessing and carelessness.

Furthermore, one of Fisher's (1935) main objectives was to extensively plan designs a priori so as to maximize the potential information that could be extracted from the data. In this study the use of an ordered triple rank (OTR) response test item design and an incremental scoring guide rubric procedure are proposed a priori in an attempt to communicate meaningful and valid inferences to the intended audience.

To summarize, this study adds to the scholarly literature in the field by developing a new multiple-choice test response format and scoring guide rubric. This new instrument facilitates polytomous measuring of vocabulary knowledge extent by being the nexus of two vocabulary aspects: size and strength. Typical vocabulary tests are concerned with only size or strength.

Theoretical Perspective

This study documents the development of a polytomously-scored Rasch model (Andrich, 1978) measurement vocabulary extent (size and strength) test. Therefore, as the test assesses extent, not only are substantive theories for the hierarchical order of item difficulty (size) outlined, but also a substantive theory for the hierarchical order of threshold difficulty (strength) is described. The vocabulary size aspect theories hypothesize about inter-item item difficulty variation, while the vocabulary strength aspect theory hypothesizes about intra-item, or more accurately, inter-threshold difficulty variation.

The first theory described is vocabulary frequency theory, which was developed by Palmer (1917) and used to study how word frequency influences difficulty. This theory indicates that there is a relative easiness of learning words that are frequently used. As applied to this study, an amendment to Palmer's (1917) theory would hold that the independent variable, loan word status, should influence the dependent variable, item difficulty, because English loan words are more frequently used than non-loanwords in Japan and are therefore easier to learn.

The second theory detailed is a vocabulary strength acquisition stage theory, which was developed by O'Connor (1940) and used to study vocabulary mistakes, or more accurately, acquisition development, made by adults using English as a first language (L1). This theory specifies five stages (from early to advanced, respectively) of learning a word's meaning: (a) sounds-like the synonym of the target word or the target word itself, (b) looks-like, (c) context words, (d) antonyms, and (e) synonyms. However, there are problems when applying O'Connor's (1940) five-stage theory to empirical testing as indicated by

Smith's (1987) disordered and reversed thresholds. The underlying problem with O'Connor's (1940) theory was that the five stages were too numerous to empirically test the constructed measures on a robust scale. Therefore, a possible solution is to reduce the five stages down to an amended four or three stages.

As applied to this study, an amendment to O'Connor's (1940) theory is that the independent variable, vocabulary acquisition stages, should influence the dependent variable, threshold difficulty. In other words, three vocabulary acquisition developmental stages should progress through words that sound or look like a target synonym (i.e., null), a form-meaning match in an imagined sentence (i.e., decontextualized) and another form-meaning match in the written item stem (i.e., synonym) in a Japanese English as a Foreign Language context. These three stages of O'Connor's (1940) original five stages are based on the results of previous pre-pilot studies. A further amendment is the ranking of these three stages as an ordered triple rank (OTR) response. Therefore, not only are the test-takers answering with a single best answer (SBA) response, but they are adding more response information in the form of metasynonymy awareness for each item. Effectively, the proposed ordered triple rank (OTR) responses bifurcate typical single best answer (SBA) responses in a three option multiple-choice test item design. In other words, the three stages were split into two each, resulting in six developmental stages of acquisition. For example, in a typical three option multiple-choice test item, *b* is the correct answer with traditional single best answer (SBA) response instructions. However, with the proposed ordered triple rank (OTR) response instructions to rank the options in a null, decontextualized and synonym sequence, *cab* is the correct answer. An ordered triple rank

(OTR) response from another test-taker of *acb* is one degree down in the correctness hierarchy. In contrast, an ordered triple rank (OTR) response of *bac* (the inverse of the *cab* correct response) from another test-taker is the incorrect answer. This new testing procedure is explained in finer detail in subsequent chapters.

The Audience for the Study

The audiences who will benefit from this study are doctoral students, vocabulary and measurement researchers, Japanese high school teachers, administrators and curriculum designers. The potential benefits for each group follow.

Doctoral students will be able to follow the documented process of designing and constructing a ranked response and polytomously-scored vocabulary test using Rasch model analyses. The conceptual process framework adopted for documenting this evidence is Wilson's (2005) construct modeling approach. This process begins with the initial stage of defining the construct of written receptive vocabulary extent by a visual map through to the final stage of making inferences from the person ability measures by another visual map.

Vocabulary researchers will be able to view vocabulary testing from a different perspective defined as vocabulary extent, that is, measuring using the typical vocabulary size and the atypical vocabulary strength consolidated into one test. This is in contrast to previous studies, for example, Ishii and Schmitt (2009), which incorporated a four test battery. Vocabulary extent is a fresh perspective as it amalgamates vocabulary size and strength into a new nexus viewpoint, which might open doors for future vocabulary research.

Measurement researchers will be able to follow the documentation of objective measurement theory analyses using the polytomous Rasch model applied to the field of written receptive vocabulary knowledge in a Japanese English as a Foreign Language (JEFL) context. The documentation of the evidence gathered in this study might be applicable to other fields of measurement research initially perceived to be unrelated.

Japanese high school teachers will be able to get an indication of what types of word synonyms (that is, contextualized or decontextualized, loanword or non-loanword, high, medium or low word frequency) are easy and difficult for learners who have matriculated into a Japanese tertiary institution. From that information, they might be able to design more effective high school English class syllabi to guide students for future English vocabulary development.

Japanese high school administrators and curriculum designers will hopefully be able to recognize the potential benefits of focusing high school classes on academic vocabulary from an internationally-recognized academic word list. As English vocabulary development has been one of the major areas of Japanese high school development under consideration in recent years, the salience of English synonymy research in Japan is very high.

Delimitations

The study is limited in scope in that the item keyword academic vocabulary was only sourced from the Academic Word List (Coxhead, 2000). Furthermore, there is a limitation that the item keyword word frequency lists were classified as high, medium and low according to the JACET 8000 academic word list (Aizawa, et al., 2008), whereby the

high frequency band was sourced from the first three word frequency levels, that is, 1,000, 2,000, and 3,000 levels, the medium frequency band was sourced from the next two word frequency levels, that is, the 4,000 and 5,000 levels and the low frequency band was sourced from the last three word frequency levels, that is, the 6,000, 7,000, and 8,000 levels.

The sample investigated is from various years of classes in tertiary-level institutions around Tokyo, Japan. The proficiency level of the test-takers covered a fairly wide range of general ability. At the high end of the proficiency level spectrum were test-takers from a university solely dedicated to language studies. On the other hand at the opposite end of the spectrum were test-takers from a two-year women's junior college. The results of this study can be generalized to Japanese learners of English as a Foreign Language because in the case of loanwords, Japanese language characters are non-alphabetic and non-genetically related to English. If the results are to be generalizable to other Asian languages, the study must be replicated. The results are not generalizable to languages where the characters are alphabetic and the language is genetically related to English.

The effect in studies can be either fixed, if all treatment conditions of researcher interest are included, or random, if treatments are drawn from only a random sample (Field, 2009). In this present study, both the item keywords and the persons were not randomly sampled. The item keywords were selected according to part of speech, that is, nouns, verbs and adjectives, loanword and non-loanword status and word frequency at a high, medium and low level from the JACET8000 academic word list. The persons were selected according to attendance in English subject related tertiary level classrooms. Therefore, because of the non-random sampling of both the items and persons, the effect in this study

is a fixed effect. Therefore, these fixed effects have limited generalization and replicability, restricted only to the conditions present in this study.

The Organization of the Study

Chapter 2, Review of the Literature, consists of two main sections: vocabulary testing and objective measurement theory. The chapter ends with a brief discussion of gaps in the current literature, a description of the purposes of the study, and the research questions, which guide the study. Chapter 3, Methods, encompasses participants, instrumentation, research design, and procedures. In Chapter 4, Item Specification and Construction Process, the detailed documentation of the test building process is presented. In Chapter 5, Results, the findings of the six research questions are presented. In Chapter 6, Discussion, an evaluation and interpretation of the study results are expanded upon. In Chapter 7, Conclusion, a summary of the findings, limitations, suggestions for future research and final conclusions are detailed.

CHAPTER 2

REVIEW OF THE LITERATURE

The literature review is divided into two main sections, vocabulary testing and objective measurement theory. This division was inspired by Wright's (1997) assertion of humankind's two greatest feats, language and numbers. Language in this study focuses on vocabulary while numbers focuses on objective measurement theory and multiple-choice testing. The aim of the present study is to bring both of these developments together so that the analyses of the test can cover both language, through the investigation of synonymy, and numbers, through the investigation of Rasch model (Rasch, 1960; Andrich, 1978) analyses and subsequently various parametric and nonparametric statistical procedures.

Vocabulary Testing

The vocabulary section is divided into two sub-sections, dichotomous size and polytomous strength. The main studies that guide the present study are displayed in Table 1. Nation's (1983) Vocabulary Levels Test is an influential and commonly used vocabulary size diagnostic test using three dichotomously-scored definitions to be matched to word forms in item clusters. Smith's (1987) Partial Knowledge of Vocabulary Test is a rarely used vocabulary strength proficiency test using a multiple-choice format with polytomously-scored single best answer synonymy responses per item. Nation and Beglar's (2007) Vocabulary Size Test is an influential and commonly used vocabulary size

proficiency test using a multiple-choice format with dichotomously-scored single best answer definition responses per item.

Table 1

Scoring Approaches of Vocabulary Tests Influencing this Study

Dichotomous size	Polytomous strength
Vocabulary Levels Test: Diagnostic	
Nation (1983): CTT/TSM	
	Partial Knowledge of Vocabulary Test: Proficiency
	Smith (1987): RMT/RSM
Nation (1990) CTT/TSM	
Beglar & Hunt (1999): RMT/DRM	
Beglar (2000): RMT/DRM	
Schmitt (2000): CTT/TSM	
Schmitt, Schmitt, & Clapham (2001): RMT/DRM	
Vocabulary Size Test: Proficiency	
Nation & Beglar (2007): Introduction	
Beglar (2010): RMT/DRM	

Note. CTT = Classical Test Theory; TSM = True Score Model; RMT = Rasch Measurement Theory; RSM = Rating Scale Model; DRM = Dichotomous Rasch Model.

Dichotomous Vocabulary Size

Even though there are many studies of vocabulary, including Meara and Buxton's (1987) yes/no test, this literature review is restricted to multiple-choice tests. The rationale underlying this decision is that yes/no tests are irrelevant to this study because even though the claim can be made that they test semantic ability, they might be testing merely word recognition. Therefore, the following literature review only includes research of direct relevance to this study.

Vocabulary can be measured in two basic ways. The first way is in terms of dimensions such as size (breadth) and strength (depth). From here on in this study, the terms *vocabulary size* and *vocabulary strength* are used. Vocabulary size refers to how many words are known by a person to one degree, while vocabulary strength refers to how well a person knows a particular word to incremental degrees. Strength can refer, for example, to knowing the multiple meanings of a polysemous word. The second way vocabulary can be measured is in terms of skills, that is, receptive use (i.e., reading or listening) or productive use (i.e., writing or speaking). The most straightforward way to measure vocabulary form-meaning match is to develop an instrument that tests written receptive vocabulary size. The rationale underlying this approach is that simpler test design methods can be employed including yes/no tests; however, yes/no tests do not necessarily measure a form-meaning match despite what claims a researcher makes. A yes/no test might merely measure the level of form recognition. Therefore, the most fundamental way of measuring form-meaning match is to develop a written receptive vocabulary size test. Indeed, many language researchers might equate the concept of a vocabulary test item exclusively with a form-meaning match, such is the matching test format's pervasiveness.

Vocabulary Levels Test (VLT)

Nation's (1983, 1990) frequency-based Vocabulary Levels Test (VLT) was the first widely accepted receptive vocabulary diagnostic instrument. The VLT consisted of five levels, 2,000, 3,000, 5,000, 10,000, (or more precisely, based on frequency, the second 1,000, the third 1,000, the fifth 1,000, and the tenth 1,000 words) and University Word

Levels, based on Xue and Nation's (1984) 836-word University Word Level (UWL) list. The five levels are hereafter referred to as the 2K, 3K, 5K, and UW levels. Each of the original five levels contained 18-items, producing a 90-item test. The VLT is a diagnostic test and the problem with using it as a size test is that multiple levels below the 10,000 word frequency level (i.e., 1K, 4K, 6K, 7K, 8K, and 9K) are not tested. These missing levels result in an underestimation of a test-taker's vocabulary size estimate. An additional problem is that the VLT is sensitive to partial knowledge because the words in each testlet cluster are not related orthographically, phonologically, or semantically. An example of a 2K testlet cluster from a Vocabulary Levels Test version (Schmitt et al., 2001, p. 82) follows.

Vocabulary Levels Test (VLT) Cluster Example

1. copy	
2. event	a. end or highest point
3. motor	b. this moves a car
4. pity	c. thing made to be like another
5. profit	
6. tip	

Therefore, even if test-takers have little knowledge of particular words they might be able to choose the correct response. Thus, this test might produce somewhat generous estimates of lexical knowledge.

A VLT amendment and validation by Beglar and Hunt (1999) parsimoniously focused on the 2K and UW levels, expanding both levels to 27-items each, over two VLT forms (Form 1 and 2) and documenting validation evidence. After Beglar and Hunt's

(1999) initial amendment and validation, Coxhead (2000) developed the 570-word Academic Word List (AWL), which superseded the UWL.

Subsequent VLT amendments and validations by Schmitt (2000), and Schmitt, Schmitt, and Clapham (2001) developed a new level, the Academic Vocabulary Level (AV) based on the AWL, and expanded the five levels, that is, the 2K, 3K, 5K, 10K, and AV word levels to 30-items each, over two VLT versions (Versions 1 and 2).

However, despite these two major amendment and validation sets, and the fact that the VLT is considered by second language (L2) vocabulary experts as the closest test our profession has to a standard and accepted vocabulary instrument, the VLT is problematic. The problems stem from the most recent amended versions of the VLT (Schmitt et al., 2001) being a very general estimate of receptive vocabulary breadth. When examining word level representation, at maximum, the AV level represents 30 out of 570 words, or 5.26%, while the other levels, that is, the 2K, 3K, 5K, and 10K word levels, represent 30 out of 1,000 words, or 3%, each. Also, each of the VLT five word levels is divided into ten clusters containing six words and three definitions. Therefore, three definitions should be matched to three words from an initial six word choices leading to the problem of dependency between items within subtest clusters.

Therefore, Nation's (1983) original VLT has had two major amendments and validations by Beglar and Hunt (1999) on one hand and Schmitt (2000) and Schmitt et al. (2001) on the other hand. Both of these study sets have led to a progressive expansion of representation, both in terms of items tested per level from 18 to 27 to 30, and the inclusion of a new academic vocabulary level based on the Coxhead's (2000) Academic Word List.

Unfortunately, because of the VLT's test design, which uses randomly selected decontextualized word forms focused on partial semantic knowledge, the general estimate provided is very generous. Despite this susceptibility to partial word knowledge, the VLT uses dichotomously scored data, with 0 assigned for an incorrect response and 1 assigned for a correct response.

In 1993, Schmitt (Schmitt et al., 2001) updated the VLT version in Nation's (1990) book version to create Version A, while additionally developing Versions B, C, and D. This seemed to solve the problem of limited variation in the original VLT, which presented only one version of the test. However, these four new versions were still only limited to having 18-items or definitions within each word level. There was still room to expand the number of items within each word level.

Beglar and Hunt (1999) solved this second criticism of the original VLT being limited by the amount of items at each word level. Beglar and Hunt (1999) analyzed the existing four versions (A, B, C, and D) and two levels (2,000-word and UWL) of the five available levels (2,000-word, 3,000-word, 5,000-word, UWL, and 10,000-word, each consisting of 18-items) of the VLT. They then revised the VLT into two forms (A and B) and two levels (2,000-word UWL), consisting of an expanded 27-items per level. The 2,000-word and UWL (836 words) levels were chosen for their importance, as they make up 87% and 8% (totaling 95%) respectively, of the words in an average text. These combined 2,836 words are close to the 3,000-word benchmark suggested by Laufer (1992) for academic reading comprehension success.

The purpose of Beglar and Hunt's (1999) study was threefold: (a) examine and compare descriptive statistics, reliability coefficients, and degree of equivalence in the existing VLT; (b) generate new, reliable and equivalent forms, and (c) validate these revised forms.

The existing 2,000-word and UWL levels of the VLT were given to 496 and 464 participants, respectively. These Japanese EFL participants came from high schools, junior colleges, and universities. An examination of the descriptive statistics indicated some large statistically significant differences in the means (2,000-word level Form C, 12.52 and Form D, 14.81), standard deviations (UWL level, Form A, 3.12 and Form C, 4.02), and reliabilities (Form A, 0.65, and Form C, 0.80) between the existing VLT forms. Repeated-measures ANOVAs suggested that the forms were statistically significantly different at both levels.

The revised two-form VLT's descriptive statistics indicated that the new means and standard deviations were closer than the existing forms. Internal consistency reliabilities using Cronbach's alpha indicated acceptable to high reliabilities for the revised forms. Equivalence using classical equating methods was established by finding no statistically significant differences between means, variances and covariances (Henning, 1987, in Beglar & Hunt, 1999). Additional results indicated that the correlations of the forms with the full TOEFL test were strong. Using Hotelling's *t*-test, the full TOEFL test and the 2,000-word forms correlations did not differ statistically significantly, while the full TOEFL and UWL forms correlations did differ significantly. The correlations of the revised VLT with the TOEFL test Reading Comprehension subsection were higher than the

other subsections correlations. Further, the revised test items were analyzed using a dichotomous Rasch analysis, the results of which were recommended to be used for equating other forms and gauging new items. After a retrial, results from repeated-measures ANOVA tests indicated a statistically significant difference for the 2,000-word forms, and no statistically significant difference for the UWL forms.

Test validity of the revised forms was established by examining three types of validity: face, content, and construct. Face validity was indicated by the test being taken seriously, with English and Japanese instructions given, and allowance for adequate completion time. Concerning content validity, the focus was on the degree the revised VLT represented the two levels. The revised 2,000-word levels represented less than 1.4% of the General Service List (GSL) (West, 1953, in Beglar & Hunt, 1999) and 3.4% of the UWL. As both these small sample sizes did not represent reliable estimates, according to Meara's (1996, in Beglar and Hunt, 1999) criterion, the use of check-list tests was suggested. Construct validity was indicated by high reliability coefficients, moderate to high correlations between forms, a dichotomous Rasch model (Rasch, 1960) analysis, revised VLT and TOEFL scores comparisons, correlations between the revised VLT, TOEFL full and TOEFL subsections, and results from a Principal Components Analysis.

The strengths of Beglar and Hunt's (1999) study were many. They took the existing VLT (4 forms, with 5 levels of 18-items), analyzed 2 of the levels (2,000-word and UWL) and created a revised VLT (2 forms, with 2 levels of 27-items). The revised VLT was a more effective and powerful test for this sample as it was shorter (54-items total instead of the existing VLT's 90-items), more reliable according to Cronbach's alpha internal

consistencies, had greater equality between forms, and took less completion time. The participants in the study encompassed a large sample, and the structure involved piloting, the main study, and re-trialing of the revised forms.

The limitations of Beglar and Hunt's (1999) study included stated small sample percentages in the revised forms, which indicated that these were not reliable estimates. As a result content validity was affected. A larger number of items to be tested were recommended especially in the case of norm-referenced testing. To achieve this, a suggestion to combine both forms to obtain 54-items at each level was given. A questionnaire was completed by three second language vocabulary experts, and the results indicated that two of these experts responded that a 27-item per level test was sufficient to give an approximate projection of vocabulary breadth. Beglar and Hunt (1999) suggested that longer test forms could achieve even greater reliability and content validity. It took another two years before Schmitt, Schmitt, and Clapham (2001) created two versions of a slightly longer test containing 30-items per word level.

However, in the meantime between 1999 and 2001, a new specialized word list evolved on the vocabulary scene: the Academic Word List (AWL) developed by Coxhead (2000). From a corpus of approximately 3.5 million headwords, the AWL was developed and included 570 word families. The AWL was divided into 10 sublists according to frequency, with sublist 1 representing the most frequently occurring academic words, and sublist 10 representing the least frequently occurring academic words. The 570 headword AWL replaced the 836 headword University Word List (UWL) (Xue & Nation, 1984) used in the original VLT. The AWL had a smaller number of headwords, was based on a larger

corpus, and was therefore more powerful than the superseded UWL. Coxhead (2000) suggested that the AWL could be used to construct suitable teaching matter, and the words in the AWL were suitable for direct learning methods.

Schmitt, Schmitt, and Clapham (2001) analyzed the original four-form, five-level VLT and developed a revised two-version (Versions 1 and 2), five-level VLT. The revised five-level VLT incorporated words from the new 570-word AWL (Coxhead, 2000), which superseded the 836-word UWL (Xue & Nation, 1984). Therefore, the new format included the 2,000-word, 3,000-word, Academic Vocabulary (AV), 5,000-word, and 10,000-word levels, consisting of 30 items per level.

The stated purpose of Schmitt et al.'s (2001) study was to fill a gap in the literature, that is, to solidly validate the VLT. They analyzed, and then revised an expanded two-version, five-level VLT, which they validated (including some post-VLT participant interviews), and attempted to equate the versions. There were 801 ESL participants from various first language backgrounds who took part.

Validity of these revised versions was determined by examining seven different aspects. Schmitt et al. (2001) perceived the first six aspects as indirect and the seventh as a direct example of validity evidence. The seven aspects were: first, native speaker results indicated that nine participants did very well, with no problems and near-maximum to maximum scores. Second, item analysis indicated that the individual items worked well. Even though it was expected items were not independent in each cluster, dichotomous Rasch model (Rasch, 1960) analyses indicated that they acted independently. Third, a profile of the sections indicated that participants learnt high-frequency words before low-

frequency words. Fourth, factor analysis indicated that there was unidimensionality, or one major underlying factor. Fifth, reliability indices indicated by Cronbach's alpha were high for both versions. Sixth, practicality was indicated by the reasonable length of 31 minutes on average to complete.

Practicality was also indicated by it being fast and easy to photocopy, grade, and requiring no special equipment. Seven, post-test interviews were conducted with 22 participants from various backgrounds. The interviews were conducted over 20-40 minutes immediately after interviewees had finished the revised written Version 1. The interviewees were asked about the reasonableness of the test, the processes they used to answer, and to define 50 target words from the 150 items of the VLT Version 1.

Equivalence was established in three ways: means, variance, and covariance (Henning, 1987, in Schmitt et al., 2001). For the whole sample, the two versions means and variances were not statistically significantly different for any level. Further, the covariances were satisfactory (Pearson's correlation coefficient values were above .90). However, when the sample was divided into three subgroups: Romance, Non-Romance, and Slovakian speakers, only the 2,000-word level was equivalent in all analyses.

Further research was recommended to answer the questions: were the items in clusters actually independent? Is this revised VLT suitable for other participants from different L1 backgrounds? What is the ambiguous benchmark used for individuals when there is the rule of no blind guessing, only guessing when they think they know?

Schmitt et al.'s (2001) stated aim was to address a gap in the research, that is, to present a primary validation of two amended and enlarged VLT versions. The amended

versions were outlined to provide an approximate guess of receptive vocabulary breadth at relevant levels identified as revealing for the researcher. However, Schmitt et al. (2001) contradictorily critiqued Beglar and Hunt (1999) who "looked at and revised the (unattributed) 2000 and University Word List sections of Versions A-D" (p. 57). Schmitt et al. (2001) subsequently reversed this position by stating if "a shorter test is desirable, it may be reasonable to exclude certain sections." (p. 71).

Schmitt et al.'s (2001) study examined participants of general and academic English from a variety of L1 backgrounds. Unfortunately no heed was paid to Nation's (1990, p. 262) caveat that the VLT was "not suitable for learners whose mother tongue is a language which has been strongly influenced by Latin," in specific reference to a Spanish participant during the piloting of the original VLT. Regardless of this warning, Schmitt et al.'s largest sub-group of participants was Spanish, consisting of 322 of the 801 participants.

Schmitt et al. (2001) referred to Messick (1989), who stated that test validity should include quantitative and qualitative evidence. For this reason, in addition to quantitative, qualitative procedures were also used by Schmitt et al., by specifically using post-test interviews. However, the use of these interviews as an example of qualitative procedures was problematic for a variety of reasons. First, Brown (2001) stated that survey research (based on interviews and questionnaires) is unique, in that it is a type of survey research that is distinct from statistical (quantitative) research and qualitative research. Second, the use of productive skill tasks like orally defining 50 items in the interview does not provide validation for the receptive skill written version of the test. The 50-item definitions data was collected quantitatively in 20-40 minutes, and rated as known or unknown, by the

raters. This left the remaining time in the interview to ask qualitatively about the participants' attitudes about the VLT, and what processes they employed to answer the written items. Schmitt et al. confided that the interviews were time-consuming and did not cover a wide sample (only 22 of the original 801 participants). However, more importantly, these interviews were not representative of the original sample as 18 of the 22 interviewees were from Asian backgrounds, while the largest original sample sub-group, Spanish (322 of the original 801 participants), were not interviewed at all. An alternative to interviews is the use of questionnaires such as those presented by Brown and Rodgers (2002). Such questionnaires can be a more efficient way to approach large-scale data collection.

Schmitt et al. (2001) critiqued Beglar and Hunt (1999) who reported their revised VLT correlated with TOEFL test scores. Schmitt et al. stated that as the TOEFL test examines general proficiency and not specifically vocabulary, then it is not an acceptable measure. However, Read (2000) included the TOEFL test as one of four well-known instruments that assess vocabulary knowledge. Read (2000) pointed out that the VLT which assesses vocabulary in isolation, is different to the TOEFL test which in its development from 1964 to 1995, has progressively contextualized vocabulary in sentences and later in texts. In reference to Beglar and Hunt's (1999) use of the TOEFL test results, Schmitt et al. (2001) stated these contradictory sentences: "It would be more informative to compare the test with other accepted tests of vocabulary size. However, as mentioned before, the Levels Test is the closest thing we have to an accepted measure" (p. 60).

Schmitt et al. (2001) in an indirect critique of Beglar and Hunt's (1999) use of Rasch model (Rasch, 1960) analyses stated that this has become an established way of

language test analysis. However, Schmitt et al. stated that local independence of items is one of the assumptions that must be met when using the Rasch model (Rasch, 1960). Schmitt et al. (2001) stated that "Strictly speaking, the vocabulary items presented together in clusters cannot be considered independent" (p. 61). In contrast, even though they expected to find no independence of items, after conducting item analysis, Schmitt et al. (2001) subsequently admitted that Rasch model analyses indicated that the items were independent of one another.

Schmitt et al. (2001) stated that one of the rules of the test was no blind guessing; the only guessing allowed was when participants thought they knew the answer. However this rule leaves a lot to be desired, as this vague distinction differs to an unknown extent between individuals. In contrast, Schmitt et al.'s item analysis indicated that guessing was not widespread, while their interview results suggested nearly 23% of interviewees admitted guessing when they did not know the vocabulary.

Equivalence was established by examining three criteria according to Henning: (1987, in Schmitt et al., 2001) means, variance, and covariance. For the full sample, equivalence was reported as the means and variances of both versions were not statistically different. Further, covariances (above .90) were satisfactory. However, when the sample was broken down into sub-groups of Romance, Non-Romance, and Slovakian L1 speakers, only the 2,000-word frequency level was equivalent.

Vocabulary Size Test

Nation and Beglar (2007) introduced the Vocabulary Size Test (VST) in a published form. The VST's purpose was to provide a proficiency measure to understand how much vocabulary was known. This was in contrast to the VLT's purpose of being a diagnostic test. The VST was a progression over the previous VLT in several aspects. The VLT had three gaps in the word frequency levels, in particular, the fourth, sixth, and ninth 1,000 word frequency levels. The VST uses all of the first 14 vocabulary levels, that is, representing 14,000 words, from the British National Corpus (BNC) (Leech, Rayson & Wilson, 2001). Instead of clustering items into bundles or testlets like the VLT, the VST separated items so that inter-item dependency issues were obsolete.

Beglar (2010) applied Rasch dichotomous model analyses to the VST. Preliminary construct validity evidence for the 140-item instrument covered six aspects: the content, substantive, structural, generalizability, responsiveness and interpretability aspects. The overall construct validity of making interpretations and inferences from the measures was found to be psychometrically sound. The test design aimed to measure the written receptive knowledge of native and non-native speakers of English on the 1,000 to 14,000 (1K to 14K) levels from the British National Corpus (BNC) (Leech, et al., 2001). The primary purpose of the study was to determine item accuracy. Three a priori hypotheses outlining item difficulty, person group ability and test unidimensionality were confirmed by the analyses. The VST overcame the shortcomings of the VLT by providing greater representation of all word frequency levels, instead of staggered VLT levels. Therefore, the VST could be classed as a vocabulary size test instead of the diagnostic label referring to

the VLT. The VST was found to increase the range of measurement and was thus a robust new test of written receptive vocabulary size. The VST has the potential to replace the VLT as the go to test for teachers wishing to not only determine vocabulary diagnostic information but further to determine vocabulary size information of their learners.

In research investigating the VST and loanwords, T. Nemoto (personal communication, July 19, 2015) indicated that many Japanese students do not recognize keywords (for example, standard, pub, microphone, strap, quiz, crab, vocabulary, cube, accessory, quilt, yoghurt, palette and cabaret) as loanwords from English when encountered in a written form only. However, the students recognize the keywords as loanwords from English when they listen to the spoken form.

Polytomous Strength

Smith's (1987) study is unique in that it applied the Rasch-Andrich rating scale model to vocabulary strength. Specifically, the study was only concerned about rating scale functioning and ignored the item difficulty hierarchical order which is usually mandatory and visually displayed by a Wright Map in most Rasch model measurement analyses. The absence of a Wright Map was conspicuous in Smith's study. Even more peculiar, as the focus was on rating scale functioning was the absence of a category characteristic curve map figure. Despite Smith's (1987) claims of moderate support for O'Connor's (1940) five developmental stage vocabulary theory, all four thresholds were disordered in this study, indicated by thresholds 1, 2, 3, and 4 displaying values of 0.33, 0.23, -0.06, and -0.50, respectively. According to Andrich's (1978) requirement of ordered empirical thresholds

for polytomous Rasch measurement, these categories should be collapsed and then a new test version should be administered to a fresh sample.

Rasch Measurement Theory

In contemporary testing theory there are three major strands: Classical Test Theory (CTT), Item Response Theory (IRT) and Rasch Measurement Theory (RMT) (Bezruczko, 2005). At their primary levels, these theories employ one or several models. CTT utilizes a True Score Model (TSM), IRT alternates between either a 2- or 3-Parameter Logistic Model (2-PLM or 3-PLM) or Samejima's (1969) polytomous model, while RMT applies the original Dichotomous Rasch Model (DRM) (Rasch, 1960) or either of two extensions of the dichotomous model in the form of polytomous models: the Rasch-Andrich Rating Scale Model (RSM) (Andrich, 1978) or Rasch-Masters Partial Credit Model (PCM) (Masters, 1982). Refer to Table 2 for a graphical representation. Note that Table 2 in this study was inspired by Bezruczko's (2005) Table 1 (p. 3).

In this study, several terms need to be defined in Table 2. Statistical theory type refers to the underlying basis of parametric testing procedures where data is perceived as sacrosanct possibly to the detriment of construct theory. Measurement theory type refers to the fundamental focus on measuring a unidimensional construct where data is not sacrosanct.

However, data is required and necessary before any measurement can commence. Classical test theory is referred to as having partial utility, or more specifically, being obtuse, that is, it is too simplistic, unclear and imprecise to the point that it is difficult to

comprehend. Item response theory is referred to having partial utility, or more specifically, as being abstruse, that is, it is too complicated, undecipherable and inexplicable to the point of being difficult to comprehend. Rasch model measurement theory is defined as being useful, that is, it is worthwhile, rewarding and helpful to the point of being easy to comprehend.

Table 2

Three Main Statistical and Measurement Theories

	Classical Test Theory (CTT)	Item Response Theory (IRT)	Rasch Measurement Theory (RMT)
Theory type	Statistical	Statistical	Measurement
Measuring utility	Partial	Partial	Useful
Model founder	Spearman, 1904	Birnbaum, 1968	Rasch, 1960
Model exemplar/s	TSM	2-PLM and 3-PLM	DRM
Polytomous model	---	Samejima, 1969	Andrich, 1978
Polytomous exemplar	---	LGRM	PRM
Model role	Fixed	Expendable	Invariant
Construct role	Subordinate to data	Subordinate to data	Predominant to data
Analysis focus	Test data	Item data	Model/Construct
Data role	Descriptive	Descriptive	Prescriptive
Data transformed	No	Yes	Yes
Data level	Ordinal	Ordinal	Interval or ratio
Parameters	2, united	1 or more, separate	2 or more, separate
Sample dependent	Yes	Yes	No

Note. TSM = True Score Model; PLM = Parameter Logistic Model; DRM = Dichotomous Rasch Model; LGRM = Logistic Graded Response Model; PRM = Polytomous Rasch Model. Adapted from *Rasch Measurement in the Health Sciences* (p. 3), by N. Bezruczko, 2005, Maple Grove, MN: JAM Press. Copyright [2005] by R.M. Smith. Adapted with permission.

Multiple-Choice Testing

Multiple-choice question tests were invented by Kelly in 1914 (Gatfield & Larmar, 2006). Otis was one of the first researchers in 1917 to put the new multiple-choice question tests to a full-scale Army selection use in 1917 (Jones & Thissen, 2007). In the century of multiple-choice popularity that has ensued since then, typical multiple-choice items tend to require only a single best answer response and a number right score. However, some studies have required a step beyond this unary response/number right score requirement.

One of the first alternative formats was proposed by Nedelsky (1954). He analyzed right (R), wrong (W) and obviously wrong (F) (gross misjudgment) responses in a physics test. In addition, he also analyzed a composite (C) score, $C = R – F/f$, where f was the average frequency of F responses in a particular item. Despite the findings being narrow in focus and no generalizations could be drawn, at least there was an attempt to measure partial knowledge. This is recommended as useful in many fields where "mastery, in the sense of getting the right judgments or conclusions, is beyond the reach of a sizable proportion of the students" (Nedelsky, 1954, 471).

Partial knowledge was assessed via three methods of analyzing four-option multiple-choice test response methods by Coombs, Milholland, and Womer (1956). First, a conventional (C) method, second, an experimental (E) method, and third, a method using both (B) (i.e., C and E). The C method required a typical single best answer response. The E method, sometimes called the elimination method, required eliminating only the incorrect distractors (maximum three, with the key left unmarked). These responses were scored on a seven-point scale from -3 to +3. One point was credited for each incorrect distractor

identified and 3 points were lost if the key was identified as an incorrect distractor. The B method required ranking (1, 2, 3) the incorrect distractors in order of recognizability (leaving the key unmarked). In addition, participants were required to circle distractors which they were confident could be scored under the E method.

Evidence was documented for partial information existence by analyzing data using the B (both) method by Coombs et al. (1956). For the vocabulary and the driver information tests, the number of participants having more than one-third of second choices correct was significantly more than the number of participants having one third or less of second choices correct. Therefore, partial information was indicated to exist. Coombs et al.'s B method is of particular relevance to this study. Several other researchers have used a similar or the same response format as Coombs et al.'s B method but often under different names.

A more sensitive alternative to conventional multiple-choice tests was presented by Willey (1960): the three-decision multiple-choice test. This was an attempt to add sophistication to the way multiple-choice tests were scored by trying to measure partial credit. For a 5-option item, test-takers were required to circle the correct response and write an X over two definitely wrong options. After responding to the test, punch marks are put through the correct options. The scoring rubric uses a penalty scheme where if the punch mark is through a circled square (correct) there is no penalty. If the punch mark is through an unmarked square there is a penalty of 1. If the punch mark is through an X-marked square there is a penalty of 3. Therefore, each item response could be scored out of a maximum of 3 points, that is, scores of 0, 2, or 3. It was found that this method favored

diligent test-takers instead of those who rushed through carelessly. Correction for guessing was internally consolidated within the test itself. Test-takers attitudes toward this method were it was fairer than the traditional single best answer response mandate because lucky guesses were reduced and partial credit could be gained for partial knowledge. Also, test-takers indicated that this test inspired more serious studying. Administration and scoring of this test took only slightly more time than a traditional multiple-choice test. A double point spread of scores is possible with this method in comparison to traditional multiple-choice single best answer response scoring. Variations for 3-option multiple-choice items (with one option marked completely wrong) and 4-option items (with two options marked completely wrong) were also suggested.

Bernhardson (1967) compared conventional multiple-choice tests, modified conventional multiple-choice (half scores given to key answers in second choice position) and three-decision multiple-choice tests. Two studies were conducted, one with items having five alternatives and another with items having four alternatives. In the first study, test-takers responded with a check mark for the correct option, a question mark for the next probably correct answer and two X-marks for definitely wrong options. Despite having the question mark for the probably correct option, there is no indication in this study of the options being written according to some theory and also no example items or options. These responses were scored according to three methods: three-decision multiple-choice (scored 0, 2, 3), modified conventional multiple-choice (scored 0, 1.5, and 3) and conventional multiple-choice (scored 0 and 3). Correlations of these scores with academic achievement scores indicated that conventional multiple-choice tests (.52) were not

improved by the two alternative scoring methods (three decision, .51 and modified conventional, .55). The second study involved the use of four alternative multiple-choice items comparing three-decision multiple-choice test scores and conventional multiple-choice test scores correlated with academic achievement. Again, the correlations indicated that conventional multiple-choice test scoring (.49) were not improved greatly by the three decision scoring, .50. Bernhardson recommended future research into the robustness of three-decision tests.

Poizner, Nicewander, and Gettys (1978) compared three response and scoring methods for multiple-choice items. The first was binary scoring or traditional multiple-choice single best answer item responses, the second was probability scoring where test-takers indicate their degree of correctness belief for each of five item options, for example, (a) .60 (b) .10, (c) .00, (d) .30, and (e) .00, and third was ordinal scoring where test-takers rank the alternatives. The ordinal scoring procedure is similar to Coombs et al.'s (1956) B method except that all options should be ranked. For example, in a five-option multiple-choice item, respondents rank the options from 1 (*least correct*) to 5 (*most correct*). In both of the two studies conducted by Poizner, Nicewander, and Gettys, the partial credit response and scoring methods, that is, probability and ordinal scoring procedures, were more reliable than the binary response and scoring procedures. They recommended that the ordinal response and scoring mode be used with adults and high school students as it is easily taught and administered and can lead to gains in reliability.

One of the aims of this study is to propose a new version of the traditional multiple-choice test format. The rationale for designing and constructing the new format is to assess

the viability of the a priori metasynonymy hypothesis based on an amended vocabulary strength theory by O'Connor (1940).

Gaps in the Literature

There are gaps in the research literature as most researchers have investigated dichotomous scoring of multiple-choice tests with single best answer (SBA) response instructions. These gaps in the research literature are a result of deficiencies at many levels. These research deficiencies are a result of test designer's perceived impossibility of incremental measurement and persistence in perpetuating only traditional test formats, a reliance on word frequency acquisition without reference to theory, limited research into loan words acquisition theory, no research in foreign language synonymy acquisition theory or foreign language test-taker guessing and carelessness.

Smith (1987) investigated partial knowledge on a multiple-choice test using the Rasch-Andrich rating scale model (Andrich, 1978). However, despite following Andrich's (1978) requirement for dominantly unimodal, that is, peaked, categories and ordered thresholds, all four of the thresholds in Smith's study were disordered. As Smith only required respondents to give one unitary 'full knowledge' response per item, 'partial knowledge' was not fully measured. Therefore, a truly incremental vocabulary knowledge test is the gap in the review of published literature. This gap is filled by constructing a new multiple-choice test design in this study.

Multiple-choice items are typically binary scored; however, in this study they are scored atypically. For example, in the case of an individual item, the single best answer (SBA) response (in sequential order) is *cab* (where c = correct).

Table 3 indicates a typical multiple-choice single best answer (SBA) response mandate scoring rubric. The DRM row is representative of Nation and Beglar (2007), while the RSM row is representative of Smith (1987).

Table 3

Typical Multiple-Choice Item Single Best Answer (SBA) Response Scoring Rubric

	c	a	b
DRM	1	0	0
RSM	2	1	0

Note. DRM = Dichotomous Rasch Model; RSM = (Rasch-Andrich) Rating Scale Model. RSM has two thresholds in this case.

Table 4 indicates an atypical multiple-choice ordered triple rank (OTR) response instruction scoring rubric. The rationale underlying ordered triple rank response instruction is to measure beyond full knowledge from a solitary single best answer response. Indeed, as

Table 4

Atypical Multiple-Choice Item Ordered Triple Rank (OTR) Response Scoring Rubric

	cab	cba	acb	abc	bca	bac
DRM	1	0	0	0	0	0
RSM	5	4	3	2	1	0

Note. DRM = Dichotomous Rasch Model; RSM = (Rasch-Andrich) Rating Scale Model. RSM has five thresholds in this case.

is explained more in detail, the ordered triple rank response instruction bifurcates typical single best answer responses. Furthermore, the ordered triple rank response instruction measures incremental partial knowledge. In this study, the combination of full and partial knowledge consolidates in incremental knowledge degrees that are scalable using polytomous Rasch model (Andrich, 1978) measurement. The ordered triple rank responses are hypothesized to form a hierarchy of metasynonymy awareness. The rationale underlying this is that test-takers are not just answering a typical single best answer response; instead, test-takers are thinking about the progression of synonymy over a developmental scale.

Purposes of the Study

The primary purposes of this study are to address the previously outlined gaps in the research. Therefore, the purposes are to address test designer's perceived impossibility of incremental measurement and persistence in perpetuating only traditional test formats, a reliance on word frequency acquisition without reference to theory, limited research into loan words acquisition theory, no research in foreign language synonymy acquisition theory and limited research on foreign language vocabulary test-taker guessing and carelessness.

A major objective of this study is to test vocabulary size and strength theories, with an emphasis on vocabulary strength theory. First, the purpose is to test vocabulary size via Palmer's (1917) vocabulary frequency theory that relates loan word status to item difficulty. The independent variable is generally defined as English loan word status, that is, non-loan and loan word classification (ordered difficult to easy. The dependent variable is generally

defined as item difficulty calibration. Further, to test Palmer's (1917) theory in regard to regular word frequency bands. Second, the purpose is to test vocabulary strength via an amendment to O'Connor's (1940) vocabulary stage theory that relates synonyms, context words and look like words (ordered difficult to easy) to the polytomous threshold calibration. The independent variable is generally defined as developmental stage categories, that is, synonym (correct), context (partially correct) and look like (incorrect). The dependent variable is generally defined as the polytomous threshold (crossover between categories) calibration. The resulting a priori metasynonymy awareness hypothesis is operationalized through the construction of the new multiple-choice test item design.

Research Questions

Applying polytomous Rasch model measurement theory, this study is an investigation of the development of a new multiple-rank response test format. Of particular interest is communicating meaningful inferences to a varied-ability, that is, novice to expert, audience. Therefore, the following six research questions guide this study:

1. Does the test meet the unidimensional Rasch model requirements?
2. Is the hypothesized easiness of loan words and difficulty of non-loan words confirmed by the data?
3. Is Palmer's (1917) word frequency acquisition theory confirmed by the data?
4. Is the a priori metasynonymy hypothesis confirmed by the data?

5. How do incremental vocabulary knowledge estimates change depending on the type of analysis utilized?
6. To what degree is guessing and carelessness apparent in the type of analysis utilized?

CHAPTER 3

METHODS

Participants

The participants in this study were tertiary-level coeducational university students. Initially, there were 403 respondents in total. However, as this study's purpose was to investigate Japanese persons, 24 non-Japanese persons were left out of the final analysis. As a result, 379 Japanese, including 196 women and 183 men, participated in the study. The age range was 99 18-year-olds, 129 19-year-olds, 93 20-year-olds, and 32 21-year-olds, 18 22-year-olds, 5 23-year-olds, 1 24-year-old, 1 25-year-old and 1 26-year-old. The grade range was 230 first-year, 92 second-year, 41 third-year, 9 fourth-year and 7 graduate students. The test-takers were selected from those enrolled in English communication classes. The participants' general English proficiency is high as they come from a range of tertiary institutions, which includes one of the higher elite private coeducational foreign language universities in the Tokyo region.

Most of the 379 persons tested, that is, 359 persons, had completed the Test of English for International Communication-Institutional Program (TOEIC-IP) test. The total score, that is, for the listening and reading comprehension sections combined, mean for these 359 persons on the TOEIC-IP test was $M = 573.0$ with a standard deviation, $SD = 136.8$ (range = 640, minimum = 260 and maximum = 900). The participants have had six years of formal English education in junior and senior high schools and were all non-English majors. The respondents were informed of the study verbally and all agreed to

voluntarily partake with the option of discontinuing at their own will. The test-takers were rewarded for their involvement in the three pre-test quizzes and the test with a chocolate gift after the completion of each administration.

Instrumentation

Polysemous Academic Vocabulary Extent Test (PAVET)

The Polysemous Academic Vocabulary Extent Test (PAVET) is designed to measure receptive incremental vocabulary extent, that is, size (breadth) and strength (depth) through a ranking task (rank ordering three options in sequence of synonymy degrees). It consists of 36 items, which have been selected from Coxhead's (2000) Academic Word List (AWL) and sequenced according to JACET8000 academic word list (Aizawa, et al., 2008) frequency bands. The rationale behind this is that the JACET8000 word list has been specifically ordered for the Japanese context, whereas the AWL has been ordered for an international context.

In regards to vocabulary size, the first half of the sequenced items, that is, the first 18 item stems include bolded loan keywords (predicted to be easier) and the last 18 item stems include bolded non-loan keywords (predicted to be more difficult). These a priori predictions are based on an amendment to Palmer's (1917) vocabulary frequency theory that more frequent words are easier to learn. Within each of these loan- word status groups, the items have been sequenced according to their ranked order in the JACET8000 vocabulary list (Aizawa, et al., 2008). The items were sequenced according to frequency, with high frequency being 1,000 to 3,000, medium frequency 4,000 to 5,000, and low

frequency being 6,000 to 8,000 academic word frequency levels. This sequenced ordering was also according to Palmer's (1917) vocabulary frequency theory.

Each multiple-choice item is comprised of a stem and three options. However, in contrast to traditional multiple-choice items, the items design and outcome space (Wilson, 2005) has a requirement that all three options should be sequenced into an ordered triple rank (OTR) response according to incremental synonymy degrees.

The three options are based on an amendment of O'Connor's (1940) first language (L1) vocabulary acquisition stage theory. O'Connor (1940) proposed five stages: synonym, antonym, context, looks-like and sounds-like (ordered difficult to easy, respectively). However, Smith's (1987) study indicated that five categories based on O'Connor's (1940) theory resulted in disordered and reversed thresholds. These disordered five thresholds suggest that five categories are too many to result in meaningful measurement. In addition, extensive pre-pilot testing before the main present study resulted in the decision to include only three of O'Connor's (1940) stages. The amended stages used were reduced to three categories: synonym, decontextualized, and null. As regards the decontextualized option, the test-takers should imagine another sentence context in which the near-synonym would match. An example item follows:

His **unique** concern is finding happiness.

unique: a. unusual b. sole c. common

The scoring of the responses is again based on O'Connor's (1940) theory. Even though the responders answer with a three-part ordered triple response, the responses are

scored according to six levels from a scoring rubric. As an example, the previously displayed Item 1of a printed test might have an item stem and options with an administrated order of decontextualized, synonym and null. Decontextualized in this case means that the keyword can match the decontextualized word in another sentence (e.g., *That is a unique (unusual) haircut*). Therefore, the completely correct ordered triple response would be null (c), decontextualized (a) and synonym (b) ordered from left to right. In contrast, the completely incorrect vector response would be null (b), decontextualized (a) and synonym (c). In between the completely correct ordered triple rank response and the completely incorrect ordered triple rank response are four other levels of incremental partial knowledge information. A scoring guide for item 1 follows:

OTR responses			OTR
Null	decontextualized	Synonym	scoring
0	1	2	guide
c	a	b	5
a	c	b	4
c	b	a	3
b	c	a	2
a	b	c	1
b	a	c	0

The rationale for O'Connor's (1940) acquisition order is that as learners progress from being a novice to a veteran, they go through several stages of synonymy acquisition. They are not errors as many people describe them. On the contrary, they are stages that are progressed through and understanding them is useful for building learners vocabularies (O'Connor, 1940). There has been no empirical evidence to support O'Connor's (1940) five stages. O'Connor's (1940) five stages amended down to three stages can be visualized

in the column labeled Synonym "2". Therefore, from top to bottom, O'Connor's stages are labeled b, b, a, a, c, and c. However, these labels would be consolidated into the three simpler labels b, a, and c in a more traditional and typically scored rubric. In other words, the rubric visualizes that the six ordered triple responses from top to bottom of cab, acb, cba, bca, abc, and bac actually bifurcate O'Connor's amended three vocabulary development stages.

Procedures

Before the test the respondents were aurally informed of the anonymous nature of the research and given the option to not participate; all respondents participated. The test was administered to the respondents during regular class time. As the test had an unfamiliar ranking response format, three weekly quizzes of nine items each were administered in the month before the test administration.

Despite having to write three responses to each of 36 items (108 responses in total), it took an average of about 30 minutes to complete the test, so it was not so demanding time-wise. After the tests were completed, the respondents were given a chocolate gift as a sign of appreciation for their participation.

The responses were entered as data via Remark Office Optical Mark Recognition (OMR) (2010) software. Scoring the responses required making a six-way rubric template so that each item's ordered triple rank responses could be scored according to six different incremental degrees. This is similar to analyzing a single best answer response with a 6-point Likert scale. A response could be completely correct, completely incorrect, or one of

four incremental shades of partial correctness degrees in between. These scores were entered into a Microsoft Excel spreadsheet and transferred through Microsoft Notepad to Winsteps (Linacre, 2015b) Rasch model measurement software. Rasch-Andrich rating scale model (Andrich, 1978) analyses were run through Winsteps and the output was analyzed.

The polytomous Rasch model (Andrich, 1978, 2010) has a dominant category approach requirement that insists on only analyzing wide, that is, more frequently observed, categories. These are graphically displayed with the resultant ordered Rasch-Andrich thresholds in a category characteristic curve figure.

Analyses

Research Question 1: Unidimensional Rasch Model Requirements

One of the purposes of this study is to use information, which has been collected in the past to make inferences about the future. Making inferences might open new paths of understanding, research and deeper meanings. "The ability to expect and so to infer is the impetus of human development, the prime tool of civilization" (Wright & Mok, 2004, p. 6). The tool to be used in this quantitative research study is a modern measurement model. The measurement model was chosen a priori based on five essential points that facilitate inference from observation. The five essential points (Wright & Mok, 2004) that the measurement model should meet are:

> (a) produce linear measures, (b) overcome missing data, (c) give estimates of precision, (d) have devices for detecting misfit, and (e) the parameters of the object

being measured and of the measurement instrument must be separable. Only the family of Rasch measurement models solve these problems. (p. 4)

Therefore, in this study the Rasch class of measurement models is used for subsequent analyses. The two members of the Rasch family employed in this study are the dichotomous Rasch model (Rasch, 1960) and the polytomous Rasch-Andrich Rating Scale Model (Andrich, 1978). The Rasch probabilistic model, that is, the mathematical equation, for the dichotomous Rasch model (Rasch, 1960) is notated as $P_{ni} = e^{\beta n - \delta i} / 1 + e^{\beta n - \delta i}$ (Andrich, 1978; Wright & Masters, 1982). This mathematical formula notation represents the probability of a person responding successfully to an item that is dichotomously scored. This probability is equal to the natural exponential function, e, to the power of person ability, βn, minus item difficulty, δi, divided by, one plus the natural exponential function, e, to the power of person ability, βn, minus item difficulty, δi. Expanding on this dichotomous model, the polytomous Rasch-Andrich Rating Scale Model (Andrich, 1978) is notated as $P_{nix} = e^{\beta n - (\delta i + \tau x)} / 1 + e^{\beta n - (\delta i + \tau x)}$ (Andrich, 1978; Wright & Masters, 1982). This mathematical formula notation represents the probability of a person responding successfully to an item threshold that is polytomously scored. This probability is equal to the natural exponential function, e, to the power of person ability, βn, minus (item difficulty, δi, plus threshold difficulty, τx), divided by, one plus the natural exponential function, e, to the power of person ability, βn, minus (item difficulty, δi, plus threshold difficulty, τx).

The unidimensional polytomous Rasch model requirements are checked by investigating a sequence of 13 steps. The steps are organized from three groups, first,

checking the four general steps that can be also be used when interpreting the results of a dichotomous Rasch analysis, second, the eight general steps when interpreting the results of a polytomous Rasch analysis and third, an extra step, step 13, which checked the unidimensionality of the Rasch analysis from a principal components analysis (PCA) of item residuals. Steps 1 to 12 were outlined by Iramaneerat, Smith, and Smith (2008). Iramaneerat, Smith, and Smith asserted that there is no standard procedural order in which to review the results of a Rasch model analysis because the order depends on the particular research objectives. Moreover, the findings produced in one step can require cycling through previous steps. Thus, the following steps are a general recommendation for analyzing polytomous Rasch model datasets. The purpose of outlining these steps is to document the evidence of the unidimensional requirements of the polytomous Rasch model (Iramaneerat, Smith, & Smith, 2008).

The thirteen steps utilized in this study are outlined in sequential order. The steps are organized from the three sourced groupings of dichotomous Rasch analyses, polytomous Rasch analyses and principal components analysis (PCA) of item residuals.

Dichotomous Rasch Analysis

The first four steps are based on the case for the most fundamental form of the Rasch model, the dichotomous case. Each of these four steps is described in more detail sequentially.

The first step involves checking the internal consistency reliability of both person ability measures and item difficulty calibrations. The separation reliability might not be the

best index to report as it "is on a nonlinear scale and suffers from ceiling effects in the same manner as KR-20 and coefficient alpha" (Iramaneerat, Smith, & Smith, 2008). Therefore, a separation index or strata index should ideally also be reported. The separation index ranges from 0 to infinity and is on a ratio scale. In contrast, the strata index displays the amount of levels at distinct statistical groups.

The second step involves checking the Wright Map for targeting. The Wright map displays distributions of person ability measures on the left and item difficulty calibrations on the right and they are both located along a common latent construct variable scale. The map is scaled into interval-level increments called logits, short for log odd units. The logits can be visualized along a vertical continuum in the center of the Wright map and forms a measuring stick-like ruler from which comparisons between items and persons can be made. The scale is divided into interval-level increments called logits, short for log odd units. The central point of the scale is centered at zero logits and aligns with the mean of the item difficulty calibration distribution. The check of targeting in the Wright Map involves visualizing and comparing the means of the person ability measure distribution and the item difficulty calibration distribution. If the mean of the item difficulty calibration distribution is located vertically higher than the mean of the person ability measure distribution, it can be determined that the items are relatively difficult for the persons. If the mean of the item difficulty calibration distribution is located vertically lower than the mean of the person ability measure distribution, it can be determined that the items are relatively easy for the persons. If the mean of the item difficulty calibration distribution is located at the same vertical location as the mean of the person ability measure distribution, it can be determined

that the items are relatively well targeted for the persons. The previous explanation of standard Wright Maps holds for the case of dichotomous Rasch model analyses, which have only one threshold per item. However, in the case of polytomous Rasch models the visualization is more complex and a standard Wright Map does not capture all the relevant information. In the case of polytomous Rasch model analyses, alternative figures such as an Andrich thresholds map (Linacre, 2015b, Winsteps Table 2.14) or a Wright Map with person and item distributions and polytomous item range can be used.

The third step involves checking the item analysis report. The items analysis report displays comprehensive information about the items in a table with 13 columns of data. The columns display information including item raw score, count, logit calibration (labeled measure), standard error, several types of fit statistics and point-biserial correlations. The easiest and most difficult items can be determined quickly from the logit calibrations (measure) and standard errors. Also, quality control statistics assessing the fit of the items to the Rasch model can be checked. "Depending on the fit criteria defined by the researcher, misfitting items should be considered for deletion to eliminate noise or data redundancy from the analysis" (Iramaneerat, Smith, & Smith, 2008, p. 59). The item fit criteria defined in this study was underfit, that is, the noisy off-dimensional misfit greater than Outfit and Infit MNSQ of 1.50 (Linacre, 2002; Boone, Staver, & Yale, 2014). An additional check not mentioned by Iramaneerat, Smith, and Smith (2008) is the inspection of the item analysis report for items with extreme scores. An item with an extreme score can be of two types, first, an item with a maximum score, which means that of the persons who responded to that particular item, everyone scored the maximum perfect score on that item; in other

words, the item was very easy. Second, in contrast is an extreme item with a minimum score, which means that of the persons who responded to that item, everyone scored the minimum zero score on that item; in other words, the item was very difficult.

The fourth step involves checking the person analysis report. The person analysis report displays comprehensive information about the persons in a table with 13 columns of data; the table column layout is basically the same as the item analysis report. The person with the lowest and highest ability can be checked from the measure column. Also, quality control statistics assessing the fit of the items to the Rasch model can be checked. The person fit criteria defined in this study was the same as that defined for item underfit, that is, the noisy off-dimensional misfit greater than Outfit and Infit MNSQ of 1.50 (Boone, Staver, & Yale, 2014; Linacre, 2002) targeted as the focus of the quality control process. An additional check not mentioned by Iramaneerat, Smith, and Smith (2008) is the inspection of the person analysis report for persons with extreme scores. A person with an extreme score can be of two types, first, a person with a maximum score, which means that of the items responded to by that particular person, they scored the maximum perfect score on those items; in other words, the person has a high ability. Second, in contrast is an extreme person with a minimum score, which means that of the items responded to by that particular person, they scored the minimum zero score on those items; in other words, the person has a low ability.

Polytomous Rasch Analysis

The next eight steps (Steps 5 to 12) are based on the case for an extended form of the Rasch model, the polytomous Rasch model. The polytomous Rasch model (PRM) as defined by Andrich (2010) includes both the Rating Scale Model (Andrich, 1978) and Partial Credit Model (Masters, 1982). Each of these eight steps is described in more detail sequentially.

The fifth step involves checking that all items are aligned with the latent trait construct. The items analysis report introduced in the third step is utilized to investigate the point biserial correlation coefficients. An item with a negative point biserial correlation coefficient indicates that particular item is not orientated with the latent trait construct. Sometimes surveys include items designed where the polarity of the scale should be reversed. In other words, high latent trait levels result in low ratings for these items. The solution to this problem is to rescore the item polarity to match the overall instrument orientation.

The sixth step involves checking the minimum of 10 observations per category. In the Winsteps (Linacre, 2015b) table, the summary of category structure table, the column labeled observed count should be checked. Within each rating category, there should be at least 10 observations. The purpose of having at least 10 observations per rating category is to calibrate accurate threshold parameters. If this guideline is not met, some solutions might require gathering extra data or collapsing categories.

The seventh step involves checking that observations are approximately evenly distributed over all categories. The observed counts column of the Winsteps table, the

summary of category structure table should be checked. If there is huge discrepancy, for example, one category is seldom selected and the next category is heavily selected, collapsing and rescoring categories might be a solution.

The eighth step involves checking that all categories are ordered. The observed average column of the Winsteps table, the summary of category structure table should be checked for monotonic advances of observed average measures. An upper category should result in higher logit category observed average measures in relation to a lower category.

The ninth step involves checking that the category outfit mean square values are under 2.0. The OUTFIT MNSQ column of the Winsteps table, the summary of category structure table should be checked. If a category's outfit mean square value is high this is a sign of unexpected responses in that particular category. The purpose of checking high outfit mean square values is that they are a threat to meaningful measurement.

The tenth step involves checking that all threshold values are ordered. The Andrich Thresholds column of the Winsteps table, summary of category structure, should be checked. As thresholds advance, the threshold values should increase. Thresholds are the crossover point between adjacent categories and can be visualized in the category characteristic curves in the probability figures under the Winsteps table, the summary of category structure table. Disordered thresholds, that is, thresholds that do not advance are detrimental to meaningful measurement (Andrich, 2013a, 2013b). A solution to the problem of disordered thresholds is to collapse and rescore categories.

The eleventh step involves checking that the thresholds advance by a positive amount. The Andrich Thresholds column of the Winsteps table, the summary of category

structure table should be checked. Previously, Linacre (2004) suggested a guideline that thresholds should advance by a minimum of 1.4 logits for a three category scale and 1.0 logits for a five category scale and these are the criteria followed by Iramaneerat, Smith, and Smith (2008). However, Linacre (2010) suggested an updated guideline that disordered thresholds can be usefully disordered between transitional categories; this new standard supersedes the Linacre (2004) guideline. However, for this study the guideline of Andrich (1978, 2013a, 2013b) requiring thresholds to be ordered by a positive value is adhered to.

The twelfth step involves checking that thresholds advance by a maximum of 5.0 logits (Linacre, 2004). The Andrich Thresholds column of the Winsteps table, summary of category structure, should be checked. The purpose for investigating this step is if thresholds are spaced too far apart, a decrease in precision can occur. A solution to this problem is to construct an extra category before the next administration of the instrument.

Research Question 2: Loanword Status

William Gosset's, also known as "Student" (1876-1937), greatest achievement was the invention of the *t* distribution and his creation of the *t*-test in 1908 (Sheskin, 2007). Gosset's advancement of the *t*-test started modern hypothesis testing theory (Sheskin, 2007). The *t*-test for two independent samples is a parametric inferential test applying either interval or ratio level data. The hypothesis assessed with the *t*-test is, are there different mean values for the two samples which represent two populations? Null and alternative (both non-directional and directional) hypothesis versions of the *t*-test exist. In the case of this study, the directional alternative hypothesis is chosen where the mean population that

group 1 (loanwords) represents is less than the mean of the population that group 2 (non-loanwords) represents. This directional alternative hypothesis is notated as, H_1: $\mu_1 < \mu_2$ and is assessed with a one-tailed test. The directional alternative hypothesis is only confirmed if the *t* sign is negative and the absolute value of *t* is greater or equal to the one-tailed critical *t* value at the a priori significance level (Sheskin, 2007). There are three assumptions: (a) the samples are selected randomly, (b) there is a normal distribution for the underlying population and (c) homogeneity of variance. Due to the iterative nature of Rasch model analyses through several pilot studies of constructing item difficulty calibrations that are located along the latent construct continuum, the item samples were not selected randomly, rather, they were selected so as to be located across a broad range of item difficulty. Therefore, the degree of meeting the random selection assumption was not high. Subsequently, only the other two assumptions of normal distribution and homogeneity of variance are checked.

Research Question 3: Word Frequency Bands

The analysis of variance, also referred to as ANOVA, is a family of parametric inferential statistical methodology developed by Ronald Fisher (Sheskin, 2007). The purpose of using an ANOVA is to assess if there is a difference between at least two means in a dataset.

The single-factor between-subjects analysis of variance, also referred to as a one-way independent ANOVA, is the most fundamental of the ANOVA procedures. Each of the sample means are used to estimate the population mean value that the sample represents.

The single-factor between-subjects analysis of variance applies either interval or ratio level data. It is based on three assumptions: (a) the samples have been chosen randomly from the representative population, (b) the population distribution from which the samples are sourced is normal, and (c) the homogeneity of variance, that is, the variances of the populations represented by the samples are equal (Sheskin, 2007). Due to the iterative nature of Rasch model analyses through several pilot studies of constructing item difficulty calibrations that are located along the latent construct continuum, the item samples were not selected randomly, rather, they were selected so as to be located across a broad range of item difficulty. Therefore, the degree of meeting the random selection assumption was not high. Subsequently, only the other two assumptions of normal distribution and homogeneity of variance are checked.

Planned comparisons, also called a priori comparisons or planned contrasts, are used when a researcher has predictions of differences between groups before data is collected. Therefore, following the omnibus ANOVA *F* test, a researcher can conduct any preplanned comparisons regardless of the tested *F* value being significant or not. When the predictor variable has three or more levels, some researchers do not conduct an initial ANOVA test. Two types of planned comparisons can be conducted, simple comparisons, also called pairwise comparisons and complex comparisons where two or more groups can be combined. In the pairwise comparisons case with for example three groups, A, B, and C, there are three possible comparisons, that is, A versus B, B versus C, and A versus C (Sheskin, 2007). Planned comparisons can also be defined in terms of orthogonal (uncorrelated or independent) and non-orthogonal comparisons. Choosing a comparison to

suit a hypothesis or hypotheses can lead the researcher to conduct non-orthogonal contrast tests (Baguley, 2012).

Research Question 4: Metasynonymy Awareness Hypothesis

A metasynonymy hypothesis spanning a six-tier hierarchy was proposed a priori. The six incremental degrees of ordered triple rank responses are ordered low to high respectively, synonym/decontextualized/null, decontextualized/synonym/null, synonym/null/decontextualized, null/synonymy/decontextualized, decontextualized/null/synonym, and null/decontextualized/synonym. These six developmental stages are based on a bifurcation of three amended stages from O'Connor's (1940) five original vocabulary acquisition developmental stages. In order to evaluate the ordering of the thresholds, that is, the adjacent borders between categories, the dataset was fit to a polytomous member of the Rasch model measurement family.

Rasch's (1960) dichotomous measurement model was extended by Andrich (1978) to include polytomously scored responses with a rating model for ordered response categories. From this point forward in this study, Andrich's (1978) model is referred to as the rating scale model. To be more specific, instead of requiring the categories to be ordered, in actual fact, the thresholds, that is, the crossover points between categories, are required to be empirically ordered. However, the ordering of thresholds might be marginal as indicated by Andrich (1996), giving an example where the distance between threshold one (-0.1) and threshold two (0.1) is an acceptable 0.2 logits. Thresholds should be described a priori to display at least one partial incremental stage between the two extremes

of a scale. Threshold discrimination is set to be the same across items. Parameters for item difficulty, person ability and threshold difficulty are estimated and quality control fit statistical tests are conducted as under the Rasch model (Rasch, 1960) family of measurement analyses. If there is evidence of empirical threshold disordering, categories can be rescored and then a subsequent analysis can be conducted.

Research Question 5: Changes in Person Measures Due to Scoring Guide

Research question 5 asked what changes happened to the person measures between the four scoring model analyses. The Wilcoxon matched-pairs signed-ranks test, a nonparametric statistical procedure, was used to evaluate the differences of pairs in the two dependent samples of each comparison (Sheskin, 2007). Each person tested had two interval measures, one generated from each scoring model. A difference score for each person pair was calculated. The directional alternative hypothesis was evaluated for each scoring model pair. The difference scores median was hypothesized a priori to be less than zero, that is, a negative number. In other words, the sum of ranks for the positive difference measures are expected to be less than the negative difference measures (Sheskin, 2007 and Field, 2009).

In order to keep the identity of the construct when comparing the same persons under different conditions, it is necessary to adjust the measurements (Wright & Masters, 1982). Putting the measurement estimates obtained from different conditions on the same playing field facilitates the investigation of objective comparisons.

Research Question 6: Guessing and Carelessness

Research question 6 asked to what degree guessing and carelessness are apparent in the type of analyses utilized. The Rasch model paradigm does not include item discrimination, guessability, or carelessness as parameters as is included in 2-, 3-, and 4-parameter logistic item response theory models.

Winsteps Rasch software can be used to calculate lower asymptote index initial approximations to detect guessing. It should be pointed out that these are initial approximations and not parameter estimates. In parallel, Winsteps Rasch software can also be used to calculate higher asymptote index initial approximations to detect carelessness. These lower and higher asymptote indices initial approximations are utilized in this study to assist in evaluation of guessing and carelessness (Linacre, 2013).

More specifically, Winsteps software can be utilized to display initial approximations of asymptotes, instead of parameters of asymptotes that would affect the item difficulty calibrations and person ability measures. Asymptotes are the lower and upper ends of an ogive curve, which usually approach a horizontal orientation. After completing an initial Winsteps analysis, within the "Specification" pull-down menu, the test developer can enter ASYMPTOTE=Yes. This procedure displays lower and upper asymptote initial approximations (not to be confused with parameters) for the first standard analysis and the subsequent cut lo and cut hi analyses (Linacre, 2013).

Further, to detect items that elicit lucky guessing and carelessness in person responses a comparison between the regular Rasch model analyses and the cutlo (guessing) and cuthi (carelessness) procedure was conducted using Winsteps (Linacre, 2015b). Cutlo

was set at -1 logits, that is, observations were eliminated when the person ability measure was 1 or more logits lower than the item difficulty calibration. In other words, the person has a 27% probability of successfully answering the item. In effect, this cuts off the responses where a low ability person takes a lucky guess at a difficult item. Cuthi was set at 2 logits, that is, observations were eliminated when the person ability measure was 2 or more logits higher than the item difficulty calibration. In other words, the person has an 88% probability of successfully answering the item. In effect, this cuts off the responses where a high ability person was careless on an easy item. Therefore a combined cut procedure with a cutlo of -1 logits and a cuthi of 2 logits was compared with the regular Rasch model analyses (Gershon, 1992a, 1992b, 1994).

Item Guessability

Item guessability refers to items where test-takers have taken part in lucky guessing. Lucky guessing typically occurs when a person of a low person ability measure encounters an item of high item difficulty calibration. Therefore, on a Rasch model Wright Map, the particular person is located on the lower end of the person ability measure distribution while the particular item is located on the upper end of the item difficulty calibration distribution. Low ability measure persons are likely to respond with a lucky guess to a high difficulty calibration item.

Bolded values in tables can be flagged as values with substantial lower asymptotes, c = 1 / (number of options) for dichotomous scoring (Linacre, 2013; Lord & Novick, 1968). Furthermore, extending this dichotomous recommendation to the polytomous scoring case,

the substantial lower asymptote critical values for the four analyses in this study are: quaternary 011223 scale, that is, RSM 0/11/22/3, c = 1/4 = 0.25*3 = 0.75 ; binary 000001 scale, that is, DRM 0/00/00/1, c = 1/4 = 0.25; ternary 001122 scale, that is, RSM 00/11/22, c = 1/3 = 0.33*2 = 0.66; and binary 000011 scale, that is, DRM 00/00/11, c = 1/3 = 0.33. A cutlo elimination of -1 was used to cut off responses where a person ability measure is 1 logit or more lower than the item difficulty calibration (Gershon, 1992b). In other words, the cutlo procedure helps identify a person that answered an item to some degree of success despite the Rasch model not expecting this response behavior.

Item Carelessness

Second, upper asymptotes related to item carelessness were examined. Carelessness occurs when a particular person of high person ability measure encounters a particular item of low item difficulty calibration. Sometimes, because of the perceived easiness of the item, the person might rush quickly with a response or take a few items at the beginning of the test to warm up and therefore make a careless mistake despite the items at the beginning of the test being not so difficult (Linacre, 2013).

The bolded values can be displayed in a table and are those with substantial upper asymptotes, d = 1 - c for dichotomous scales (Waller & Reise, 2010). Therefore, extending these critical values to the polytomous scales, the substantial lower asymptote critical values for the four analyses are respectively: quaternary 011223 scale, that is, RSM 0/11/22/3, d = (1 – .25)*3 = 2.25; binary 000001scale, that is, DRM 0/00/00/1, d = 1 - .17 = 0.83; ternary 001122 scale, that is, RSM 00/11/22, d = (1 - .33)*2 = 1.34; and binary

000011 scale, that is, DRM 00/00/11, d = 1 - .33 = 0.67. A cuthi elimination of 2 is used to cut off responses where a person ability measure is 2 logits or more higher than the item difficulty calibration (Gershon, 1992b). In other words, the cuthi procedure helps identify a person that answered an item to some degree of failure despite the Rasch model not expecting this response behavior.

CHAPTER 4

ITEM SPECIFICATION AND CONSTRUCTION PROCESS

In this chapter, the item specification and construction process utilized in this study is described in detail. Overall, this process is conceptualized as a heuristic technique whereby knowledge and evidence about new test items can be specified, developed, and documented. The rationale underlying this documentation is to provide evidence for construct validity of inferences and interpretations formed from the test.

There has been partial overlap between specific researchers regarding the conceptualization of the item specification and construction process. Three conceptual frameworks displaying this fractional parallelism include Davidson and Lynch's (2002) Testcraft Language Test Specifications, Wilson's (2005) Construct Modeling Approach, and Wolfe and Smith's (2007) Test Specifications (see Table 5).

Table 5

Item Specification and Construction Process Conceptual Framework

Stage	Davidson & Lynch (2002)	Wilson (2005)	Wolfe & Smith (2007)
1	General Description	Construct Maps	Construct Definition
2	Prompt Attributes	Items Design	Construct Models
3	Response Attributes	Outcome Space	Test Content Specification
4	Sample Item	Measurement Model	Scoring Model Creation
5	Specification Supplement		Scaling Model Selection

As displayed in Table 5, both Davidson and Lynch's (2002) and Wolfe and Smith's (2007) frameworks have five components while Wilson's (2005) framework has four

components. To alleviate redundancy and present the frameworks in a parsimonious structure, Davidson and Lynch's (2002) fourth component, sample item, is incorporated into other components. Further, for a parsimonious structure, Wolfe and Smith's (2007) first and second components, construct definition and construct models, are joined together into one encompassing component, construct definition/models. Despite the fact that the components have been rearranged slightly, the descriptions of the various components are intact. These amendments are displayed in Table 6.

Table 6

Reduced Item Specification and Construction Process Conceptual Framework

Stage	Davidson & Lynch (2002)	Wilson (2005)	Wolfe & Smith (2007)
1	General Description	Construct Maps	Construct Definition/Models
2	Prompt Attributes	Items Design	Test Content Specification
3	Response Attributes	Outcome Space	Scoring Model Creation
4	Specification Supplement	Measurement Model	Scaling Model Selection

The three conceptual frameworks are fundamentally similar structurally; however they differ in that Davidson and Lynch (2002) do not refer to Rasch model analyses, while Wilson (2005) and Wolfe and Smith (2007) do. Despite this difference, there is adequate overall structural agreement to consolidate the aspects into a unified conceptual framework of organization for this chapter.

Davidson and Lynch's (2002) Testcraft Language Test Specifications

Davidson and Lynch's (2002) components of Testcraft language test specifications were "adapted from Popham (1978, 1981) and his associates at the Instructional Objectives

Exchange (IOX)" (p. 20). Four of the five components presented by Davidson and Lynch (2002) are examined here. The one component left out of Davidson and Lynch's description is the Sample Item component. This component was omitted because several sample items were included within other components. Devoting a component's solitary focus to a sample item was deemed slightly excessive; therefore, sample items are provided as examples in other components.

The first component, The General Description, is a concise yet comprehensive outline of the construct (or the "what") to be tested. Davidson and Lynch's (2002) focus on criterion-referenced measurement facilitates a detailed analysis of the construct being tested. The construct description and definition should be detailed adequately so that external test writers can use it to write additional test items. Points to cover include the purpose, rationale, mandate, and theoretical background underlying the construct. This seminal foundation facilitated by the General Description component provides a basic blueprint for the subsequent components. Interaction between the subsequent three components provides valuable information for later revisions. Therefore, a sequential progression starting from Component 1 and progressing through to Component 4 is not necessary. For instance, in the case of this study's focus on receptive vocabulary, the purpose of the Vocabulary Size Test is to measure the written receptive English vocabulary size of native speakers and non-native speakers of English (Nation, 2012).

The second component, The Prompt Attributes, describes what is presented to the test-takers. For instance, in vocabulary testing this might be a multiple-choice item format based on a sentence stem's key word. The Prompt Attributes component, which answers

the question of how test-takers are prompted to provide an answer, should be linked back to the General Description Component outlined previously. The Prompt Attributes of the basic Vocabulary Size Test items are in a typical multiple-choice format of a sentence stem that contains a bolded key word followed by four options.

The third component, The Response Attributes, describes how the test-takers respond. The Response Attributes can interact with the previous (second) component, Prompt Attributes. The Response Attributes answers the question of what should occur when the test-takers respond. For instance, in vocabulary testing this might be that test-takers select a single best answer from three multiple-choice options.

The fourth component, The Specification Supplement, enables the description of any additional information without having to potentially overload the other three component descriptions. An example of a Specification Supplement for a vocabulary test is a list of all the key words that are assessed. In another case, an example of a Specification Supplement for a reading test is a list of the books from which the reading text passages were sampled. The Specification Supplement component is an optional component. It is only written when extra detailed issues are described.

Wilson's (2005) Construct Modeling Approach

Wilson's (2005) Construct Modeling Approach is categorized into four building blocks consisting of Construct Maps, Items Design, Outcome Space, and Measurement Model. Each of these four building blocks is described sequentially.

The first building block, Construct Maps, is visually represented by construct map figures. Construct maps are hierarchical sequenced concrete diagrams that are visualized representations of an underlying abstract construct, that is, a concept. Created a priori, construct maps facilitate the conceptualization of a theory (or theories) related to a given construct domain. The construct maps building block answers the question, what exactly is the concept being measured. This conceptualization facilitates research-based theoretical models about the test-takers' latent cognition manifested in diagrams. Construct maps are more exact than the idea of a construct because constructs, such as chemistry or biology ability, can have a rather basic form covering only one end of a continuum to another.

Wilson (2005) described a high-school chemistry curriculum (called the Living by Chemistry project) where the chemistry concept is divided into three strands: matter, change, and stability. Table 7 (adapted from Wilson, 2005, p. 8) displays the construct of matter only. Specifically, the definition of matter focuses on descriptions at the atomic and molecular levels. The developmental stages of the matter construct are categorized into two sub-strands: visualizing and measuring. In terms of success, high school students are expected to progress through the various levels starting at a novice level (Success Level 1: describing). At this success level, they can typically describe properties of matter (visualizing matter) and amounts of matter (measuring matter). As students develop through this mastery progression, they eventually reach the highest level of success displayed in Table 7 (Success Level 5: integrating) at the college or graduate level. At this success level, they would typically be able to integrate bonding and relative reactivity (visualizing matter) and integrate models and evidence (measuring matter). The Living by

Chemistry matter strand construct map as displayed in Table 7 is comprehensive, yet it is untested at the higher success levels.

When construct maps are first conceived, they might not be as fully developed as the one displayed in Table 7. This is not problematic, as construct maps are a foundational step to be taken before construct validation can take place. On the other hand, construct maps facilitate qualitative levels that can describe the incremental stages between a

Table 7

Construct Map for the Matter Strand of the Living by Chemistry Project

Levels of success	Atomic and molecular views	Measuring and model refinement
5: Integrating	Bonding and relative reactivity	Models and evidence
4: Predicting	Phase and composition	Limitations of models
3: Relating	Properties and atomic views	Measured amounts of matter
2: Representing	matter with chemical symbols	Mass with a particular view
1: Describing	Properties of matter	Amounts of matter
	A: Visualizing matter	B: Measuring matter

construct continuum's two far ends. An example of this is displayed in Table 7 where the matter strand of chemistry is compartmentalized vertically into five hierarchical levels of success. Also, the construct map in Table 7 is further compartmentalized horizontally into two sub-strands, atomic and molecular views (realized through visualizing matter) and measuring and model refinement (realized through measuring matter).

Construct maps can be refined through a four-step process. The first step uses concrete construct maps to describe the abstract construct to colleagues. The second step involves developing items that elicit responses confirming the construct maps levels. The third step requires pilot testing these items with a small but representative sample of test-

takers. The fourth step analyzes the data to determine whether the results confirm the initial aspirations displayed in the construct maps.

The second building block, Items Design, can be differentiated from the first building block, which is based on the underlying abstract construct. In contrast, the items design focuses on the concretely manifested items themselves; thus the question answered here is how to prompt responses from test-takers. Given enough time and development, the test designer perceives the items as representations of the underlying latent construct. For instance, an items design of a question from the Vocabulary Size Test (Nation & Beglar, 2007) consolidates a sentence stem containing a boldfaced key word and four options in a typically formatted multiple-choice item design. The following Vocabulary Size Test example is from the 5,000 most frequent word level of the British National Corpus (Nation & Beglar, 2007, p. 11).

1. miniature: It is a miniature.
    a. a very small thing of its kind
    b. an instrument for looking at very small objects
    c. a very small living creature
    d. a small line to join letters in handwriting

Test designers need to focus on a test item design that covers a wide range of knowledge. Interpretations providing criterion-referenced insights from low to high across the construct are essential. Think aloud and exit protocol interviews with responders can provide valuable feedback during the item trialing process.

The third building block, Outcome Space, concerns how to classify observations and score them according to the underlying construct. For instance, observations can be

examined for the Vocabulary Size Test item example 1 outlined in the previous building block, Items Design, targeting the key word *miniature*. Nation and Beglar (2007) used a dichotomous scoring model for item analysis, that is, items were binary scored 1 point for a correct response and binary scored 0 points for an incorrect response. Thus, for the previously discussed example item 1 (in the second building block, Items Design) with the boldfaced key word of *miniature*, the observed response, a) *a very small thing of its kind*, is classified as correct and scored 1. Further, the three other possible observed responses, b) *an instrument for looking at very small objects*, c) *a very small living creature* and d) *a small line to join letters in handwriting*, are all classified as incorrect and scored 0.

Finite classifications of a clearly defined, ordered, theoretically based set of categories are essential for a concise definition. The previous building block, Items Design, and this building block, Outcome Space, are often inseparable, especially in multiple-choice test format designs. As a result, in the case of multiple-choice test item development they are typically conflated together.

The fourth building block, Measurement Model, details a procedure to link all the four building blocks together, as the Items Design and Outcome Space are linked back to the Construct Maps through the Measurement Model. The advantages of using Wright Maps are that they potentially display person and item distributions, persons-items targeting, person predictive validity, item construct validity, person and item inference (Linacre, 2012). This building block answers the question of which measurement model to employ. Fundamentally, the answer to this question depends on an individual's definition of the term measurement. To simplify the definition, a choice between two approaches to

measurement is presented. One approach focuses on items and is represented by the Rasch model paradigm. The other approach, which focuses on scores and the instrument, is represented by classical test theory and true score theory. The approach chosen by Wilson (2005) is the Rasch model paradigm. The visualization of the Measurement Model when using the Rasch model paradigm is facilitated by the use of a Wright Map in Figure 1.

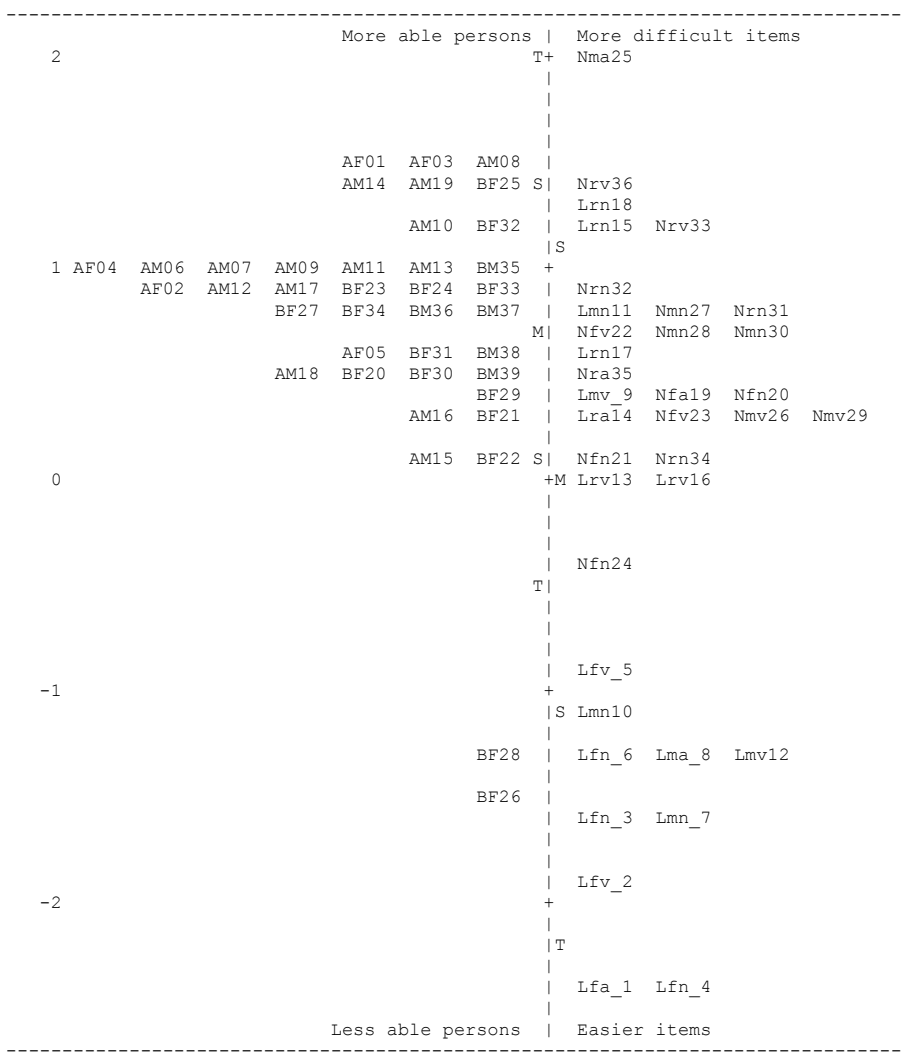


Figure 1. Wright map example.

Figure 1 displays a Wright Map in which both the person ability measure hierarchy on the left side and the item difficulty calibration hierarchy on the right side are placed on a central common metric scale measured in logits (log odds units).

The final Wright Map can be compared back to the initial Construct Map to visualize how a posteriori empirical results displayed in a Wright Map and a priori theoretical hypotheses displayed in Construct Maps are similar or different. This circular completion of the four building block cycle ideally enables the construction of a fundamentally robust test.

Wolfe and Smith's (2007) Test Specifications

Wolfe and Smith (2007) describe a conceptual framework calling for evidence-based documentation towards the goal of instrument measure validation. They outline four test specification categories that are categorized into Construct Definition/Models, Test Content Specification, Scoring Model Creation, and Scaling Model Selection. Each of these four test specification categories is described sequentially.

The first test specification category, Construct Definition/Models, is a conflation of the two separate categories, Construct Definition and Construct Models, originally outlined by Wolfe and Smith (2007). For the conceptual framework utilized in this study, it made conceptual sense to consolidate these two construct related ideas together into one cohesive whole.

The first test specification category, Construct Definitions/Models, first, defines the construct concisely and specifically. Second, this category connects the construct to

relevant theory-based models in the research literature. The construct definition process enables test developers to clarify the intended construct, display how it differs from other constructs and how it can involve relationships with other constructs. A domain analysis is an essential part of the construct definition. Construct models are a crucial part of this process and outline the construct under a theoretical perspective. Examples of construct models include internal, external, developmental, and processing models. The models utilized in this study, that is, Palmer's (1917) amended word frequency models and O'Connor's (1940) amended synonymy acquisition model are examples of developmental models under this classification.

The second test specification category, Test Content Specification, requires making a large pool of indicators that are relevant to the construct. The ultimate goal of making a test blueprint should be preceded by test developers deciding on three prior issues related to the test item format: Test developers should specify a universe of potential indicators, specify a content and process sampling plan, and select the item format.

The third test specification category, Scoring Model Creation, involves deciding how to utilize raw scores. A decision must be made to either score these codes dichotomously or polytomously at a numeric level. The resulting codes at a numeric level are on an ordinal scale of measurement. The codes are on an ordinal scale because they can be ordered, but the intervals between numeric codes might not be equal. This definition comes from Stevens' (1946/2006) theory of measurement scales, which was classified into four scale levels: nominal, ordinal, interval and ratio. These four broad scale levels are hierarchical where the nominal scale level is conceptualized as the most basic or lowest

level and ratio scale level is at the highest level. The definition of measurement is according to a rule where numerals can be assigned. The nominal scale is the broadest as it allows the most open assignment of numerals. An example of this scale is the categorization of gender where men are classed into a group, notated as 0, and women are classed into a group, notated as 1. Next, the ordinal scale facilitates the ordering of values. An example of this scale is the hardness of minerals. Then, crossing over into quantitative measuring, the interval scale includes standardized metric units between values and an arbitrary zero on the scale's continuum. An example of this scale is the degrees Celsius temperature scale. Finally, the ratio scale includes an absolute zero on the scale. An example of this scale is the degrees Kelvin temperature scale with an absolute zero set at 0 degrees Kelvin (-273 degrees Celsius) as a theoretical null point (Stevens, 1946/2006).

The fourth test specification category, Scaling Model Selection, describes the choice of a Rasch model from the Rasch model family. This category outlines the Rasch model process that consolidates and maps the ordinal level numeric codes onto a linear ruler that represents the incremental measures of the intended construct. At the basic level, these measures represent parameters of both item difficulty and person ability. Using one of the members of the family of Rasch models available, both the item and person parameters can be mapped onto the latent construct in the form of a visual Wright Map. The advantages of displaying the results in a Wright Map are many. Wright Maps provide a clear representation of the distributions of persons and items, targeting of persons with items, predictive validity of persons, construct validity of items, and inference of persons and items (Linacre, 2012).

Commonly applied versions of Rasch models include the Dichotomous Rasch Model (DRM) developed by Rasch (1960), the Rating Scale Model (RSM) developed by Andrich (1978), the Partial Credit Model (PCM) developed by Masters (1982), and the Multifaceted Rasch Model (MFRM) developed by Linacre (1989).

Conceptual Framework Rationale

To summarize this item specification and construction process conceptual framework, there is considerable research overlap between the three approaches described by Davidson and Lynch (2002), Wilson (2005), and Wolfe and Smith (2007). Wilson's (2005) construct modeling approach employing the four building block categorization is adopted in this study because of the clarity and foundational stability of using four distinct building blocks. Further, Wilson's (2005) construct modeling approach facilitates displaying clear graphical visualizations representing the key aspects of both the initial and final building blocks. The use of an initial map and a final map is important, as it facilitates a comparison of the theoretical beginning and the empirical finish. However, in a sense it is not an end, as this process also assists in the linking of the present iteration to initiate the next iteration of the process. Therefore, measurement develops and progresses. In the seminal construct maps stage, theoretically based concrete construct maps represent the abstract latent construct, while in the final measurement models stage, empirically based concrete Wright maps and Category Characteristic Curve maps represent a polytomously scored abstract latent construct. Figure 2 displays an example of a Category Characteristic

Curve map that has three category curves and two thresholds, that is, crossover points where two adjacent category curved lines meet.

One of the advantages of employing user-friendly visual maps is that such figures facilitate the item specification and construction process used in this study to be communicated to a wider audience. This advantage can assist in one of the long-term

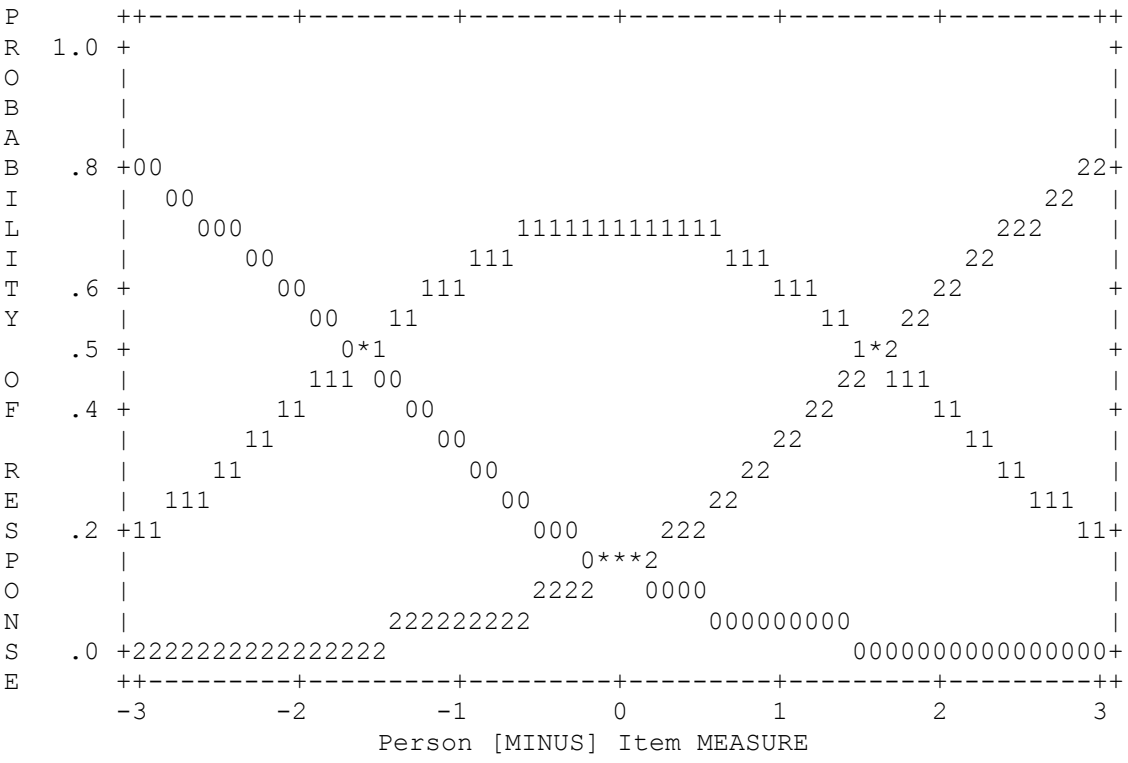


Figure 2. Category characteristic curve map example.

goals of this study, which is to communicate the benefits of Rasch model measurement analyses to as large an audience as possible. However, where appropriate, aspects are adopted from David and Lynch's (2002) and Wolfe and Smith's (2007) conceptual

classifications. More specifically, sample items were included from Davidson and Lynch's classification. Wolfe and Smith's classification provided the incentive to include two developmental models, Palmer's (1917) word frequency theory and O'Connor's (1940) synonymy acquisition theory. Additionally, Wolfe and Smith's framework provided the idea to include a test blueprint figure visualizing the main domains in the study.

Construct Maps

Vertically orientated construct maps are typically presented as a hypothesized a priori hierarchical order of the item difficulty calibration parameter and/or the person ability measure parameter. For instance, when vocabulary size is being measured in this study, a vertically orientated construct map can be used to conceptualize the a priori inter-(between) item difficulty calibration hierarchy as illustrated in Table 8.

Table 8

A Priori Inter-Item Difficulty Parameter Vertical Construct Map

Difficulty hierarchy	Loanword status	Word frequency
Difficult	Non-loanwords	Low frequency
Medium		Medium frequency
Easy	Loanwords	High frequency

Both the loanword status and word frequency are theoretically based on Palmer's (1917) word frequency theory. Keywords were categorized on an incremental continuum from high to low frequency. Palmer (1917) stated, "the more frequently used words will be the more easily learnt" (p. 123). An amendment to Palmer's word frequency theory is

proposed in this study. English-sourced loanwords into the Japanese language are more frequently used than regular English words (non-loanwords) in the Japanese English as a foreign language context. Therefore, in the Japanese context, English-sourced loanwords are easier to learn than English-sourced non-loanwords. The number of foreign loanwords used in Japan can be contextualized by Tomoda's (2009) volume, *The Foreign Word Tsunami*. Especially, in the post-World War II era, there was huge influx of foreign loanwords into the Japanese society and lexicon.

Modern frequency research conducted since Palmer (1917) includes that of Ellis (2002), who investigated frequency effects in language processing. Ellis (2002) asserted that input frequency is closely connected to language processing. Frequency is an essential key for language learning because patterns emerge from learning life spans of analysis due to the distributional aspects of language input. As learners advance, language comprehension is governed by the implicit processing of statistical data of the input's characteristics. Language comprehension depends on the ability to discern the most likely syntactic and semantic analyses of new words based on a foundation of frequencies of prior word analyses. Therefore, when people consciously use conversations to communicate, they unconsciously gain knowledge of language aspect frequencies (Ellis, 2002).

As learners' counts of language features are unconscious, the construction of knowledge into categories is sometimes ambiguous. Research into learner categorization ability is one of the largest research fields in cognitive science. However, one unanswered question is whether counting is scaled in token (event) or type (individual) units.

Traditionally, researchers have focused on token events; however, recently there has been a surge of research interest in type frequency (Ellis, 2002).

Language learners discern statistical and distributional language aspects (similar to many corpus linguistic studies) cognitively, as their comprehension is determined by this lexical knowledge. These statistical and distributional analyses require gradual learning and practice over tens of thousands of hours. Language knowledge is due to a great collection of prior encounters. "Learners analyze input that they are exposed to; practice makes perfect" (Ellis, 2002, p. 178).

However, frequency is not the sole determinant of learning; language learning is complex, therefore, multiple variables make the prediction of outcomes difficult to determine. The counterpoint of Ellis' (2002) frequency hypothesis is illustrated by Gass and Mackey (2002). Gass and Mackey (2002) propose that although frequency is an important factor in second language acquisition but it is not the only factor to the extent that if input is not frequent acquisition is not affected. In terms of this present study, polysemous words of high frequency have context-dependent meanings, therefore, they are not easy to acquire.

The role of frequency in language processing and acquisition has been ignored in mainstream research for 40 years (Ellis, 2002). It is time for a revival of its salience in research. After all, "it is the language learner who counts" (Ellis, 2002, p. 179).

Another example of modern research into frequency effects is Gries (2008) investigation into corpus methods and second language acquisition. The most fundamental tool of corpus-linguistics is the word frequency list. Word frequency lists are useful in

pedagogical settings because frequency lists are constructed with the underlying rationale that it is better for second language learners to study words that are important. In the corpus linguistics research field, more important is equivalent to more frequent. Additionally, in cognitive linguistics, frequency lists are theoretically important because the more frequent a word, the more entrenched it is believed to be (Gries, 2008).

As a result of frequency's role, the most direct use of frequency lists in second language acquisition is in curriculum development (Gries, 2008). That is, frequency lists are used in sequencing specific words in a particular order in textbooks, classroom activities, homework and examinations. Further, word frequency lists are used "to quantify and/or compare the attainment of language proficiency" (Gries, 2008, 415).

The additional focus in this polytomously scored study of vocabulary extent is receptive vocabulary strength (depth), that is, as conceptualized by the a priori intra-item difficulty calibration hierarchy of ranked item options. O'Connor (1940) originally proposed a five-stage synonymy acquisition theory that outlined the form-meaning mistakes a vocabulary learner develops through on their way to acquiring the meaning of new words. At the first stage, the novice vocabulary learner associates the meaning of a word heard with a words of similar sound. At the second stage, the learner associates the meaning of a word written with a word of similar written form. At the third stage, the learner associates the meaning of a word with a word in a different context. In the fourth stage, the learner associated the meaning of a word with an antonym. Finally, in the fifth stage, the learner associates the meaning of a word with a synonym. Thus, O'Connor's (1940) five synonymy acquisition stages can be described as sound like, look like, context,

antonym and synonym. First language learners of English are expected to follow these stages of vocabulary acquisition on the road to synonym mastery. Examples of some of the five options used in O'Connor (1940) are used to construct an example item. The example multiple-choice item consists of a sentence stem with a boldfaced key word noun followed by five options. The test-takers are instructed to select the single best answer (SBA) from the five available options.

She was in a **depression** for 3 months.

depression: a) decry b) dent c) sadness d) elation e) machine

Novice vocabulary learners of English as a native language in the first vocabulary acquisition stage would confuse the key word *depression* with a similar sounding word such as *press*. Thus, they would select the synonym of *press* which is option e) machine. In the second vocabulary acquisition stage, slightly higher ability vocabulary learners would confuse the key word *depression* with the written word such as *depreciation*. Thus, they would select the synonym of *depreciation* which is option a) decry. In the third vocabulary acquisition stage, other slightly higher ability vocabulary learners would confuse the key word *depression* in a different context. For example, they would think that it was the same as placing the key word *depression* into the following (inappropriate) sentence.

There is a **depression** in the car door where the rock hit it.

Therefore, they would select option b) dent. In the fourth stage, vocabulary learners would choose the antonym of *depression*, d) elation. In the fifth and final stage, a vocabulary learner of superior ability would select the synonym of *depression* as c) sadness in the original item sentence stem.

However, numerous pre-pilot study results indicated that O'Connor's five stages were too numerous to be operationalized as five categories. This finding of disordered thresholds was corroborated by Smith's (1987) results of vocabulary depth using O'Connor's (1940) five stages. In Smith's (1987) study, the use of five categories resulted in disordered thresholds. According to Andrich's (2013a, 2013b) recommendations, disordered thresholds are a violation of statistical sufficiency and meaningful measurement. Furthermore, there is no need for five theoretical stages because words such as high frequency concrete nouns can be learned efficiently without passing through five stages.

Furthermore, it is difficult for one acquisition theory to account for all word parts of speech, for example, nouns, verbs and adjectives. Miller and Fellbaum's (1991) study of semantic networks, primarily in terms of synonymy, examined a relational approach between words. They concluded that nouns are structurally organized into hierarchical levels. Nouns are usually represented as having a generic level at the top of the hierarchy moving down to a specific level at the bottom of the hierarchy. Miller and Fellbaum (1991) stated that the relations between adjectives were found to be different to those between nouns. The primary semantic relation between adjectives is antonymy. Uniquely, adjectives can be used to describe opposite ends of a continuum for value attributes. Verbs are probably the most salient language lexical category (Miller & Fellbaum, 1991) because all

English sentences have at least one verb. However, in English there are much fewer verbs than nouns. The verb lexicon is mostly organized according to troponymy, that is, manner or fashion, of verb hyponymy. Verb classification hierarchies have less number of levels than noun hierarchies. Typically, in verb hierarchies one verb lexical level is denser than the other verb levels. Thus, a single synonymy acquisition theory cannot apply to all individual parts of speech, such as nouns, adjectives and verbs.

Therefore, in this present study only three stages were included from O'Connor's (1940) original five-stage synonymy acquisition theory. The first stage, labeled sound like/look like, is a consolidation of the original sound like and look like stages. The second stage, labeled context, is the context stage. Finally, the third stage, labeled synonym, is the sense synonym stage. However, each of these three stages is bifurcated, that is, split, into an upper and a lower division. The rationale underlying this splitting of stages into halves is an attempt to measure more of the available response data information. As an example, test-takers encounter a typical multiple-choice item sentence stem with a boldfaced key word followed by three options. Instead of selecting a single best answer, as instructed in typical multiple-choice tests, test-takers are instructed to rank order all three options according to a rule. The rule is a form-meaning match (ordered left to right) in: no sentences, another imagined sentence, and this written sentence. An example item follows.

The sun is very **bright** today.

a) short b) radiant c) intelligent

The metasynonymy awareness hypothesis predicts that novice vocabulary learners respond with an ordered triple rank that is incorrect, b) radiant < c) intelligent < c) short, that is, at the sound like/look like lower level. In this case, the test-takers confused the key word *bright* for the similar looking and sounding word *brief.* A synonym of the word *brief* is *short*. A slightly higher ability vocabulary learner responds with an ordered triple rank, c) intelligent < b) radiant < c) short, that is, at the sound like/look like upper level. Slightly higher ability vocabulary learners respond with an ordered triple rank, b) radiant < a) short < c) intelligent, that is, at the context lower level. Slightly higher ability vocabulary learners respond with an ordered triple rank, a) short < b) radiant < c) intelligent, that is, at the context upper level. Slightly higher ability vocabulary learners respond with an ordered triple rank, c) intelligent < a) short < b) radiant, that is, at the synonym lower level. Finally, a superior ability vocabulary learner responds with an ordered triple rank, a) short < c) intelligent < b) radiant, that is, at the synonym upper level. The highest level vocabulary learners identified option a), short, as incorrect in any sentence, option c), intelligent, as possible in a different contextualized sentence for instance, *she is a bright student*, and option b), radiant, as the sense synonym in the written sentence stem.

A horizontally orientated construct map of the threshold difficulty parameter hierarchy is displayed in Table 9. Observing Table 9 from left to right, the a priori horizontal construct map proposes a threshold difficulty parameter hierarchical order of sound- or look-like (null synforms) being divided into two subdivisions (lower and upper), two context subdivisions (lower and upper) and two synonym subdivisions (lower and upper).

Table 9

A Priori Intra-Item Difficulty Parameter Horizontal Construct Map

Construct Labels	S-/L-Like Lower	S-/L-Like Upper	Context Lower	Context Upper	Synonym Lower	Synonym Upper
Category	0	1	2	3	4	5
Correctness	Incorrect	Partial 1	Partial 2	Partial 3	Partial 4	Correct
Difficulty	Easy					Difficult
Threshold	1	2	3	4	5	

Note. S-/L-Like = Sound-/Look-Like Synform Null Option.

These vertically orientated and horizontally orientated construct maps can be visualized on a single academic vocabulary extent (size and strength) conceptual framework as in Table 10. Joining both of these aspects into one vocabulary extent scale consolidates the conceptualization of the academic vocabulary construct into a unidimensional hierarchical order. This approach also helps with conceptualizing the rationale underlying three of the six research questions selected to guide this study. The macro-analysis level loanword aspect is addressed by research question 2. The meso-

Table 10

Academic Vocabulary Extent (Size and Strength) Construct Framework

Item type	Analysis level	Construct theory	System granularity	Vocabulary capacity	Focal point	M. model evidence
Inter-	Macro	Loanword status	Very coarse	Size	Keyword	Wright map
(Between)	Meso	Word frequency	Coarse	Size	Keyword	Wright map
Intra-	Micro	Synonymy	Fine	Strength	Option	CCC map
(Within)	Nano	Syn: Bifurcated	Very fine	Strength	Option	CCC map

Note. Syn = Synonymy; M. model evidence = Measurement Model Evidence; CCC map = Category characteristic curve map.

analysis level word frequency aspect is addressed by research question 3. The micro- and nano-analysis level synonymy aspects are addressed by research question 4.

Items Design/Outcome Space

Wilson (2005) detailed a special case for multiple-choice items and Likert-type items. The differentiation between the items design and outcome space for these items is often ambiguous; therefore, the items design and outcome space building blocks are often conflated. Typically, the responses that are categorized by the respondents and the scores assigned to the categories have been determined beforehand. Therefore, the items design and outcome space building blocks are consolidated in this section.

In this study, the Items Design/Outcome Space is organized into a sequence of seven stages in the protocol as displayed in Table 11. The seven-stage protocol in Table 11 is structured to answer seven questions relevant to the structure of the combined Items Design/Outcome Space building block. Procedure 1, preliminaries, asks and answers the question, what is the initial description of the items design and outcome space? Procedure 2, options arrangement, asks and answers the question, how many written option arrangements should be prepared for the test? Procedure 3, permissible answers, asks and answers the question, how many answers are permissible for each item? Procedure 4, answer key, asks and answers the question, what is the correct answer key for each individual item? Procedure 5, anticipated responses, asks and answers the question, what are the anticipated person response patterns to each item? Procedure 6,

Table 11

Items Design/Outcome Space Building Blocks Protocol

Building block	Procedure	Action	Method	Focal point
A. Items design	1. Preliminaries	Define	Describe	Test developer
	2. Options arrangement	Enumerate	Count	Test developer
	3. Permissible answers	Enumerate	Count	Test-taker
	4. Answer key	Allocate	Arrange	Test developer
B. Outcome space	5. Anticipated responses	Predict	Create	Test-taker
	6. Scoring guide	Calculate	Categorize	Test developer
	7. Data compression	Transform	Consolidate	Test developer

scoring guide, asks and answers the question, which score should be assigned to each response to individual items? Procedure 7, data compression, asks and answers the question, how to compress the dataset usefully for applying it to rectangular Winsteps software control files? Each of these procedures is examined sequentially.

Preliminaries

Procedure 1, preliminaries, asks and answers the question, what is the initial description of the items design and outcome space? This study utilizes 36 typical multiple-choice item stems and three options per item in a paper question booklet. Each stem consists of a written English sentence where one of the words is a bolded key word. A test blueprint indicating the parts of speech for the 36 bolded key words in the test is displayed in Table 12. This ratio of vocabulary testing using three nouns to two verbs to one adjective was based on the proportions in the Lancaster-Oslo/Bergen (LOB) and Brown corpora (Johansson & Hofland, 1989; Hudson, 1994).

In this study, the keywords were selected from the Academic Word List (Coxhead, 2000). Initially, keywords were determined as having loanword status if they were included

in the listings of a katakana (Japanese loanword character) word dictionary published by Gakken (2007). Subsequently, two other katakana dictionaries published by Sanseido (2008) and Jiyu (2011) were used to confirm the keyword's loanword status. If the keyword was present in at least two of the three dictionary listings, then the keyword was classified

Table 12

Test Blueprint for the 36 Bolded Keywords in the Sentence Stems

Domain 1 loanword status	Domain 2 J8K word frequency band	Part of speech			
		Nouns	Verbs	Adjectives	Total
Non-loanword	Low	3	2	1	6
	Medium	3	2	1	6
	High	3	2	1	6
Loanword	Low	3	2	1	6
	Medium	3	2	1	6
	High	3	2	1	6
Total		18	12	6	36

Note. J8K = JACET 8000 academic word list.

as a loanword as indicated in Appendix A. Regarding representation of word coverage in katakana dictionaries, Gakken's contained 29,000 words, Sanseido's contained 55,200 words and Jiyu's contained 39,000 words.

The loanwords and non-loanwords were then stratified into three word frequency bands based on the JACET 8000 academic word list covering the most frequent 8,000 words occurring in a corpus of Japanese based educational contexts (Aizawa, et al., 2008). The rationale for using the JACET 8000 word list is that it is specific to the Japanese English as a Foreign Language (JEFL) learning context. Specifically, keywords listed in the

1,000 to 3,000 most frequent academic word levels were categorized as high frequency, keywords listed in the 4,000 to 5,000 most frequent academic word levels were categorized as medium frequency and keywords listed in the 6,000 to 8,000 most frequent academic word levels were categorized as low frequency as indicated in Appendix A. The rationale underlying these high, medium, and low frequency cut points was to allow a sufficient pool of available words in each hierarchical level to select keywords from. Overall, the Academic Word List (Coxhead, 2000) has a total of 570 words. Once these 570 words have been divided into loanwords and non-loanwords and further divided into JACET8000's eight levels of word frequency, the pools of available keywords are rather small and therefore impractical for test item selection. Thus, a decision was made to consolidate the potential keywords into JACET 8000 high, medium, and low word frequency level bands. These classifications are visualized in Appendix A.

Set Theory.

An introduction to the mathematical concept of set theory is fundamental to the core of this investigation. Set theory is a language that connects many diverse areas of mathematics together. The differentiation between a set and a list is essential to understanding this study. A set is a collection of unordered elements that is notated within braces, for example {c, b, a}. On the other hand, a list is a sequence of ordered elements that is notated within parentheses, for example (0, 1, 2). A list with 3 ordered elements can also be described as an ordered triple or a 3-tuple. The crucial differences between sets and lists are that sets are comprised of an unordered collection of elements and notated inside

braces. In contrast, lists are comprised of an ordered sequence of elements and are notated inside parentheses. Ordered lists containing 2 elements, that is, 2-tuples or ordered pairs, and lists containing 3 elements, that is, 3-tuples or ordered triples with no repetition of elements are utilized in this study (Breuer, 2006).

Within set theory, the terminology of a set can be classified into the definition of a finite set and an infinite set. Once again, set theory notation is the important key in distinguishing between the two types. A finite set is defined as a collection of a countable number of elements, for example {b, c, a} is a finite set of three unordered but countable elements. In contrast, an infinite set is defined as a collection of an uncountable number of elements, for example {0, 1, 2, 3, 4, …} is an infinite set in which the elements continue to infinity and therefore contain an uncountable number of elements. Only finite ordered lists are utilized in this study (Breuer, 2006).

Mathematics can be conceptualized as concerning two categories, discrete and continuous mathematics. In discrete mathematics, the data elements have separate, distinct, and countable values. Discrete mathematics examines countable objects that can be represented by integers, for example, there are 27 students in my Monday second period English Communication class. In other words, there are not 27.54 or 26¼ students in my class (Borowski & Borwein, 2006).

In contrast, within continuous mathematics the data elements have a gradual, that is, smooth, change in their variation. Continuous mathematics examines measurable real numbers, for example, the heights of my three best friends using a soundly calibrated measuring instrument. The heights of my three best friends are: Friend A is 181.26

centimeters, Friend B is 165.06 centimeters, and Friend C is 172.84 centimeters tall. In other words, the measuring units are not restricted to whole centimeters, that is, Friend A is 181 centimeters, Friend B is 165 centimeters, and Friend C is 173 centimeters tall. An extension of this line of thought is that the measuring units are not in blocks of tens centimeters, that is, Friend A is 180 centimeters, while both Friend B and C are 170 centimeters. Discrete mathematics is utilized in the initial sections of this chapter when raw scores are used for counting. However, continuous mathematics is utilized in the latter sections when mathematical transformations are used for measuring.

Arranging multiple-choice item options in a written a), b), c) list of left to right ordering is a convenient and traditional convention. However, on a more abstract level, the a), b), c) letter symbols represent the content of the latent construct under investigation. In this study's case, the construct's content for the three options are look-/sound-like, context, and synonym, in respective low to high hierarchical order (O'Connor, 1940, amended). More specifically and for the purposes of clarity, these three construct representations can be relabeled null, decontextualized, and synonym, in respective hierarchical order. Further, these three construct labels can be shortened to n, d, and s. These three construct labels form the three latent elements of a set that is presented to test-takers within each item. Figure 3 a) displays a visual representation of how options can be arranged for a particular item in a listed written order from left to right of synonym, decontextualized, and null, that is, s, d, and n.

The instructions for taking a typical multiple-choice test require the test-takers to select a single best answer. Therefore, the test-takers are actually instructed to create a one-

element subset of the three-element set they are presented with in the written options of the test item. In Figure 3 a), the location to move this chosen one-element response is represented by a narrow shelf with a solitary "1" (maximum score) displayed at the front. Under the dichotomous Rasch model (Rasch, 1960) measurement paradigm, a test developer would expect test-takers of various ability levels to respond in different ways to an individual item. High ability level test-takers are expected to respond with a correct answer of *synonym*, that is, s, as displayed in Figure 3 b). Middle ability level test-takers are expected to respond with a partially correct answer of *decontextualized*, that is, d, as

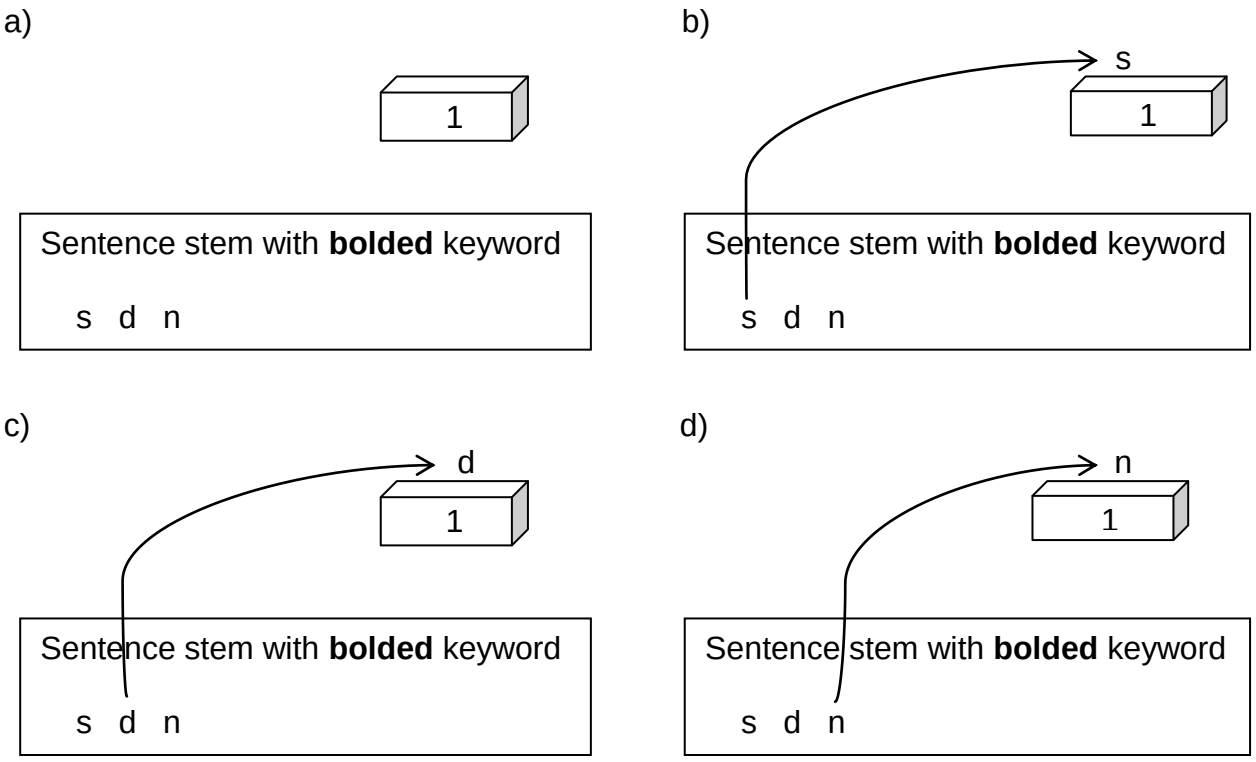


Figure 3. Three listed construct options with single best answer (SBA) instruction and responses.

displayed in Figure 3 c). Low ability level test-takers are expected to respond with an incorrect answer of *null*, that is, n, as displayed in Figure 3 d).

Furthermore, under the dichotomous Rasch model (Rasch, 1960) measurement paradigm, test developers would expect individual test-takers to respond differently to items of various difficulties. An individual, Person A, would be expected to respond with *synonym* to a very easy item such as that displayed in Figure 3 b). The same individual would be expected to respond with *decontextualized* to a medium difficulty level item such as that displayed in Figure 3 c). Finally, Person A, would be expected to respond with *null* to a very difficult item such as that displayed in Figure 3 d).

The instructions of the atypical multiple-choice test proposed in this study require the test-takers to rank all three options with no repetition of elements according to a synonymy correctness rule. Therefore, the test-takers are instructed to form a three-element ordered list (an ordered triple) arrangement of the randomly ordered triple they are presented with in the written options of the individual test item. As the ordered triples constructed by the test developer and the test-takers have the same cardinality (they both have three elements) and they share the same construct elements of null, decontextualized, and synonym with no repetitions of elements, they form a bijective function, that is, a bijection, which is used subsequently for further analyses. In Figure 4 a), the location to move the three-element response is represented by a relatively wider shelf with ordered spaces for the three elements indicated by the labels 0, 1, and 2 (from minimum to maximum score) displayed at the front. Under the polytomous Rasch model (Andrich, 1978) measurement paradigm, a test developer would expect test-takers of various ability

levels to respond in different ways to a single item. High ability level test-takers are expected to respond with a correct answer of (left to right) *null*, *decontextualized*, and *synonym*, that is, n, d, and s, as displayed in Figure 4 b). Middle ability level test-takers are expected to respond for example, with a partially correct answer of (left to right) *null*, *synonym*, and *decontextualized*, that is, n, s, d, as displayed in Figure 4 c). Low ability level test-takers are expected to respond with an incorrect answer of (left to right) *synonym*,

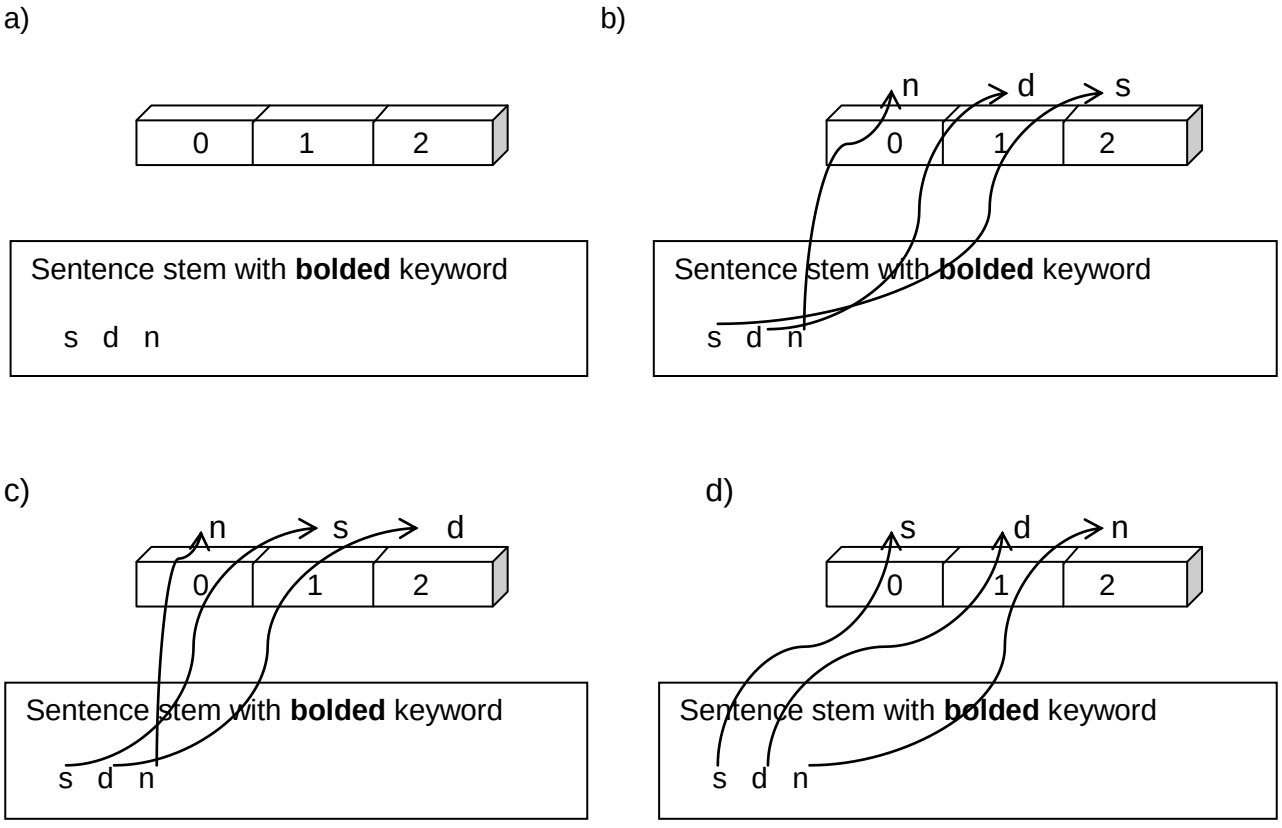


Figure 4. Three listed construct options with ordered triple rank (OTR) instruction and responses.

decontextualized and *null*, that is, s, d, n, as displayed in Figure 4 d). Note that this incorrect response, s, d, n, is the counterpoint to the correct response, n, d, s.

Furthermore, under the polytomous Rasch model (Andrich, 1978) measurement paradigm, a test developer would expect individual test-takers to respond differently to items of various difficulties. Person A would be expected to respond (left to right) with *null*, *decontextualized*, *synonym* to a very easy item such as that displayed in Figure 4 b). Also, the same individual, Person A, would be expected to respond with *null*, *synonym*, *decontextualized* to a medium difficulty level item such as that displayed in Figure 4 c). Finally, Person A, would be expected to respond with *synonym*, *decontextualized*, *null* to a very difficult item such as that displayed in Figure 4 d).

Options Arrangement

Procedure 2, options arrangement, asks and answers the question, how many written option arrangements with no repetition of elements should be prepared for the test? Within the mathematical field of set theory, enumerating 3-element arrangements of an ordered triple provides the answer to this question. Initially, the test developer starts with an ordered list (n, d, s) representing the construct option elements of *null*, *decontextualized* and *synonym*. Next, the test developer could use one arrangement of this, for example, the ordered triple (s, d, n) as the option arrangement for a particular item. To conceptualize the other possible arrangements of listings of options in the written left to right format, the test developer can perceive the other arrangements of all 3-element lists of the original (n, d, s)

list. Figure 5 displays a graphic representation of all six 3-element arrangements, that is, the ordered triples, *N1*, *N2*, *D1*, *D2*, *S1*, and *S2*. When ordering these six arrangements together, they can also be conceptualized as a sextuple, that is, a 6-tuple ordered list, ((n, d, s), (n, s, d), (d, n, s), (d, s, n), (s, n, d), and (s, d, n)).

(n, d, s)

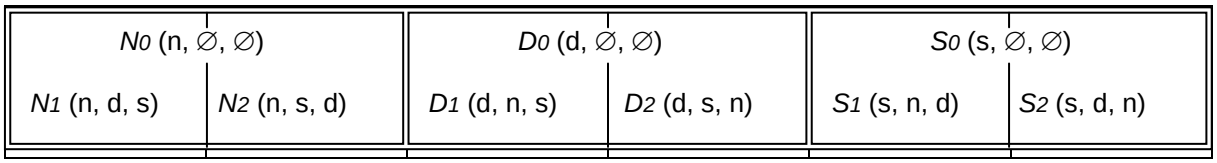


Figure 5. Ordered 3-element arrangements of the list (n, d, s) resulting in sextuples.

Therefore, the test developer should prepare six option arrangements for the test: *N1* (n, d, s), *N2* (n, s, d), *D1* (d, n, s), *D2* (d, s, n), *S1* (s, n, d) and *S2* (s, d, n). Note that the naming of these lists, for example, *N1* (n, d, s), starts with the element n, according to traditional left to right written convention.

Permissible Answers

Procedure 3, permissible answers, asks and answers the question, how many answers with no repetition of elements are permissible for each item? The answer to this question is assisted by the mathematical discipline of combinatorics.

Within the discrete mathematics discipline there is a branch of research called combinatorics. Within the combinatorics research field there is a mathematical concept

called permutations. Permutations are defined in this study as ordered arrangements of elements from a set where choice order is ruled as a requirement. More specifically for this study, the research field of enumerative permutations facilitates the investigation of counting the number of ways particular patterns can be formed from the arrangement of elements from a finite set (Borowski & Borwein, 2006).

Permutations are enumerated in this study using four methods. The first method uses a decision tree, which is a visualization or graphical method. The second method uses the principle of multiplication of choices. The third method uses a permutation formula, that is, when choices, *r*, are less than or equal to the total elements, *n*. The fourth method uses the permutation formula when choices, *r*, are equal to the total elements, *n*.

Decision Tree.

The following decision tree, Figure 6, illustrates the process that typical test-takers might make in completing the task of putting the three options into a synonymy rank-order. From left to right, test-takers are required to complete a form-meaning match by placing the letter representations a, b, and c for a sound-/look-like synform in no sentences (0); for a context in another imagined sentence (1) and for a synonym in this written sentence (2). As there are three options to choose from, this process is an example of a ternary, also called a trinary, rank-order. There is an instruction that repeating the letters a, b, and c is not allowed. In contrast to the western cultural tradition of reading and writing text from left to right, test-takers might typically start with choosing a letter (a, b, or c) from the far right side (the synonym form-meaning match for this written sentence). In other words, when

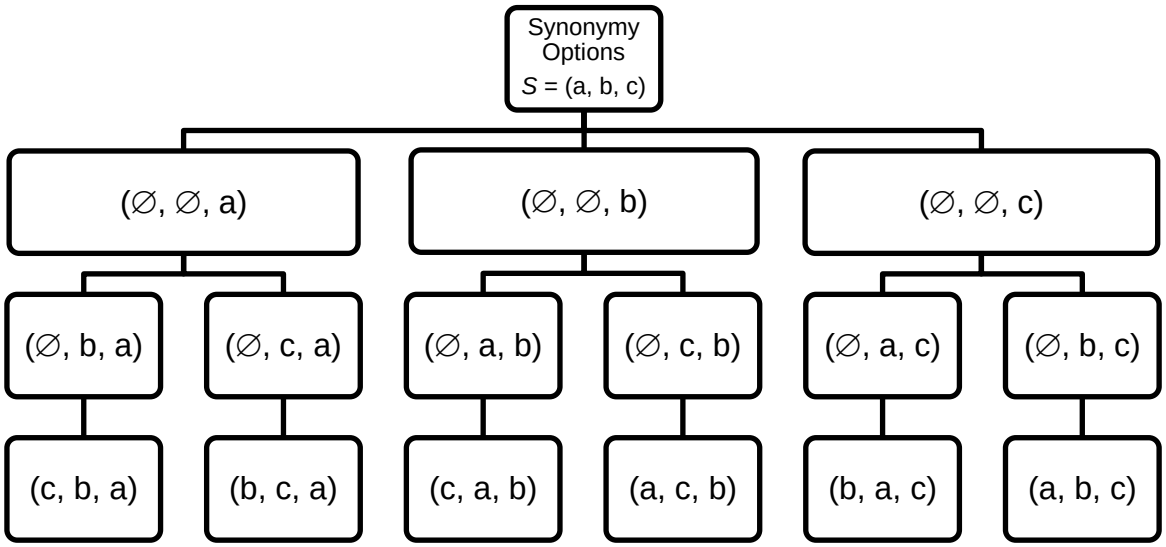


Figure 6. Decision tree for ordered triple responses to three options.

conceptualized as counted from the left, this is data column number 3. Then, they might move leftwards to the middle data column (data column number 2) for the context form-meaning match for another imagined sentence. Finally, they might move to the far left data column (data column number 1) for the sound-/look-like synform form-meaning match in no written or imagined sentences. The rationale underlying this three-stage sequencing is starting with the familiarity of conventional single best answer instruction in the right side column. Overall, the responders' decisions represent their responses to the synonymy rank-ordering of three multiple-choice options per item sentence stem. The decision tree in Figure 6 illustrates this decision choice process method graphically.

The top of this decision tree represents the root node or decision node. The root node in this case is the synonymy options represented by the ordered triple S = (a, b, c). The branches represent paths of choices that are made. Test-takers are instructed to construct a three-element ordered triple response without letter (i.e., a, b, c) repetitions. The construction of the response is hypothesized to involve three stages of decision making. At Stage 1, test-takers select from three alternatives, a, b, and c, to create an ordered list, (∅, ∅, a), (∅, ∅, b), or (∅, ∅, c) from the synonymy option set S = (a, b, c). Using set theory notation, ∅ denotes no elements, that is, an empty list ( ). Therefore, the cardinality, that is, the number of elements, of $|\varnothing| = 0$. However, a list (∅, ∅, a), has 3 elements within parentheses. Set theory notation also defines that $a \in S$, which means a is an element of the list S. Therefore, at Stage 1, the three available choices are notated as a, b, $c \in S$. Therefore, at Stage 1 the three permissible lists are notated as (∅, ∅, a), (∅, ∅, b) and (∅, ∅, c).

At Stage 2, test-takers make a context option choice from the two remaining alternatives to create one of six permissible ordered sequence lists notated as (∅, b, a), (∅, c, a), (∅, a, b), (∅, c, b), (∅, a, c), or (∅, b, c).

At Stage 3, test-takers only have one alternative (sound/look like option) left. The endpoints or leaf nodes at the bottom of Figure 6 represent outcomes. The leaf count of six possibilities represents a six decision rule count, that is, six classification rules. In this decision tree, the outcomes of the six ordered triple rank decision rules are left to right: (c, b, a), (b, c, a), (c, a, b), (a, c, b), (b, a, c), and (a, b, c). Therefore, at Stage 3, the six permissible lists are notated as (c, b, a), (b, c, a), (c, a, b), (a, c, b), (b, a, c), and (a, b, c).

Therefore, by definition even the last ordered triple rank response (a, b, c) is conceptualized as an arrangement of list *S* (a, b, c) despite the parallelism.

The Principle of Multiplication of Choices.

The principle of multiplication of choices is also referred to as the fundamental multiplication principle and the product rule. The principle is "if there are *r* successive choices to be made, and for $1 \leq i \leq r$, the *i*th choice can be made in n_i ways, then the total number of ways of making these choices is n_1 x n_2 x … x n_r" (Allenby & Slomson, 2011, 19). Referring back to the decision tree in Figure 1, there were three choices ($n_1 = 3$) at Stage 1, two choices ($n_2 = 2$) at Stage 2, and one choice ($n_3 = 1$) at Stage 3. Therefore, the total number of ways of making these choices in this study is the multiplication of 3 x 2 x 1 = 6 total number of alternatives.

Permutation Formula: When Number of Choices ≤ Number of Elements.

The permutation formula is defined as "for all nonnegative integers *r*, *n* with $r \leq n$, $P(n, r) = n! / (n - r)!$" (Allenby & Slomson, 2011, 19). This formula introduces the factorial *n*, which is symbolized as *n*! In permutations research, there is the convention that 0! = 1. In this study's case, the permutation is $P(n, r) = P(3, 3) = 3! / (3 - 3)!$. Therefore, $P(3, 3) = 6 / (0)! = 6 / 1 = 6$ uniquely ordered permutation arrangements.

This permutation can be visualized in Figure 7, which illustrates the permutations process with the six permutation arrangement outcomes. In other words, *3*! = factorial *n*

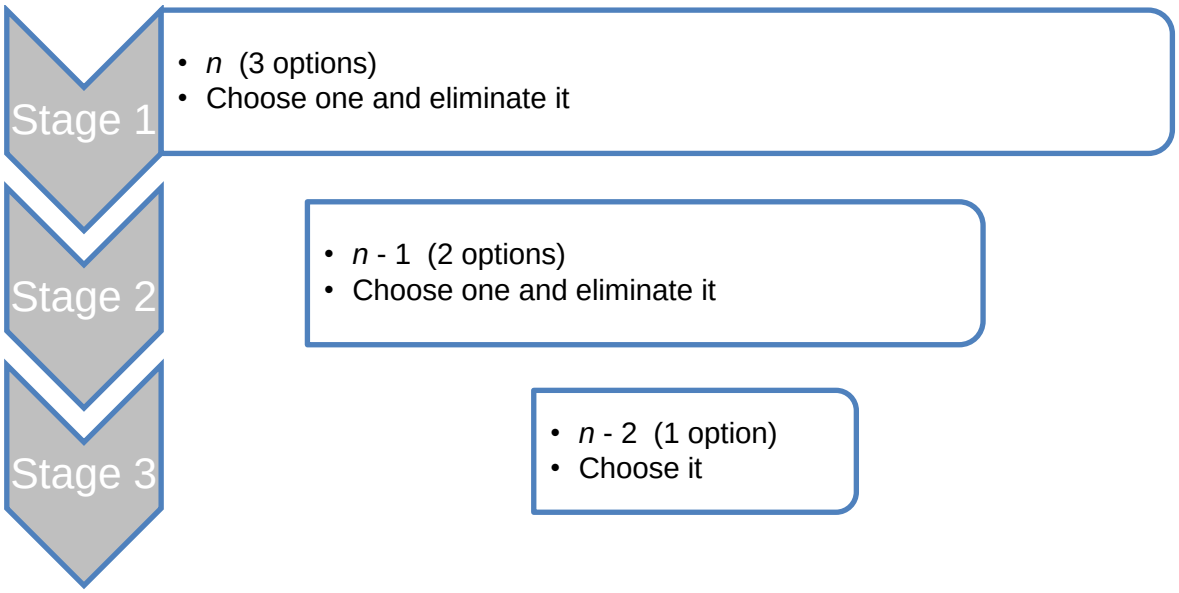


Figure 7. Permutations decision process without repetition.

= 3 x 2 x 1 (Stage 1 x Stage 2 x Stage 3) = 6 permutation arrangements of the 3 elements (a, b and c) of the set. This can also be visualized in the third row of the column *n*! in Table 9.

Permutation Formula: Number of Choices = Total Number of Elements.

The permutation formula when the number of choices, *r*, is equal to the number of elements, *n*, is ${}_nP_n = n!$ (Borowski & Borwein, 2006). Therefore, in this study there are 3! = 6 uniquely ordered arrangements. The number of permutations increases greatly with a small incremental increase in the number of elements as displayed in Table 13.

From the far right column of Table 13, Useful, the number of permutations or polytomous categories that are useful for vocabulary testing studies is only two cases. The first case, is for $n = 2$ (that is, $n! = 2$ permutations), in other words, two categories or the equivalent of one threshold (calculated from the number of categories minus 1 equals the

Table 13

Rank-Ordered Options Permutations

n		*n*!	Categories/ Permutations	Thresholds (Categories -1)	Useful
1	1	1	1	0	No
2	2 * **1**	2 * 1!	2	1	Yes
3	3 * **2 * 1**	3 * 2!	6	5	Yes
4	4 * **3 * 2 * 1**	4 * 3!	24	23	No
5	5 * **4 * 3 * 2 * 1**	5 * 4!	120	119	No
6	6 * **5 * 4 * 3 * 2 * 1**	6 * 5!	720	719	No
7	7 * **6 * 5 * 4 * 3 * 2 * 1**	7 * 6!	5,040	5,039	No
8	8 * **7 * 6 * 5 * 4 * 3 * 2 * 1**	8 * 7!	40,320	40,319	No
9	9 * **8 * 7 * 6 * 5 * 4 * 3 * 2 * 1**	9 * 8!	362,880	362,799	No
10	10 * **9 * 8 * 7 * 6 * 5 * 4 * 3 * 2 * 1**	10 * 9!	3,628,800	3,628,799	No

Note. *n* = number of elements; *n*! = factorial *n*.

number of thresholds) typically used in dichotomously-scored tests such as a vocabulary yes/no checklist test or a multiple-choice test. The second case, is for $n = 3$ (that is, $n! = 6$ permutations), in other words, six categories or the equivalent of five thresholds, as used in this study. Incidentally, these six categories and equivalent five thresholds are also within the range suggested by Miller (1956) in his seminal study of the number seven plus or minus two. Miller (1956) suggested that around seven (that is, from five to nine) might be an ideal number of scale categories within which people have a capacity to transmit and process information.

To conclude this section, the preceding four methods all arrived at the same result. There are six uniquely ordered arrangements, that is, permutations, that can be listed from the three elements (the alphabetical/lexical representations a, b and c) of the multiple-choice three-option set. Note that these six permissible answers are in parallel with the previous procedure, option arrangement.

Answer Key

Procedure 4, answer key, asks and answers the question, what is the correct answer key for each individual item? As explained in Procedure 3, there are six correct answers possible for the items in this test. For each of the 36 items, six permutations are prescribed as correct answers. In other words, the 36 items used in the test were assigned a correct answer in one of six ways as illustrated in Table 14. These six correct answer permutations are all variants of the ordered triple A_n (a, b, c). A typical multiple-choice test with three options and a single best answer response there are only three possible correct answers, a, b, or c. In contrast, Table 14 displays the six permissible correct answers for the test items in

Table 14

A Priori Six Permutations as a Correct Answer Key

List *An* (a, b, c)					
		Correct	Answer		
1	2	3	4	5	6
(0, 1, 2)	(0, 1, 2)	(0, 1, 2)	(0, 1, 2)	(0, 1, 2)	(0, 1, 2)
(c,b,a)	(b,c,a)	(c,a,b)	(a,c,b)	(b,a,c)	(a,b,c)
a = Synonym	a = Synonym	b = Synonym	b = Synonym	c = Synonym	c = Synonym
b = Context	c = Context	a = Context	c = Context	a = Context	b = Context
c = S-/L-Like	b = S-/L-Like	c = S-/L-Like	a = S-/L-Like	b = S-/L-Like	a = S-/L-Like

Note. (0, 1, 2) = Ordered Triple Rank (OTR) representing a category response instruction that sequences 3 form-meaning matches in: 0 = no sentences, 1 = another imagined sentence and 2 = this written sentence; S-/L-Like = Sound-/Look-Like option.

this study. These correct answers are assigned randomly so that there are six instances each of the ordered triple ranks (a, b, c), (b, a, c), (a, c, b), (c, a, b), (b, c, a), and (c, b, a) assigned to the 36 items in the test. These correct answers are randomly assigned to the individual items as displayed in Table 15 and also in Appendix B.

Table 15

Correct Answer Key for 36 Items

Item	Correct answer	Item	Correct answer	Item	Correct answer
1	(b, c, a)	13	(c, a, b)	25	(b, c, a)
2	(a, b, c)	14	(b, c, a)	26	(c, b, a)
3	(c, a, b)	15	(a, c, b)	27	(a, c, b)
4	(b, a, c)	16	(c, b, a)	28	(a, b, c)
5	(c, b, a)	17	(c, a, b)	29	(a, c, b)
6	(a, c, b)	18	(a, b, c)	30	(b, c, a)
7	(b, a, c)	19	(b, a, c)	31	(b, a, c)
8	(a, c, b)	20	(c, b, a)	32	(c, b, a)
9	(c, b, a)	21	(b, c, a)	33	(a, b, c)
10	(b, c, a)	22	(c, a, b)	34	(c, a, b)
11	(b, a, c)	23	(a, b, c)	35	(b, a, c)
12	(a, b, c)	24	(c, a, b)	36	(a, c, b)

Anticipated Responses

Procedure 5, anticipated responses, asks and answers the question, what are the anticipated person response patterns for each individual item? First, the rationale underlying the outcome space response matrix provided to the test-takers is examined.

Response Matrix (3-by-3) Rationale.

The test-takers are required to answer in a 3-by-3 response matrix design for each item on a separate Optical Mark Recognition (OMR) answer sheet. Within this design, the test-takers are instructed not to repeat the letters a, b, and c within each item response. In other words, they are allowed to use only one a, one b, and one c per item response. The letters a, b, and c, which are alphabetical/lexical representations of multiple-choice word options, must be ordered according to a rule. The rule is a sequenced (left to right) form-

meaning match to the bolded keyword in: 0, which represents no match in any sentence, 1, which represents a context match in another imagined sentence, and 2, which represents a synonym match in the written sentence stem of the printed test item. High ability test-takers are expected to be able to respond to an easy item with a sequenced ranking of the three options left to right accordingly: the sound-/look-like (null synform) option in column 0, the context option in column 1, and the synonym option in column 2.

The rationale underlying this 3-by-3 response matrix design is connected to mathematical relations and the Cartesian product rule (Makinson, 2012). In this study, the ordered triple response *A* (a, b, c) represents all six possible item option content correct answers. List *A* can be manifested as six permutations of ordered triples for the correct item answers. For example, one possible permutation of the set *A* {a, b, c} is the ordered triple response *R* (a, b, c) that displays a mirrored parallelism. The conventional alphabetically ordered option labels in the set *L* {a, b, c} and the ordered ranking rule, *N* (0, 1, 2) both have the same cardinality of 3 elements (where *L* = letters and *N* = numbers). Mathematical relations between the set *L* and list *N*'s elements are represented as ordered pairs, for example (a, 0) is one example. The Cartesian product of the set *L* and the ordered triple *N* is mathematically equivalent to (Makinson, 2012):

$L \times N = \{a, b, c\} \times (0, 1, 2)$

$= \{(a, 0), (a, 1), (a, 2), (b, 0), (b, 1), (b, 2), (c, 0), (c, 1), (c, 2)\}$

Therefore, according to the Cartesian product rule, there are nine, that is, three times three, possible ordered pair lists. This underlying Cartesian product rule is visually illustrated in Table 16.

Table 16

Cartesian Product Rule Rubric

		0	1	2
Multiple-choice	a	(a, 0)	(a, 1)	(a, 2)
Option lexical	b	(b, 0)	(b, 1)	(b, 2)
Representation	c	(c, 0)	(c, 1)	(c, 2)

Note. Ordered triple *N* = (0, 1, 2) rank (OTR) representing a category response instruction that sequences 3 form-meaning matches in: 0 = No sentences, 1 = Another imagined sentence, 2 = This written sentence; Set *L* {a, b, c} = multiple-choice options (unordered set according to content).

However, in this study there is a mandate that no repetitions of the letters a, b, and c are allowed. Referring back to Figure 6 and the cell values in Table 16, there are six ordered triple rank response lists permissible from the nine Cartesian product ordered pair lists that can be enumerated for each item as illustrated in Table 17. The six permissible list response elements have been underlined and bolded in the appropriate cells of Table 17.

Table 17

Underlying Mathematical Relations Via the Cartesian Product Rule Lists Resulting in Six Permissible Responses

	0	1	2
1. a	(a, 0)	(a, 1)	***(a, 2)***
b	(b, 0)	***(b, 1)***	(b, 2)
c	***(c, 0)***	(c, 1)	(c, 2)
Response	c	b	a

	0	1	2
3. a	(a, 0)	***(a, 1)***	(a, 2)
b	(b, 0)	(b, 1)	***(b, 2)***
c	***(c, 0)***	(c, 1)	(c, 2)
Response	c	a	b

	0	1	2
5. a	(a, 0)	***(a, 1)***	(a, 2)
b	***(b, 0)***	(b,1)	(b, 2)
c	(c, 0)	(c, 1)	***(c, 2)***
Response	b	a	c

	0	1	2
2. a	(a, 0)	(a, 1)	***(a, 2)***
b	***(b, 0)***	(b, 1)	(b, 2)
c	(c, 0)	***(c, 1)***	(c, 2)
Response	b	c	a

	0	1	2
4. a	***(a, 0)***	(a, 1)	(a, 2)
b	(b, 0)	(b, 1)	***(b, 2)***
c	(c, 0)	***(c, 1)***	(c, 2)
Response	a	c	b

	0	1	2
6. a	***(a, 0)***	(a, 1)	(a, 2)
b	(b, 0)	***(b, 1)***	(b, 2)
c	(c, 0)	(c, 1)	***(c, 2)***
Response	a	b	c

Note. 0, 1, 2 = Ordered Triple Rank (OTR) representing a category response instruction that sequences 3 form-meaning matches in: 0 = no sentences, 1 = another imagined sentence and 2 = this written sentence; a, b, c = multiple-choice options.

In Table 17, the six permissible responses are: 1. (c, b, a) [(c, 0) + (b, 1) + (a, 2)], 2. (b, c, a) [(b, 0) + (c, 1) + (a, 2)], 3. (c, a, b) [(c, 0) + (a, 1) + (b, 2)], 4. (a, c, b) [(a, 0) + (c, 1) + (b, 2)], 5. (b, a, c) [(b, 0) + (a, 1) + (c, 2)] and 6. (a, b, c) [(a, 0) + (b, 1) + (c, 2)]. The ordered triple rank (OTR) responses displayed in Table 17 are also able to be perceived as subsets of the Cartesian product of set L x N (Eccles, 2007). For example, the first permissible subset pattern consisting of three ordered pairs, is notated as, (c, 0), (b, 1), (a, 2) $\subseteq L$ x N in Table 17.

The six permissible subset patterns are displayed more succinctly in Table 18. Table 18 displays example response pattern subsets that have the underlying mathematical relations of ordered pairs as the rationale which forms the Cartesian product rule. The test-takers in this study are given the instruction of no repetitions of letters (L), a, b, c, and

Table 18

Six Permutations of Permissible Correct Answer Patterns Based on Cartesian Product Rule Subset Patterns

	0	1	2
1. a	○	○	●
b	○	●	○
c	●	○	○
Response	c	b	a

	0	1	2
3. a	○	●	○
b	○	○	●
c	●	○	○
Response	c	a	b

	0	1	2
5. a	○	●	○
b	●	○	○
c	○	○	●
Response	b	a	c

	0	1	2
2. a	○	○	●
b	●	○	○
c	○	●	○
Response	b	c	a

	0	1	2
4. a	●	○	○
b	○	○	●
c	○	●	○
Response	a	c	b

	0	1	2
6. a	●	○	○
b	○	●	○
c	○	○	●
Response	a	b	c

Note. (0, 1, 2) = Ordered Triple Rank (OTR) representing a category response instruction that sequences 3 form-meaning matches in: 0 = no sentences, 1 = another imagined sentence and 2 = this written sentence; ● = example test-taker response.

numbers (*N*), 0, 1, 2, in their order triple rank (OTR) responses. Therefore, there are only six allowable responses to each item as displayed in Table 18.

The information explained in the previous section on subsets of Cartesian product rule lists is relevant to the design of the Optical Mark Recognition (OMR) mark sheet administered to test-takers in this study. On the Optical Mark Recognition (OMR) mark sheet that is administered to test-takers, the 3-by-3 response matrix design is manifested visually for 6 items as shown in Figure 8. However, the mark sheet allows 36 ordered triple rank responses to the 36 items in the test. For each item, the test-takers must rank all three options in order of synonymy correctness. The ranking can be conceptualized as left to right: incorrect (0), partially correct (1), and correct (2), respectively.

Example Mark Sheet Matrix Before Test-Taker Responses

	⓿	❶	❷		⓿	❶	❷		⓿	❶	❷
1.	a	a	a	3.	a	a	a	5.	a	a	a
	b	b	b		b	b	b		b	b	b
	c	c	c		c	c	c		c	c	c
2.	a	a	a	4.	a	a	a	6.	a	a	a
	b	b	b		b	b	b		b	b	b
	c	c	c		c	c	c		c	c	c

Example Mark Sheet Matrix After Test-Taker Responses

	⓿	❶	❷		⓿	❶	❷		⓿	❶	❷
1.	a	a	●	3.	a	●	a	5.	a	●	a
	b	●	b		b	b	●		●	b	b
	●	c	c		●	c	c		c	c	●
2.	a	a	●	4.	●	a	a	6.	●	a	a
	●	b	b		b	b	●		b	●	b
	c	●	c		c	●	c		c	c	●

Figure 8. Optical mark recognition 3-by-3 matrix item response design example.

In Figure 8, an example is given of one hypothetical person's Optical Mark Recognition (OMR) responses to items 1 to 6. Reading from left to right, the person responded using the form-meaning match ordered rule of: 0 = no match, 1 = context match in another imagined sentence, and 2 = synonym match in this written sentence. Therefore, from the responses to the first six items, the following raw data was obtained for item 1: c, b, a; item 2: b, c, a; item 3: c, a, b; item 4: a, c, b; item 5: b, a, c; and item 6: a, b, c.

Bifurcating Typical Single Best Answer Responses into Ordered Triple Rank Responses.

In this study, typical single best answer outcome spaces for multiple-choice items are modified to form atypical ordered triple rank outcome spaces. Figure 6's decision tree for synonym rank-order options, Stage 1 (single best answer outcome space) and Stage 3 (ordered triple rank outcome space) serves as an introduction to this section.

As displayed in Figure 9, the single best answer outcome space design is in a single horizontal row orientation.

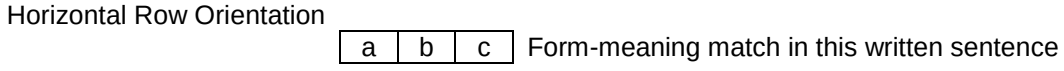


Figure 9. Outcome space design for typical single best answer (SBA) item response.

This row can be rotated 90 degrees clockwise to form Column ❷ of the ordered triple rank outcome space design in Figure 10. In Figure 10, the vertical column orientation

is sequenced across the top according to the form-meaning match rule, 0 = in no sentences, 1 = in another imagined sentence, and 2 = in this written sentence.

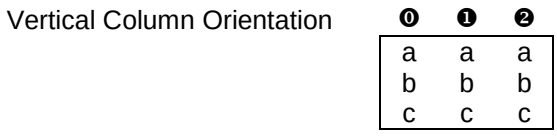


Figure 10. Outcome space design for atypical ordered triple rank (OTR) item responses.

In this study, typical single best answer responses can be bifurcated, that is, split into two, to form ordered triple rank responses of upper and lower levels. For instance, the following example responses in this section all refer to the same correct answer (a, b, c), which forms the ordered triple rank sequenced left to right. Under a single best answer instruction, a correct synonym response, dichotomously scored 1, is shown in Figure 11.

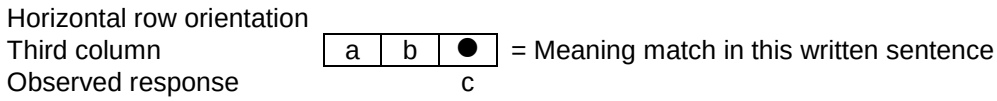


Figure 11. Typical single best answer (SBA) item response: “c”.

However, when bifurcated into ordered triple rank responses, the finer grained ability levels of the person’s responses can be ascertained. In Figure 12, on the left side is a synonym upper response from a relatively higher ability person, while on the right is a synonym lower response from a relatively lower ability person.

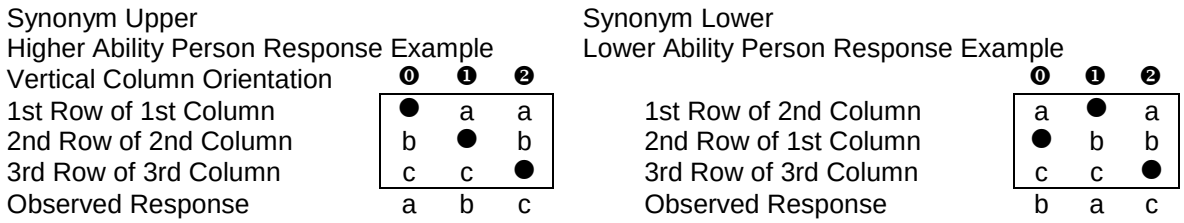


Figure 12. Atypical ordered triple rank (OTR) item responses: Bifurcation of SBA “c”.

Under a single best answer instruction, an incorrect context distractor response, which is dichotomously scored 0, is displayed in Figure 13. There is an observed response of b.

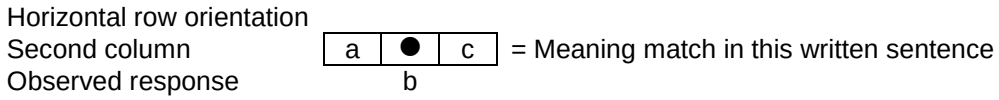


Figure 13. Typical single best answer (SBA) item response: “b”.

However, when bifurcated into ordered triple rank responses, the finer grained ability levels of the person’s responses can be ascertained. In Figure 14, on the left side is a context upper response from a relatively higher ability person, while on the right is a context lower response from a relatively lower ability person.

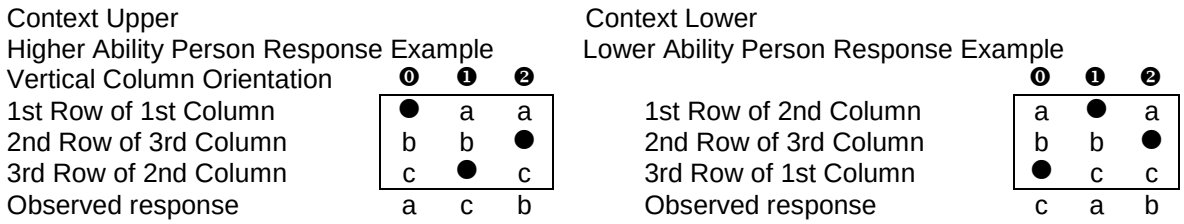


Figure 14. Atypical ordered triple rank (OTR) item responses: Bifurcation of SBA “b”.

Under a single best answer instruction, an incorrect sound-/look-like synform distractor response, which is dichotomously scored 0, is displayed in Figure 15. There is an observed response of a.

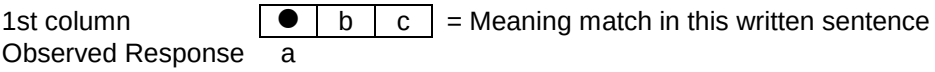


Figure 15. Typical single best answer (SBA) item response: “a”.

However, when bifurcated into ordered triple rank responses, the finer grained ability levels of the person’s responses can be ascertained. In Figure 16, on the left side is a sound-/look-like upper response from a relatively higher ability person, while on the right is a sound-/look-like lower response from a relatively lower ability person.

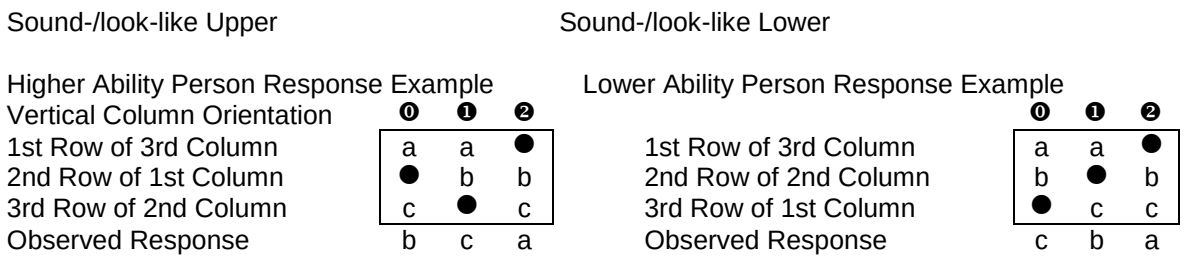


Figure 16. Atypical ordered triple rank (OTR) item responses: Bifurcation of SBA “a”.

An alternative approach to visualizing the response decision process to an ordered triple rank item is displayed in Figure 17. There are two observed ordered triple rank responses that represent bifurcations of the single best answer response of a.

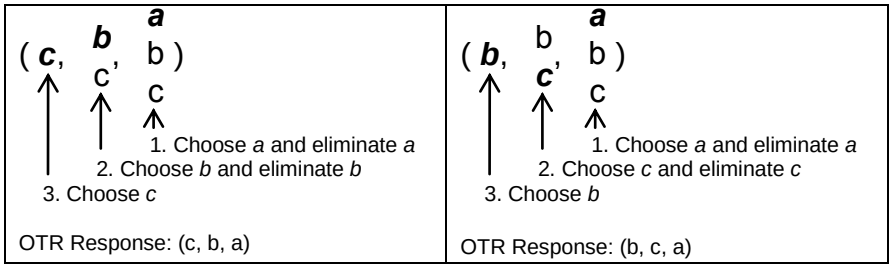


Figure 17. Ordered triple rank (OTR) response decision process without repetition.

The decision processes in Figure 17 illustrate an example describing the steps in creating a (c, b, a) response on the left and a (b, c, a) response on the right. These two processes are amalgamated to create a proxy single best answer response in Figure 18.

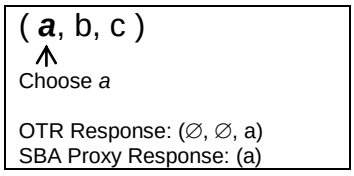


Figure 18. Proxy single best answer (SBA) response.

An alternative approach to visualizing the response decision process to an ordered triple rank item is displayed in Figure 19. There are two observed ordered triple rank responses that represent bifurcations of the single best answer response of b.

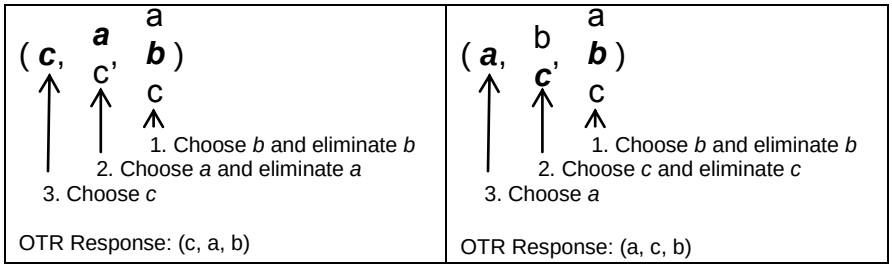


Figure 19. Ordered triple rank (OTR) response decision process without repetition.

The decision processes in Figure 19 illustrate an example describing the steps in creating a (c, a, b) response on the left and a (a, c, b) response on the right. These two processes are amalgamated to create a proxy single best answer response in Figure 20.

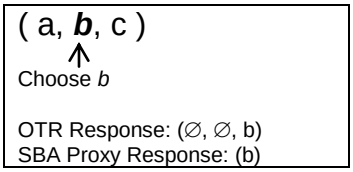


Figure 20. Proxy single best answer (SBA) response.

An alternative approach to visualizing the response decision process to an ordered triple rank item is displayed in Figure 21. There are two observed ordered triple rank responses that represent bifurcations of the single best answer response of c.

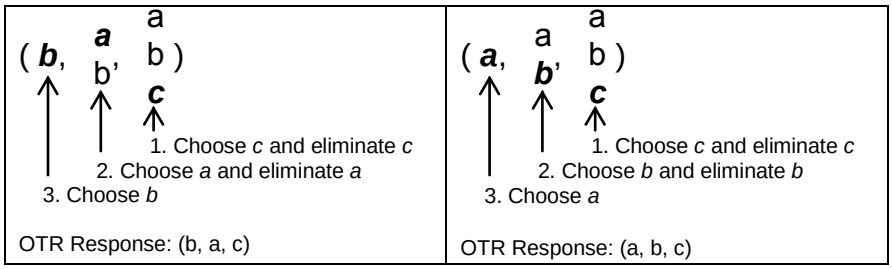


Figure 21. Ordered triple rank (OTR) response decision process without repetition.

The decision processes in Figure 21 illustrate an example describing the steps in creating a (b, a, c) response on the left and a (a, b, c) response on the right. These two processes are amalgamated to create a proxy single best answer response in Figure 22.

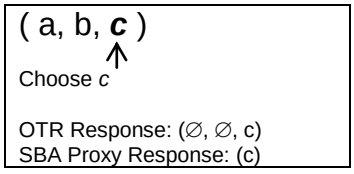


Figure 22. Proxy single best answer (SBA) response.

Scoring Guide

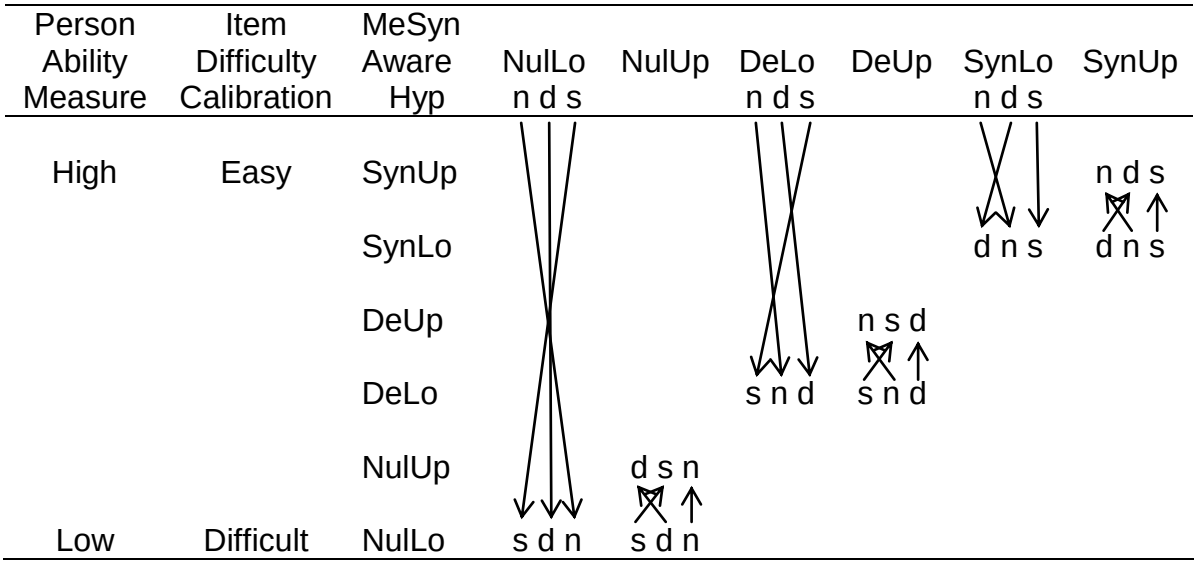
Procedure 6, scoring guide, asks and answers the question, which score should be assigned to each response to individual items? The scoring guide rubric was determined a priori and based on reversed engineering of permutations of the ordered set (null, decontextualized, synonym) abbreviated as (n, d, s) in Table 19.

In Table 19, the fourth column displays the counterpoint of the correct answer (n, d, s) as the incorrect response (s, d, n). A person of a slightly higher ability, or, an item that was slightly easier, would be expected to elicit a response such as in the fifth column (d, s, n). A person of slightly higher ability, or, an item that was slightly easier, would be expected to elicit a response such as in the sixth column (s, n, d). A person of slightly higher ability, or, an item that was slightly easier, would be expected to elicit a response such as in the seventh column (n, s, d). A person of slightly higher ability, or, an item that

was slightly easier, would be expected to elicit a response such as in the eighth column (d, n, s). In the last stage, a person of highest ability, or, an item that was easiest, would be

Table 19

Reverse Engineering the Permutational Word Sense Stages of the Metasynonymy Awareness Hypothesis



Note. MeSyn Aware Hyp = Metasynonymy Awareness Hypothesis; NulLo = Null Lower; NulUp = Null Upper; DeLo = Decontextualized Lower; DeUp = Decontextualized Upper; SynLo = Synonym Lower; SynUp = Synonym Upper.

expected to elicit a response such as in the ninth column (n, d, s). A further description of Table 19 is that the (s, d, n) response in column 4 is a null lower response, the (d, s, n) response in column 5 is a null upper response, the (s, n, d) response in column 6 is a decontextualized lower response, the (n, s, d) response in column 7 is a decontextualized

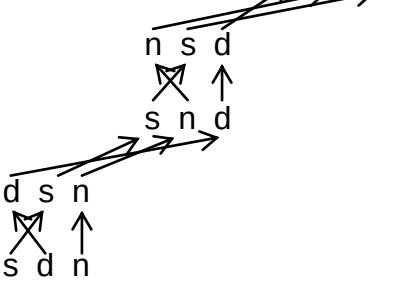
upper response, the (d, n, s) response in column 8 is a synonym lower response and the (n, d, s) response in column 9 is a synonym upper response.

Applying mathematical permutations to a scoring guide for this present study is presented in Table 20. In Table 20 there are three cyclic permutations of particular relevance. A cyclic permutation is a permutation subset where elements exchange places

Table 20

Metasynonymy Awareness Hypothesis Represented as Permutational Word Sense Stages with Scoring Guide

Person Ability Measure	Item Difficulty Calibrat	MeSyn Aware Hyp	Null Lo > Up	Decont-ualized Lo > Up	Synonym Lo > Up	Letter Score	Number Score
High	Easy	SynUp			n d s	z	5
		SynLo			d n s	y	4
		DeUp		n s d		x	3
		DeLo		s n d		w	2
		NulUp	d s n			v	1
Low	Difficult	NulLo	s d n			u	0



Note. Item Difficulty Calib = Item Difficulty Calibration; MeSyn Aware Hyp = Metasynonymy Awareness Hypothesis; NulLo = Null Lower; NulUp = Null Upper; DeLo = Decontextualized Lower; DeUp = Decontextualized Upper; SynLo = Synonym Lower; SynUp = Synonym Upper; Lo = Lower; Up = Upper; Lo > Up = from Lower to Upper.

with one another (Weisstein, 2003). The three cyclic permutations examined have a transposition where two elements from an ordered list trade places with each other while other elements stay the same (Weisstein, 2003). The three cyclic permutations are, first, the

ordered list of null lower (s, d, n) that has two elements *s* and *d* exchanged to form the ordered list of null upper (d, s, n) represented as (s, d)(n), second, the ordered list of decontextualized lower (s, n, d) that has two elements *s* and *n* exchanged to form the ordered list of decontextualized upper (n, s, d) represented as (s, n)(d), and third, the ordered list of synonym lower (d, n, s) that has two elements *d* and *n* exchanged to form the ordered list of synonym upper (n, d, s) represented as (d, n)(s). The hierarchy of the three cyclic permutations is based on an amendment to O'Connor's (1940) vocabulary acquisition theory in order of null, decontextualized and synonym, respectively. However, in this present study the cyclic permutations within each stage form a new hierarchy, that is, null lower, null upper, decontextualized lower, decontextualized upper, synonym lower and synonym upper, respectively.

Columns 7 and 8 of Table 20 display the scoring guide as letters and numbers respectively. At the bottom of Table 20 is the null lower response (s, d, n) and scored 0 that is expected from a person with low ability or elicited from a difficult item. At the top of Table 20 is the synonym upper response (n, d, s) and scored 5 that is expected from a person with high ability or elicited from an easy item. Ascending between these two extremes of the metasynonymy awareness hypothesis scale are incrementally ordered steps.

In advanced mathematical applications, permutations can also be conceptualized as bijective functions or bijections, that is, the permutations are bijective functions or mappings of one-to-one correspondence between two lists. One and only one element of a list is paired with an element of the other list. As explained previously, there is a slight but important distinction between the set theory notation used. The notation for a set, that is, an

unordered set, is indicated by enclosing the set inside braces as notated by {}, for example, Set *Ln* {a, b, c}. Despite the fact that the set elements are alphabetically ordered left to right by Western convention in a written format, the definition of a set is a collection of unordered elements. Therefore, the Set *Sn* {n, d, s} represents the synonymy scale option contents that could be encountered in any order to test-takers in this test's multiple-choice item options. The key point here is that the option contents are randomly ordered. The traditionally used and convenient alphabetically labeled representations of a), b), and c) remain in the same written order for all 36 items. Therefore, hypothetically the written arrangement for item 1 could be a) *s*, b) *d*, c) *n* and for item 2 could be a) *n*, b) *s*, c) *d*. In contrast to the definition of a set, an ordered triple in this study's case has three sequenced elements, also called a 3-tuple, and is indicated by enclosing the listed sequence inside parentheses as notated by ( ), for example, Correct Answer 1 (c, b, a) in the first column of Table 21 (Borowski & Borwein, 2006).

Subsequently, related to each of the six correct answer types, there are six hierarchically ordered levels of scores that can be assigned to a permutation-scored response depending on the level of incremental correctness. Table 21 illustrates this six-category permutations hierarchy scoring guide in a 6-by-6 matrix scoring rubric design.

Table 21 displays two hierarchical levels of permutations. The first and higher level of permutations is denoted in the top heading as Correct Answer 1 (c, b, a), 2 (b, c, a), 3 (c, a, b), 4 (a, c, b), 5 (b, a, c), and 6 (a, b, c). These six ordered triples are the six possible correct answer codes that were assigned randomly to the 36 items and were introduced in Table 15.

The second and lower level of permutations in Table 21 are represented as extensions of the first column, Response, with the labels *Z, Y, X, W, V, U*. Therefore, in the second column of Table 21 the individual elements of the response Z_1 (c,b,a) are paired to Correct Answer 1 (c,b,a) which is the targeted correct response for that particular item. As there is an exact one-to-one correspondence for each element of both sets, this permutation is raw scored at a value of 5 (the maximum score for this item response). In contrast, the

Table 21

A Priori Six Category Responses-Focused Hierarchical Scoring Guide

	List *Sn* (a, b, c)						
			Correct	Answer			
	1	2	3	4	5	6	
Response	(0, 1, 2) (c,b,a)	(0, 1, 2) (b,c,a)	(0, 1, 2) (c,a,b)	(0, 1, 2) (a,c,b)	(0, 1, 2) (b,a,c)	(0, 1, 2) (a,b,c)	Raw score
Z	*Z1* (c,b,a)	*Z2* (b,c,a)	*Z3* (c,a,b)	*Z4* (a,c,b)	*Z5* (b,a,c)	*Z6* (a,b,c)	5
Y	*Y1* (b,c,a)	*Y2* (c,b,a)	*Y3* (a,c,b)	*Y4* (c,a,b)	*Y5* (a,b,c)	*Y6* (b,a,c)	4
X	*X1* (c,a,b)	*X2* b,a,c)	*X3* (c,b,a)	*X4* (a,b,c)	*X5* (b,c,a)	*X6* (a,c,b)	3
W	*W1* (a,c,b)	*W2* (a,b,c)	*W3* (b,c,a)	*W4* (b,a,c)	*W5* (c,b,a)	*W6* (c,a,b)	2
V	*V1* (b,a,c)	*V2* (c,a,b)	*V3* (a,b,c)	*V4* (c,b,a)	*V5* (a,c,b)	*V6* (b,c,a)	1
U	*U1* (a,b,c)	*U2* (a,c,b)	*U3* (b,a,c)	*U4* (b,c,a)	*U5* (c,a,b)	*U6* (c,b,a)	0

Note. (0, 1, 2) = Ordered Triple Rank (OTR) representing a category response instruction that sequences 3 form-meaning matches in: 0 = no sentences, 1 = another imagined sentence and 2 = this written sentence; *Z* = Synonym Upper; *Y* = Synonym Lower; *X* = Context Upper; *W* = Context Lower; *V* = Sound-/Look-Like Upper; *U* = Sound-/Look-Like Lower.

individual elements of the response U_1 (a,b,c) (which is the counterpoint of Set Z_1) are paired to the elements of Correct Answer 1 (c, b, a). As the elements of the response U_1 are in the inverse order of Correct Answer 1's elements, this permutation is raw scored at a value of 0 (the minimum score for this item response).

An alternative way of notating the permutations is to perceive them as circular permutations (Nelson, 2008). The notation for the permutation (or, substitution) group from the second column of Table 19 follows: Z_1 $\binom{c\ b\ a}{c\ b\ a}$, Y_1 $\binom{c\ b\ a}{b\ c\ a}$, X_1 $\binom{c\ b\ a}{c\ a\ b}$, W_1 $\binom{c\ b\ a}{a\ c\ b}$, V_1 $\binom{c\ b\ a}{b\ a\ c}$ and U_1 $\binom{c\ b\ a}{a\ b\ c}$. According to the rightmost column of Table 21, these permutations would be scored 5, 4, 3, 2, 1, and 0 respectively for a test item that had Correct Answer 1 (c,b,a) as the correct response.

Visualizing the scoring guide rubric can alternatively be illustrated in Table 22. Table 22 indicates a focus on the scores in the center of the scoring guide. This is in contrast to having the focus on the response permutations in the center of the scoring

Table 22

A Priori Six Category Scores-Focused Hierarchical Scoring Guide

			List *Sn* (a, b, c)			
O-Triple Response	Answer *1* (c, b, a)	Answer *2* (b, c, a)	Answer *3* (c, a, b)	Answer *4* (a, c, b)	Answer *5* (b, a, c)	Answer *6* (a, b, c)
(c, b, a)	5	4	3	1	2	0
(b, c, a)	4	5	2	0	3	1
(c, a, b)	3	1	5	4	0	2
(a, c, b)	2	0	4	5	1	3
(b, a, c)	1	3	0	2	5	4
(a, b, c)	0	2	1	3	4	5

Note. O-Triple = Ordered Triple; Answer = Correct Answer.

guide as illustrated previously in Table 21. The values in these cells are also useful to conceptualize the next procedure in the protocol, data compression.

An alternative approach to visualizing the scoring guide is displayed in Table 23. The first column on the left lists the six possible correct answers that are permutations of

the ordered triple S_n, (a, b, c). Columns two to seven define the scoring code (0, 1, 2, 3, 4, 5) as an ordered sextuple, that is, a sequence of six ordered elements. The first row underneath the score heading lists the six permissible permutation responses for each correct answer type. Thus, there are three levels of permutations: one, from the test

Table 23

A Priori Six Sextuples-Focused Hierarchical Scoring Guide

List *Sn*			Raw	Score		
(a, b, c)	(0,	1,	2,	3,	4,	5)
(c, b, a)	((a, b, c)	(b, a, c)	(a, c, b)	(c, a, b)	(b, c, a)	(c, b, a))
(b, c, a)	((a, c, b)	(c, a, b)	(a, b, c)	(b, a, c)	(c, b, a)	(b, c, a))
(c, a, b)	((b, a, c)	(a, b, c)	(b, c, a)	(c, b, a)	(a, c, b)	(c, a, b))
(a, c, b)	((b, c, a)	(c, b, a)	(b, a, c)	(a, b, c)	(c, a, b)	(a, c, b))
(b, a, c)	((c, a, b)	(a, c, b)	(c, b, a)	(b, c, a)	(a, b, c)	(b, a, c))
(a, b, c)	((c, b, a)	(b, c, a)	(c, a, b)	(a, c, b)	(b, a, c)	(a, b, c))

designer who assigns one of six permutations as a correct answer to each item, two, from the test-takers who respond with one of six permutations as an ordered triple rank response and three, from the test designer who scores the responses using one of six permutations from the scoring guide.

Table 24 displays the ordered triple rank response scoring guide for Item 5 (correct answer (c, b, a)) when it is polytomously and dichotomously scored. The polytomous senary scoring facilitates a potential maximum measurement information retrieval while the dichotomous binary scoring, with a collapsing of categories, entails a potential loss of measurement information.

Table 24

Ordered Triple Rank Responses Scoring Guide for Item 5 with Correct Answer (c, b, a)

Polytomous senary (6) response	Category label	Category score	Dichotomous binary (2) response	Category label	Category score
(c, b, a)	Syn Upper	5	(c, b, a)	Syn Upper	1
(b, c, a)	Syn Lower	4	(b, c, a)	Distractor	0
(c, a, b)	Con Upper	3	(∅, ∅, b)	Distractor	0
(a, c, b)	Con Lower	2			
(b, a, c)	Like Upper	1	(∅, ∅, c)	Distractor	0
(a, b, c)	Like Lower	0			

Note. Syn Upper = Synonym Upper; Syn Lower = Synonym Lower; Con Upper = Context Upper; Con Lower = Context Lower; Like Upper = Sound-/look-like Upper; Like Lower = Sound-/look-like Lower.

Table 25 displays the changes to the scoring guide after amalgamating the six narrower ordered triple rank response categories into three wider proxy single best answer response categories. For Item 5 this further loss of measurement information progresses from polytomous scoring on the left of Table 25 down to dichotomous scoring on the right of Table 25.

Table 25

Proxy Single Best Answer Responses Scoring Guide for Item 5

Polytomous ternary (3) response	Category label	Category score	Dichotomous binary (2) response	Category label	Category score
(∅, ∅, a)	Synonym	2	(∅, ∅, a)	Synonym	1
(∅, ∅, b)	Context	1	(∅, ∅, b)	Distractor	0
(∅, ∅, c)	Sd/Lk Like	0	(∅, ∅, c)	Distractor	0

Note. Sd/Lk Like = Sound-/look like.

Table 26 displays the ordered triple rank response scoring guide for Item 1 (correct answer (b, c, a)) when it is polytomously and dichotomously scored. The polytomous

senary scoring facilitates a maximum measurement information retrieval while the dichotomous binary scoring, with a collapsing of categories, entails a loss of measurement information.

Table 26

Ordered Triple Rank Responses Scoring Guide for Item 1 with Correct Answer (b, c, a)

Polytomous senary (6) response	Category label	Category score	Dichotomous binary (2) response	Category label	Category score
(b, c, a)	Syn Upper	5	(b, c, a)	Syn Upper	1
(c, b, a)	Syn Lower	4	(c, b, a)	Distractor	0
(b, a, c)	Con Upper	3	(∅, ∅, c)	Distractor	0
(a, b, c)	Con Lower	2			
(c, a, b)	Like Upper	1	(∅, ∅, b)	Distractor	0
(a, c, b)	Like Lower	0			

Note. Syn Upper = Synonym Upper; Syn Lower = Synonym Lower; Con Upper = Context Upper; Con Lower = Context Lower; Like Upper = Sound-/look-like Upper; Like Lower = Sound-/look-like Lower.

Table 27 displays the changes to the scoring guide after amalgamating the six narrower ordered triple rank response categories into three wider proxy single best answer response categories. For Item 1 this further loss of measurement information progresses

Table 27

Proxy Single Best Answer Responses Scoring Guide for Item 1

Polytomous ternary (3) response	Category label	Category score	Dichotomous binary (2) response	Category label	Category score
(∅, ∅, a)	Synonym	2	(∅, ∅, a)	Synonym	1
(∅, ∅, c)	Context	1	(∅, ∅, c)	Distractor	0
(∅, ∅, b)	Sd/Lk Like	0	(∅, ∅, b)	Distractor	0

Note. Sd/Lk Like = Sound-/look like.

from polytomous scoring on the left of Table 27 down to dichotomous scoring on the right of Table 27.

Table 28 displays the ordered triple rank response scoring guide for Item 3 (correct answer [c, a, b]) when it is polytomously and dichotomously scored. The polytomous senary scoring facilitates maximum measurement information retrieval while the dichotomous binary scoring, with a collapsing of categories, entails a loss of measurement information.

Table 28

Ordered Triple Rank Responses Scoring Guide for Item 3 with Correct Answer (c, a, b)

Polytomous senary (6) response	Category label	Category score	Dichotomous binary (2) response	Category label	Category score
(c, a, b)	Syn Upper	5	(c, a, b)	Syn Upper	1
(a, c, b)	Syn Lower	4	(a, c, b)	Distractor	0
(c, b, a)	Cont Upper	3	(∅, ∅, a)	Distractor	0
(b, c, a)	Cont Lower	2			
(a, b, c)	Like Upper	1	(∅, ∅, c)	Distractor	0
(b, a, c)	Like Lower	0			

Note. Syn Upper = Synonym Upper; Syn Lower = Synonym Lower; Con Upper = Context Upper; Con Lower = Context Lower; Like Upper = Sound-/look-like Upper and Like Lower = Sound-/look-like Lower.

Table 29 displays the changes to the scoring guide after amalgamating the six narrower ordered triple rank response categories into three wider proxy single best answer response categories. For Item 3 this further potential loss of measurement information progresses from polytomous scoring on the left of Table 29 down to dichotomous scoring on the right of Table 29.

Table 29

Proxy Single Best Answer Responses Scoring Guide for Item 3

Polytomous ternary (3) response	Category label	Category score	Dichotomous binary (2) response	Category label	Category score
(∅, ∅, b)	Synonym	2	(∅, ∅, b)	Synonym	1
(∅, ∅, a)	Context	1	(∅, ∅, a)	Distractor	0
(∅, ∅, c)	Sd/Lk Like	0	(∅, ∅, c)	Distractor	0

Note. Sd/Lk Like = Sound-/look like.

Table 30 displays the ordered triple rank response scoring guide for Item 6 (correct answer (a, c, b)) when it is polytomously and dichotomously scored. The polytomous senary scoring facilitates maximum measurement information retrieval while the dichotomous binary scoring, with a collapsing of categories, entails a loss of measurement information.

Table 30

Ordered Triple Rank Responses Scoring Guide for Item 6 with Correct Answer (a, c, b)

Polytomous senary (6) response	Category label	Category score	Dichotomous binary (2) response	Category label	Category score
(a, c, b)	Syn Upper	5	(a, c, b)	Syn Upper	1
(c, a, b)	Syn Lower	4	(c, a, b)	Distractor	0
(a, b, c)	Cont Upper	3	(∅, ∅, c)	Distractor	0
(b, a, c)	Cont Lower	2			
(c, b, a)	Like Upper	1	(∅, ∅, a)	Distractor	0
(b, c, a)	Like Lower	0			

Note. Syn Upper = Synonym Upper; Syn Lower = Synonym Lower; Con Upper = Context Upper; Con Lower = Context Lower; Like Upper = Sound-/look-like Upper; Like Lower = Sound-/look-like Lower.

Table 31 displays the changes to the scoring guide after amalgamating the six narrower ordered triple rank response categories into three wider proxy single best answer

response categories. For Item 6 this further loss of measurement information progresses from polytomous scoring on the left of Table 31 down to dichotomous scoring on the right of Table 31.

Table 31

Proxy Single Best Answer Responses Scoring Guide for Item 6

Polytomous ternary (3) response	Category label	Category score	Dichotomous binary (2) response	Category label	Category score
(∅, ∅, b)	Synonym	2	(∅, ∅, b)	Synonym	1
(∅, ∅, c)	Context	1	(∅, ∅, c)	Distractor	0
(∅, ∅, a)	Sd/Lk Like	0	(∅, ∅, a)	Distractor	0

Note. Sd/Lk Like = Sound-/look like.

Table 32 displays the ordered triple rank response scoring guide for Item 4 (correct answer (b, a, c)) when it is polytomously and dichotomously scored. The polytomous senary scoring facilitates maximum measurement information retrieval while the

Table 32

Ordered Triple Rank Responses Scoring Guide for Item 4 with Correct Answer (b, a, c)

Polytomous senary (6) response	Category label	Category score	Dichotomous binary (2) response	Category label	Category score
(b, a, c)	Syn Upper	5	(b, a, c)	Syn Upper	1
(a, b, c)	Syn Lower	4	(a, b, c)	Distractor	0
(b, c, a)	Cont Upper	3	(∅, ∅, a)	Distractor	0
(c, b, a)	Cont Lower	2			
(a, c, b)	Like Upper	1	(∅, ∅, b)	Distractor	0
(c, a, b)	Like Lower	0			

Note. Syn Upper = Synonym Upper; Syn Lower = Synonym Lower; Con Upper = Context Upper; Con Lower = Context Lower; Like Upper = Sound-/look-like Upper; Like Lower = Sound-/look-like Lower.

dichotomous binary scoring, with a collapsing of categories, entails a loss of measurement information.

Table 33 displays the changes to the scoring guide after amalgamating the six narrower ordered triple rank response categories into three wider proxy single best answer response categories. For Item 4 this further loss of measurement information progresses from polytomous scoring on the left of Table 33 down to dichotomous scoring on the right of Table 33.

Table 33

Proxy Single Best Answer Responses Scoring Guide for Item 4

Polytomous ternary (3) response	Category label	Category score	Dichotomous binary (2) response	Category label	Category score
(∅, ∅, c)	Synonym	2	(∅, ∅, c)	Synonym	1
(∅, ∅, a)	Context	1	(∅, ∅, a)	Distractor	0
(∅, ∅, b)	Sd/Lk Like	0	(∅, ∅, b)	Distractor	0

Note. Sd/Lk Like = Sound-/look like.

Table 34 displays the ordered triple rank response scoring guide for Item 2 (correct answer (a, b, c)) when it is polytomously and dichotomously scored. The polytomous senary scoring facilitates maximum measurement information retrieval while the dichotomous binary scoring, with a collapsing of categories, entails a loss of measurement information.

Table 34

Ordered Triple Rank Responses Scoring Guide for Item 2 with Correct Answer (a, b, c)

Polytomous senary (6) response	Category label	Category score	Dichotomous binary (2) response	Category label	Category score
(a, b, c)	Syn Upper	5	(a, b, c)	Syn Upper	1
(b, a, c)	Syn Lower	4	(b, a, c)	Distractor	0
(a, c, b)	Cont Upper	3	(∅, ∅, b)	Distractor	0
(c, a, b)	Cont Lower	2			
(b, c, a)	Like Upper	1	(∅, ∅, a)	Distractor	0
(c, b, a)	Like Lower	0			

Note. Syn Upper = Synonym Upper; Syn Lower = Synonym Lower; Con Upper = Context Upper; Con Lower = Context Lower; Like Upper = Sound-/look-like Upper; Like Lower = Sound-/look-like Lower.

Table 35 displays the changes to the scoring guide after amalgamating the six narrower ordered triple rank response categories into three wider proxy single best answer response categories. For Item 2 this further loss of measurement information progresses from polytomous scoring on the left of Table 35 down to dichotomous scoring on the right of Table 35.

Table 35

Single Best Answer Responses Scoring Guide for Item 2

Polytomous ternary (3) response	Category label	Category score	Dichotomous binary (2) response	Category label	Category score
(∅, ∅, c)	Synonym	2	(∅, ∅, c)	Synonym	1
(∅, ∅, b)	Context	1	(∅, ∅, b)	Distractor	0
(∅, ∅, a)	Sd/Lk Like	0	(∅, ∅, a)	Distractor	0

Note. Sd/Lk Like = Sound-/look like.

Another alternative to visualizing the processes is to consolidate two of the five procedures together. The two procedures of anticipated responses and scoring guide are visualized together in this section through a series of figures.

To facilitate the description of this section, the mathematical Pigeonhole Principle is amended. The pigeonhole principle is also referred to as Dirichlet's Box Principle. The principle was first proposed by Dirichlet in 1834 who stated that when there are *n* boxes and *m* greater than *n* objects, at a minimum one box must contain more than one object (Weisstein, 2003). In other words, this basic enumeration principle examines how a set *n* of objects, for instance, pigeons, fly into *p* pigeonholes and where $1 \leq p < n$, then at least a pigeonhole must have two pigeons inside (Nelson, 2008).

In the case of this study, the set *L* {a, b, c) has equal cardinality (three elements) as the list *N* (0, 1, 2). However, the pigeonhole principle can be adapted to suit the purpose of explanation of this study's case. In this study, the test-takers task is to fly the three elements of the set *L* (a, b, c "pigeons") into the three elements of list *N* (0, 1, 2 "pigeonholes").

Figure 23 a) displays the advanced mathematics concept of a permutation as a bijective function, also referred to as a bijection. Bijections are both injective and surjective.

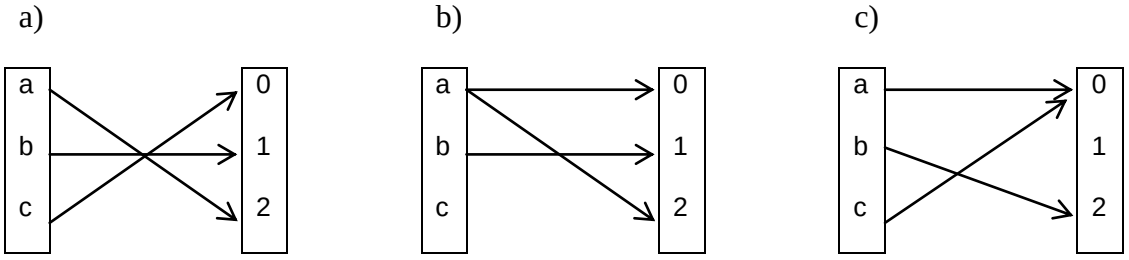


Figure 23. (a) A bijective function (bijection) between the 3 elements of the 2 sets; (b) and (c) are non-bijections.

An injective function is a one-to-one function. A surjective function is an onto function. In other words, in a bijection all the elements of the first list are mapped on to all the elements of the second list. List elements are matched separately without any sharing of mappings. An example bijection is displayed in Figure 23 a).

The bijection in Figure 23 a) displays how every element in the domain (the left side set) is mapped to an element in the codomain (the right hand list). The non-bijection example in Figure 23 b) displays how the element *a* in the domain is mapped to both elements *0* and *2* in the codomain. Also, element *c* in the domain is not mapped to any element in the codomain. The non-bijection example in Figure 23 c) displays how the elements *a* and *c* in the domain are mapped to the element *0* in the codomain. Also, there is no mapping to the element *1* in the codomain.

In terms of the pigeonhole principle displayed in Figures 24 to 29, the respectively parallel bijective and non-bijective examples displayed in Figures 23 a), b) and c) can be visualized in Figure 24 a), b) and c). However, instead of displaying the set *L* and list *N* as parallel as in Figures 23 a), b) and c), Figure 24 a), b) and c) displays the list *L* and list *N* as perpendicular. Despite this difference, the principle of forming mappings between the set *L* and list *N* remain the same.

Figure 24 a) displays a permissible ordered triple rank response that forms a bijection. Each element of the domain has been mapped on to an element of the codomain. Figure 24 b) displays a non-permissible response that does not form a bijection. The domain element *a* has been mapped on to both the codomain elements *0* and *2*. The domain element *c* is not mapped to any codomain element. Figure 24 c) displays a non-permissible

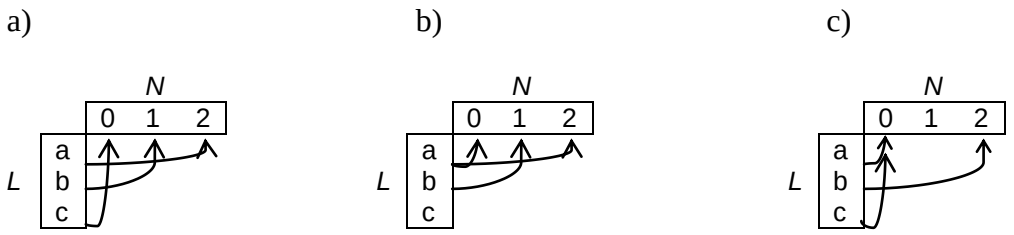


Figure 24. (a) A permissible ordered triple rank (OTR) response; (b) and (c) are impermissible ordered triple rank responses.

response that does not form a bijection. Both the domain elements *a* and *c* are mapped on to the codomain element *0*. The codomain element *1* has no mapping on to it from a domain element.

Therefore, test-takers are instructed to respond with no repetitions of letters or numbers in their ordered triple rank responses. In other words, test-takers are instructed to respond with ordered triple rank (OTR) response bijections.

Applying the bijection mandate to the pigeonhole principle facilitates a visualization of permissible ordered triple rank (OTR) responses. As shown in previous sections this can be achieved in one of six ways for each item response. Figures 25 to 30 display the permissible ordered triple rank responses with an accompanying scoring guide underneath.

Figure 25 displays the possible ordered triple rank responses and scores for item 2 with the correct answer key of (a, b, c). Note that each element of set *L* is mapped to an element of list *N* and therefore forms a bijection.

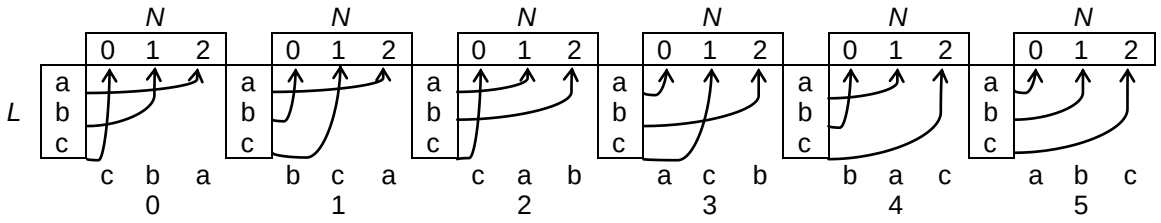


Figure 25. Ordered triple rank (OTR) responses for item 2 where (a, b, c) is correct.

Figure 26 displays the possible ordered triple rank responses and scores for item 4 with the correct answer key of (b, a, c). Note that each element of set *L* is mapped to an element of list *N* and therefore forms a bijection.

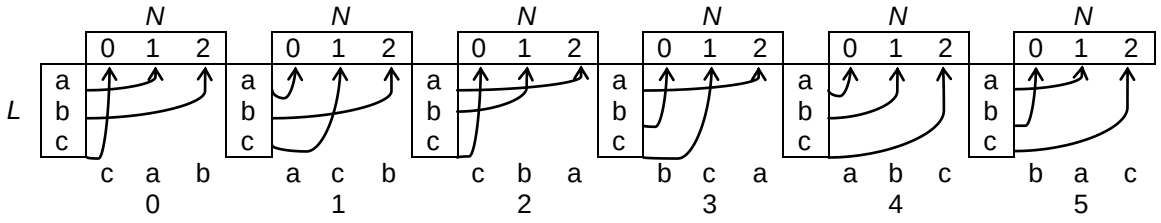


Figure 26. Ordered triple rank (OTR) responses for item 4 where (b, a, c) is correct.

Figure 27 displays the possible ordered triple rank responses and scores for item 6 with the correct answer key of (a, c, b). Note that each element of set *L* is mapped to an element of list *N* and therefore forms a bijection.

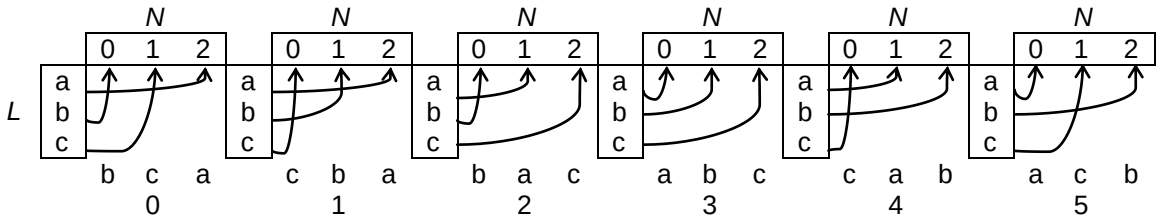


Figure 27. Ordered triple rank (OTR) responses for item 6 where (a, c, b) is correct.

Figure 28 displays the possible ordered triple rank responses and scores for item 3 with the correct answer key of (c, a, b). Note that each element of set *L* is mapped to an element of list *N* and therefore forms a bijection.

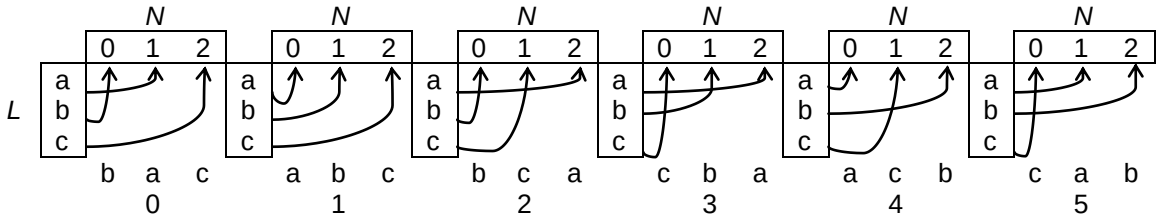


Figure 28. Ordered triple rank (OTR) responses for item 3 where (c, a, b) is correct.

Figure 29 displays the possible ordered triple rank responses and scores for item 1 with the correct answer key of (b, c, a). Note that each element of set *L* is mapped to an element of list *N* and therefore forms a bijection.

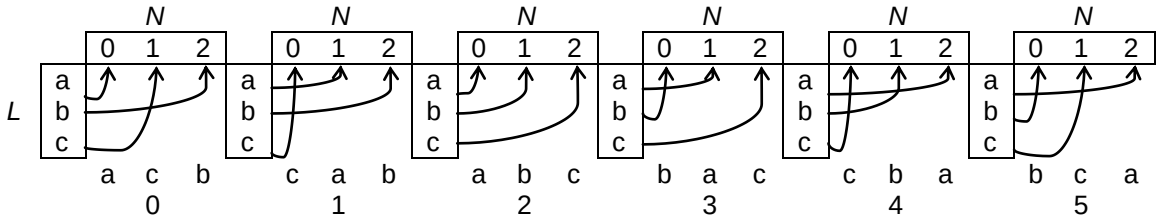


Figure 29. Ordered triple rank (OTR) responses for item 1 where (b, c, a) is correct.

Figure 30 displays the possible ordered triple rank responses and scores for item 5 with the correct answer key of (c, b, a). Note that each element of set *L* is mapped to an element of list *N* and therefore forms a bijection.

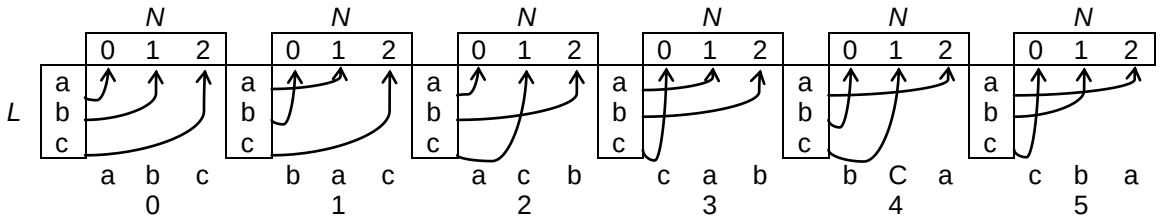


Figure 30. Ordered triple rank (OTR) responses for item 5 where (c, b, a) is correct.

An alternative approach to visualizing the scoring guide is the use of scoring guide trees. Figure 31 displays a scoring guide tree for item 5 with the correct answer (c, b, a). At Stage 1, scores are displayed for single best answer responses (scored with binary (B0 or B1) and ternary (T0, T1 or T2) scales) and at Stage 3, scores are displayed for ordered triple rank responses (scored with binary (B0 or B1) and senary (S0, S1, S2, S3, S4 or S5) scales).

In order to attain the ultimate objective of spreading person ability measures over a wide range as possible, an ordered triple rank (OTR) response senary scoring mandate in Stage 3 of Figure 31 is proposed as the ultimate categorization for capturing polytomous measurement information from this dataset. Therefore, at an optimal and ideal measurement level the proposed ordered triple rank (OTR) responses and senary (S) scoring guide are respectively: (a, b, c) = 0, (b, a, c) = 1, (a, c, b) = 2, (c, a, b) = 3, (b, c, a) = 4 and (c, b, a) = 5. From this senary scoring baseline, subsequent analyses are enabled in which categories can be collapsed or recoded to optimize information from the dataset.

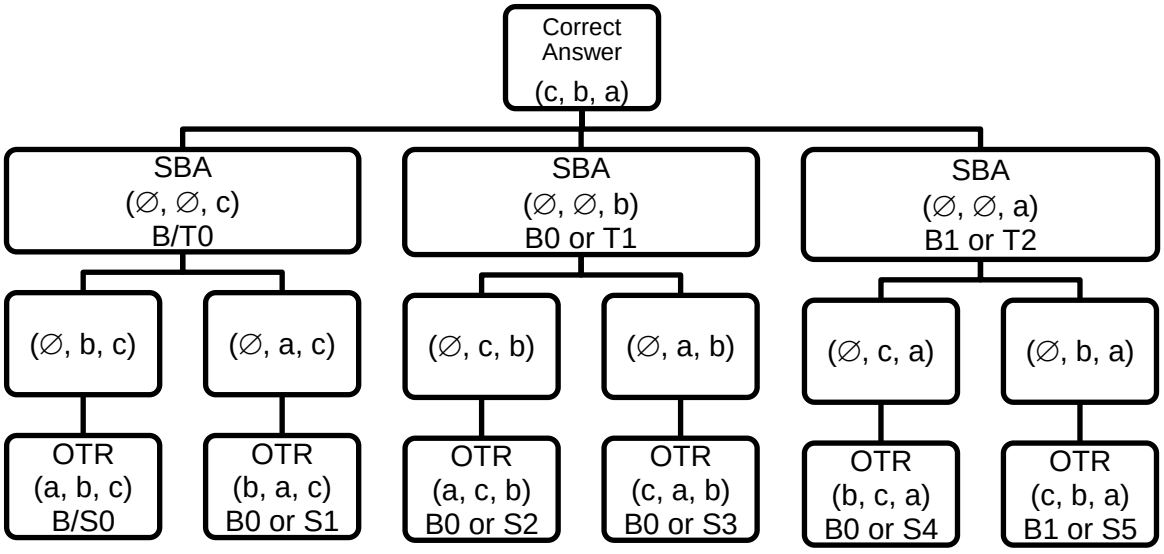


Figure 31. Scoring guide tree for item 5 with correct answer (c, b, a), indicating permissible single best answer (SBA) and ordered triple rank (OTR) responses and respective scores. B/T0 = Binary/Ternary score 0; B0 = Binary score 0; B1 = Binary score 1; T1 = Ternary 1; T2 = Ternary 2; B/S0 = Binary/ Senary score 0; S1 = Senary score 1; S2 = Senary score 2; S3 = Senary score 3; S4 = Senary score 4; S5 = Senary score 5; ∅ = empty element.

Figure 32 displays a scoring guide tree for item 1 with the correct answer (b, c, a). At Stage 1, scores are displayed for single best answer responses (scored with binary (B0 or B1) and ternary (T0, T1 or T2) scales) and at Stage 3, scores are displayed for ordered triple rank responses (scored with binary (B0 or B1) and senary (S0, S1, S2, S3, S4 or S5) scales).

In order to attain the ultimate objective of spreading person ability measures over a wide range as possible, an ordered triple rank (OTR) response senary scoring mandate in Stage 3 of Figure 32 is proposed as the ultimate categorization for capturing polytomous measurement information from this dataset. Therefore, at an optimal and ideal measurement level the proposed ordered triple rank (OTR) responses and senary

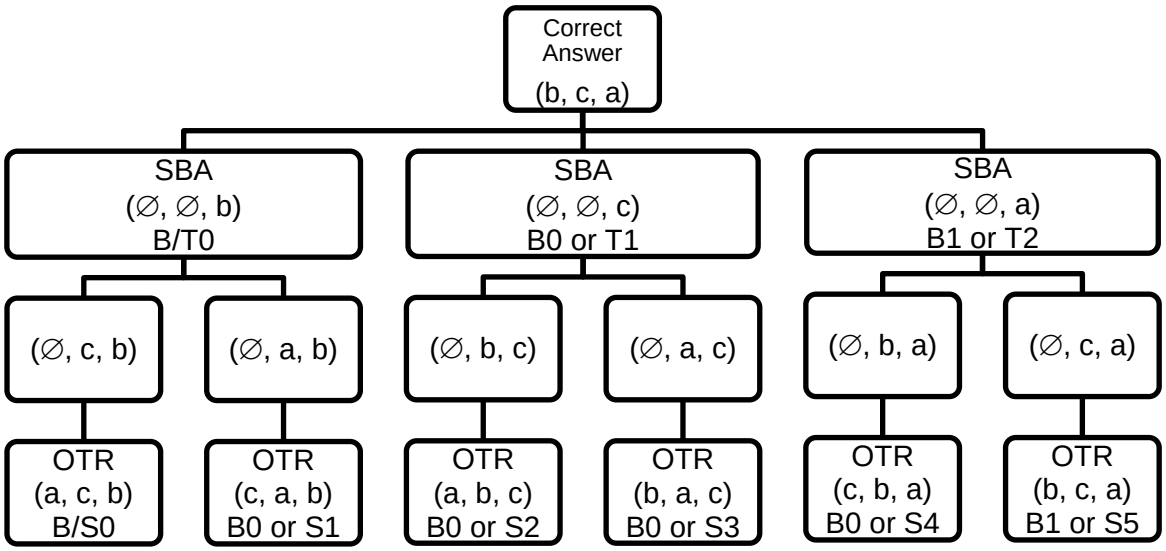


Figure 32. Scoring guide tree for item 1 with correct answer (b, c, a), indicating permissible single best answer (SBA) and ordered triple rank (OTR) responses and respective scores. B/T0 = Binary/Ternary score 0; B0 = Binary score 0; B1 = Binary score 1; T1 = Ternary 1; T2 = Ternary 2; B/S0 = Binary/ Senary score 0; S1 = Senary score 1; S2 = Senary score 2; S3 = Senary score 3; S4 = Senary score 4; S5 = Senary score 5; ∅ = empty element.

(S) scoring guide are respectively: (a, c, b) = 0, (c, a, b) = 1, (a, b, c) = 2, (b, a, c) = 3, (c, b, a) = 4 and (b, c, a) = 5. From this senary scoring baseline, subsequent analyses are enabled in which categories can be collapsed or recoded to optimize information from the dataset.

Figure 33 displays a scoring guide tree for item 3 with the correct answer (c, a, b). At Stage 1, scores are displayed for single best answer responses (scored with binary (B0 or B1) and ternary (T0, T1 or T2) scales) and at Stage 3, scores are displayed for ordered triple rank responses (scored with binary (B0 or B1) and senary (S0, S1, S2, S3, S4 or S5) scales).

In order to attain the ultimate objective of spreading person ability measures over a wide range as possible, an ordered triple rank (OTR) response senary scoring mandate in

Stage 3 of Figure 33 is proposed as the ultimate categorization for capturing polytomous measurement information from this dataset. Therefore, at an optimal and ideal measurement level the proposed ordered triple rank (OTR) responses and senary (S) scoring guide are respectively: (b, a, c) = 0, (a, b, c) = 1, (b, c, a) = 2, (c, b, a) = 3, (a, c, b) = 4 and (c, a, b) = 5. From this senary scoring baseline, subsequent analyses are enabled in which categories can be collapsed to optimize information from the dataset.

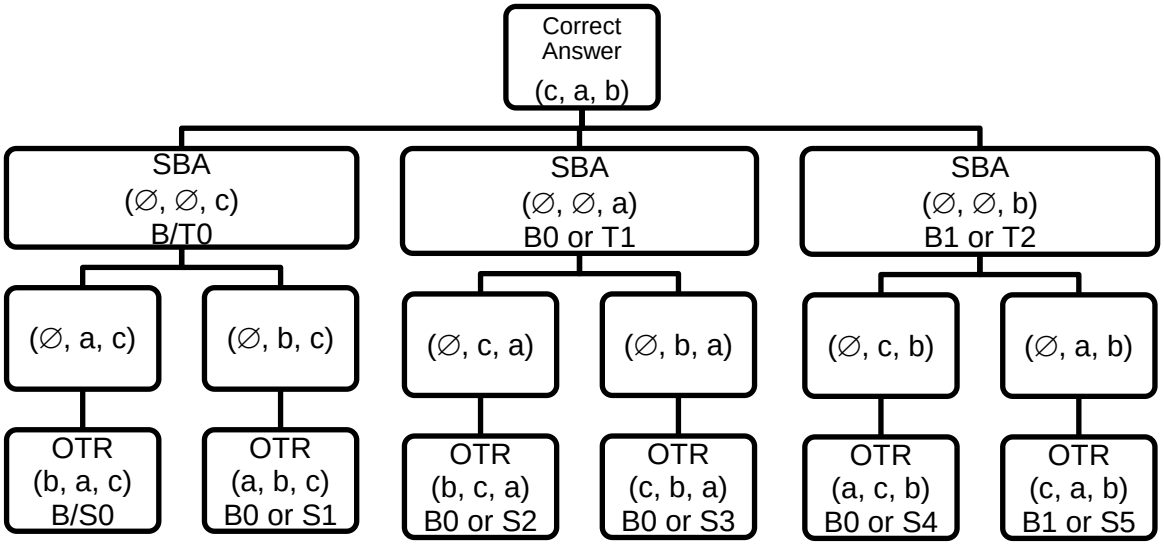


Figure 33. Scoring guide tree for item 3 with correct answer (c, a, b), indicating permissible single best answer (SBA) and ordered triple rank (OTR) responses and respective scores. B/T0 = Binary/Ternary score 0; B0 = Binary score 0; B1 = Binary score 1; T1 = Ternary 1; T2 = Ternary 2; B/S0 = Binary/ Senary score 0; S1 = Senary score 1; S2 = Senary score 2; S3 = Senary score 3; S4 = Senary score 4; S5 = Senary score 5; ∅ = empty element.

Figure 34 displays a scoring guide tree for an item 6 with the correct answer (a, c, b). At Stage 1, scores are displayed for single best answer responses (scored with binary (B0 or B1) and ternary (T0, T1 or T2) scales) and at Stage 3, scores are displayed for ordered

triple rank responses (scored with binary (B0 or B1) and senary (S0, S1, S2, S3, S4 or S5) scales).

In order to attain the ultimate objective of spreading person ability measures over a wide range as possible, an ordered triple rank (OTR) response senary scoring mandate in Stage 3 of Figure 34 is proposed as the ultimate categorization for capturing polytomous measurement information from this dataset. Therefore, at an optimal and ideal measurement level the proposed ordered triple rank (OTR) responses and senary

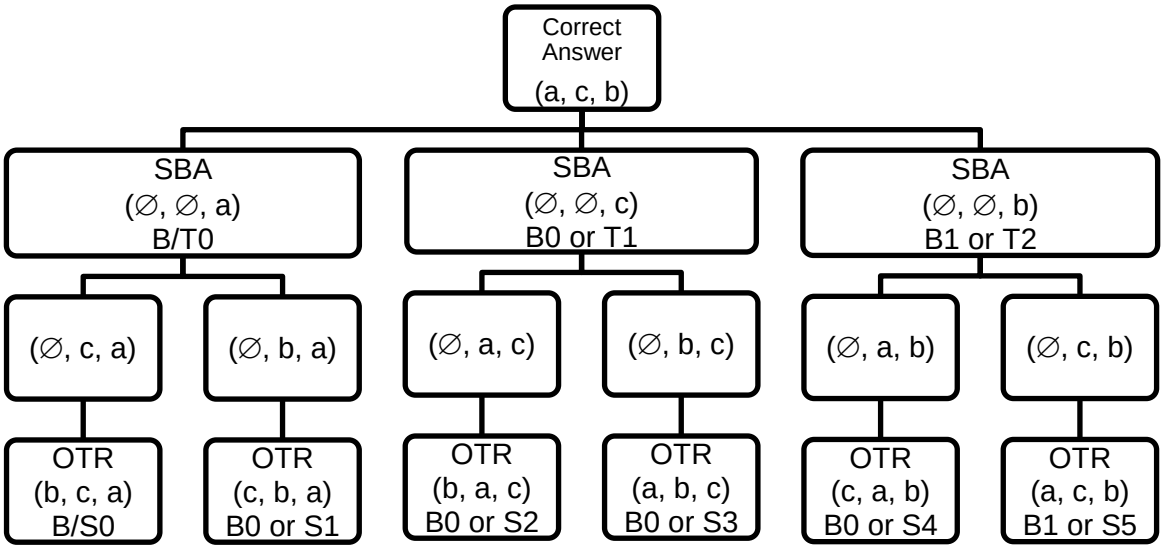


Figure 34. Scoring guide tree for item 6 with correct answer (a, c, b), indicating permissible single best answer (SBA) and ordered triple rank (OTR) responses and respective scores. B/T0 = Binary/Ternary score 0; B0 = Binary score 0; B1 = Binary score 1; T1 = Ternary 1; T2 = Ternary 2; B/S0 = Binary/ Senary score 0; S1 = Senary score 1; S2 = Senary score 2; S3 = Senary score 3; S4 = Senary score 4; S5 = Senary score 5; ∅ = empty element.

(S) scoring guide are respectively: (b, c, a) = 0, (c, b, a) = 1, (b, a, c) = 2, (a, b, c) = 3, (c, a, b) = 4 and (a, c, b) = 5. From this senary scoring baseline, subsequent analyses are enabled in which categories can be collapsed or recoded to optimize information from the dataset.

Figure 35 displays a scoring guide tree for item 4 with the correct answer (b, a, c). At Stage 1, scores are displayed for single best answer responses (scored with binary (B0 or B1) and ternary (T0, T1 or T2) scales) and at Stage 3, scores are displayed for ordered triple rank responses (scored with binary (B0 or B1) and senary (S0, S1, S2, S3, S4 or S5) scales).

In order to attain the ultimate objective of spreading person ability measures over a wide range as possible, an ordered triple rank (OTR) response senary scoring mandate in Stage 3 of Figure 35 is proposed as the ultimate categorization for capturing polytomous measurement information from this dataset. Therefore, at an optimal and ideal measurement level the proposed ordered triple rank (OTR) responses and senary (S) scoring guide are respectively: (c, a, b) = 0, (a, c, b) = 1, (c, b, a) = 2, (b, c, a) = 3, (a, b, c) = 4 and (b, a, c) = 5. From this senary scoring baseline, subsequent analyses are enabled in which categories can be collapsed or recoded to optimize information from the dataset.

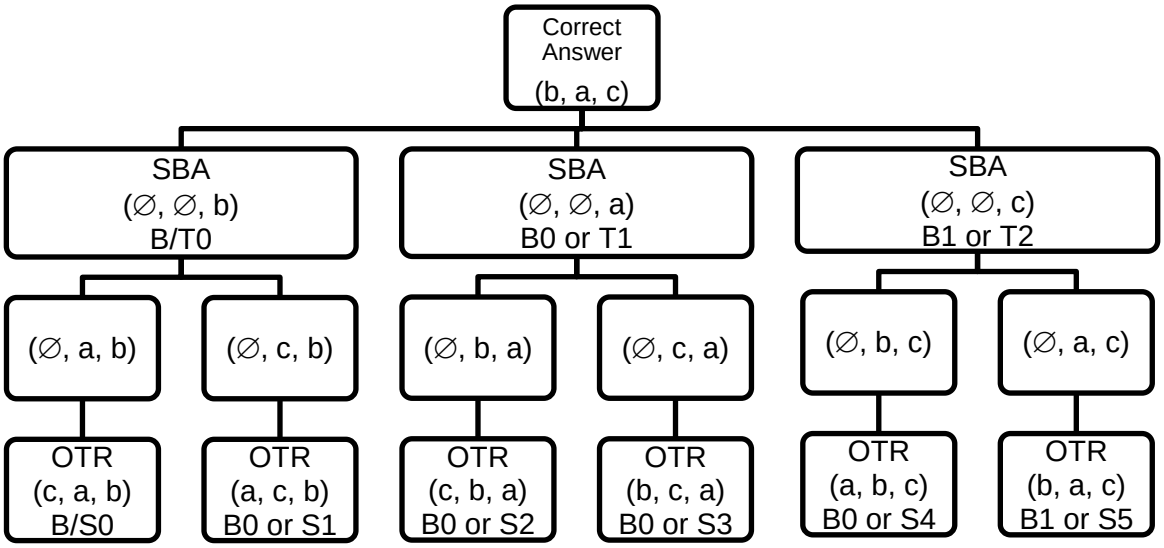


Figure 35. Scoring guide tree for item 4 with correct answer (b, a, c), indicating permissible single best answer (SBA) and ordered triple rank (OTR) responses and respective scores. B/T0 = Binary/Ternary score 0; B0 = Binary score 0; B1 = Binary score 1; T1 = Ternary 1; T2 = Ternary 2; B/S0 = Binary/ Senary score 0; S1 = Senary score 1; S2 = Senary score 2; S3 = Senary score 3; S4 = Senary score 4; S5 = Senary score 5; ∅ = empty element.

Figure 36 displays a scoring guide tree for item 2 with the correct answer (a, b, c). At Stage 1, scores are displayed for single best answer responses (scored with binary (B0 or B1) and ternary (T0, T1 or T2) scales) and at Stage 3, scores are displayed for ordered triple rank responses (scored with binary (B0 or B1) and senary (S0, S1, S2, S3, S4 or S5) scales).

In order to attain the ultimate objective of spreading person ability measures over a wide range as possible, an ordered triple rank (OTR) response senary scoring mandate in Stage 3 of Figure 36 is proposed as the ultimate categorization for capturing polytomous measurement information from this dataset. Therefore, at an optimal and ideal measurement level the proposed ordered triple rank (OTR) responses and senary

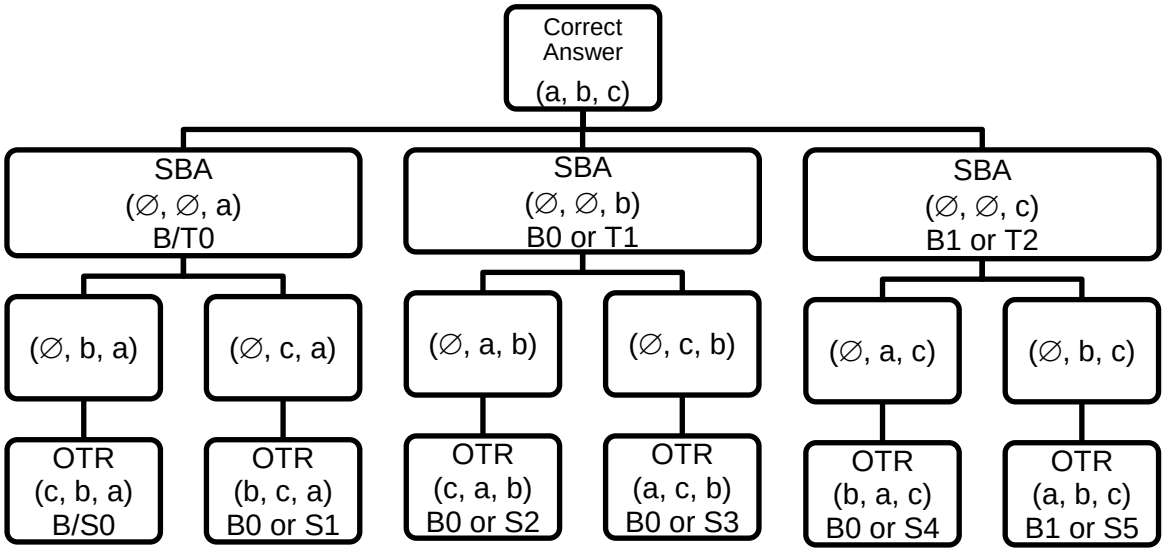


Figure 36. Scoring guide tree for item 2 with correct answer (a, b, c), indicating permissible single best answer (SBA) and ordered triple rank (OTR) responses and respective scores. B/T0 = Binary/Ternary score 0; B0 = Binary score 0; B1 = Binary score 1; T1 = Ternary 1; T2 = Ternary 2; B/S0 = Binary/ Senary score 0; S1 = Senary score 1; S2 = Senary score 2; S3 = Senary score 3; S4 = Senary score 4; S5 = Senary score 5; ∅ = empty element.

(S) scoring guide are respectively: (c, b, a) = 0, (b, c, a) = 1, (c, a, b) = 2, (a, c, b) = 3, (b, a, c) = 4 and (a, b, c) = 5. From this senary scoring baseline, subsequent analyses are enabled in which categories can be collapsed or recoded to optimize information from the dataset.

Data Compression

Procedure 6, data compression, asks and answers the question, how to compress the dataset usefully for applying it to Winsteps software control files? In order to submit a rectangular dataset to Winsteps software, person responses were modified from Ordered Triple Rank (OTR) responses (3 alphabetical letters format) to a single data value in either an alphabetic or numeric format. The alphabetic dataset format facilitated the use of

Winsteps with a Dichotomous Rasch Model and the numeric dataset format facilitated the use of Winsteps with the Rasch-Andrich Rating Scale Model. Data compression can be classified into two general types: lossless and lossy. Lossless data compression means that there is no loss of information when the data has been compressed and then reconstructed. Lossy data compression means that there is a loss of some information when the data has been compressed and then reconstructed (Sayood, 2012). In this study, the ordered triple rank response dataset is an example of lossless data compression. In contrast, the single best answer dataset is an example of lossy data compression.

Figure 37 a) displays the ordered triple rank response scores for ordered triple response lossless data compression and Figure 37 b) displays single best answer proxy response lossy data compression for item 5 with a correct answer (c, b, a). The composition of functions f and g is notated by $g \circ f$ (Makinson, 2012). In Figure 37 a), each element of domain set R, function f aligns with an element of the codomain set L. Then, for an element of domain set L, function g aligns with an element of codomain set $B\,S$. Alternatively, for an element of domain set R, the composition of functions f and g, that is, $g \circ f$, aligns with an element of codomain set $B\,S$. In other words, the ordered triple rank response dataset can be compressed into a Winsteps software rectangular control file ready format. Figure 37 b) displays parallel compositions for single best answer (SBA) proxy responses via lossy data compression.

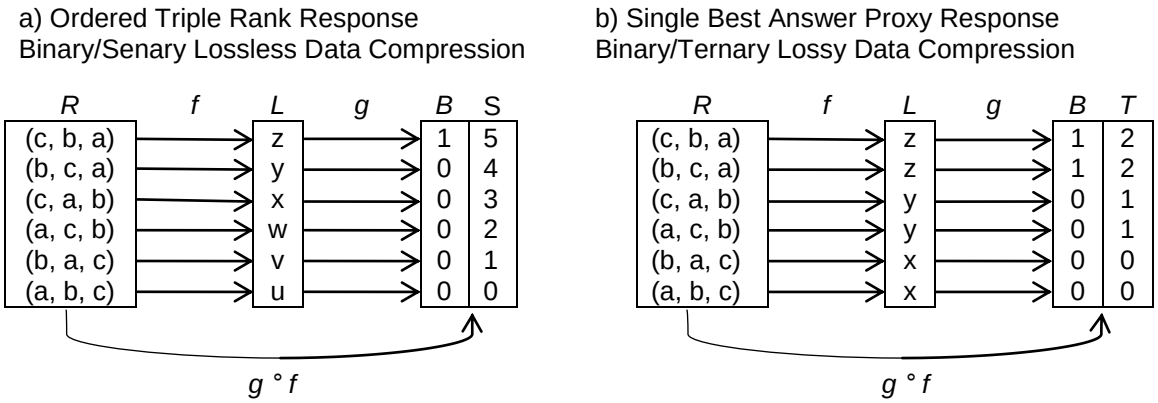


Figure 37. a) Function composition representing lossless and b) lossy data compression for item 5 with correct answer (c, b, a). *R* = response; *L* = letter score; *B* = binary score; *S* = senary score; *T* = ternary score; *f* = function *f*; *g* = function *g*; *g* ° *f* = composition of functions *f* and *g*.

Figure 38 a) displays the ordered triple rank response scores for ordered triple response lossless data compression and Figure 38 b) single best answer proxy response lossy data compression for item 1 with a correct answer (b, c, a). The composition of functions *f* and *g* is notated by *g* ° *f* (Makinson, 2012). In Figure 38 a), each element of domain set *R*, function *f* aligns with an element of the codomain set *L*. Then, for an element of domain set *L*, function *g* aligns with an element of codomain set *B S*. Alternatively, for an element of domain set *R*, the composition of functions *f* and *g*, that is, *g* ° *f*, aligns with an element of codomain set *B S*. In other words, the ordered triple rank response dataset can be compressed into a Winsteps software rectangular control file ready format. Figure 38 b) displays parallel compositions for single best answer (SBA) proxy responses via lossy data compression.

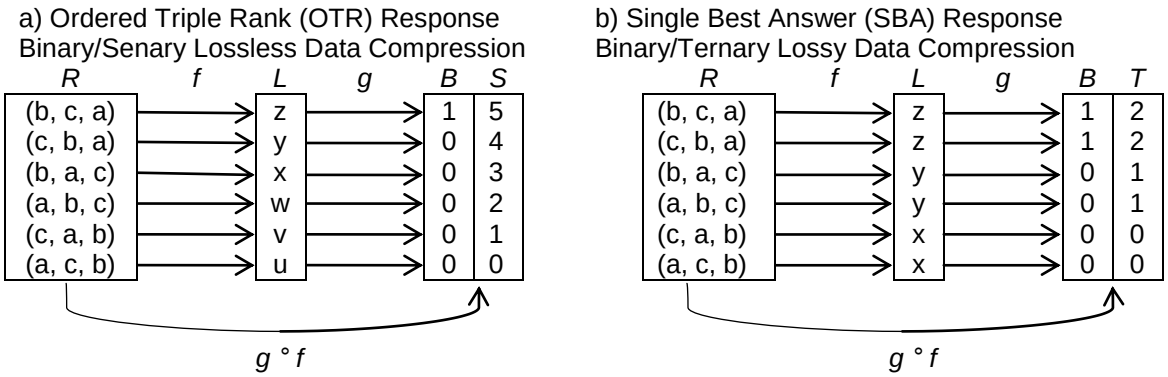


Figure 38. a) Function composition representing lossless and b) lossy data compression for item 1 with correct answer (b, c, a). *R* = response; *L* = letter score; *B* = binary score; *S* = senary score; *T* = ternary score; *f* = function *f*; *g* = function *g*; *g* ° *f* = composition of functions *f* and *g*.

Figure 39 a) displays the ordered triple rank response scores for ordered triple response lossless data compression and Figure 39 b) single best answer proxy response lossy data compression for item 3 with a correct answer (c, a, b). The composition of functions *f* and *g* is notated by *g* ° *f* (Makinson, 2012). In Figure 38 a), each element of domain set *R*, function *f* aligns with an element of the codomain set *L*. Then, for an element of domain set *L*, function *g* aligns with an element of codomain set *B S*. Alternatively, for an element of domain set *R*, the composition of functions *f* and *g*, that is, *g* ° *f*, aligns with an element of codomain set *B S*. In other words, the ordered triple rank response dataset can be compressed into a Winsteps software rectangular control file ready format. Figure 39 b) displays parallel compositions for single best answer (SBA) proxy responses via lossy data compression.

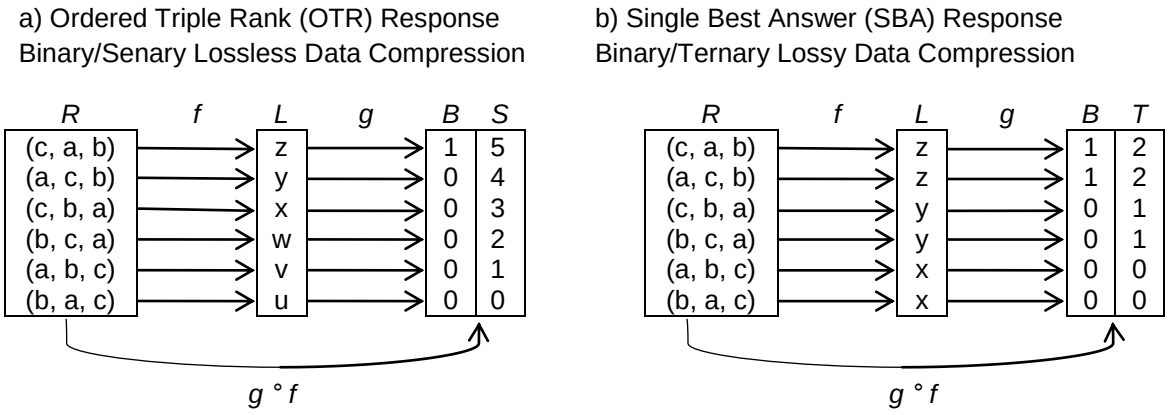


Figure 39. a) Function composition representing lossless and b) lossy data compression for item 3 with correct answer (c, a, b). *R* = response; *L* = letter score; *B* = binary score; *S* = senary score; *T* = ternary score; *f* = function *f*; *g* = function *g*; *g* ° *f* = composition of functions *f* and *g*.

Figure 40 a) displays the ordered triple rank response scores for ordered triple response lossless data compression and Figure 40 b) displays single best answer proxy response lossy data compression on the right for item 6 with a correct answer (a, c, b). The composition of functions *f* and *g* is notated by *g* ° *f* (Makinson, 2012). In Figure 40 a), each element of domain set *R*, function *f* aligns with an element of the codomain set *L*. Then, for an element of domain set *L*, function *g* aligns with an element of codomain set *B S*. Alternatively, for an element of domain set *R*, the composition of functions *f* and *g*, that is, *g* ° *f*, aligns with an element of codomain set *B S*. In other words, the ordered triple rank response dataset can be compressed into a Winsteps software rectangular control file ready format. Figure 40 b) displays parallel compositions for single best answer (SBA) proxy responses via lossy data compression.

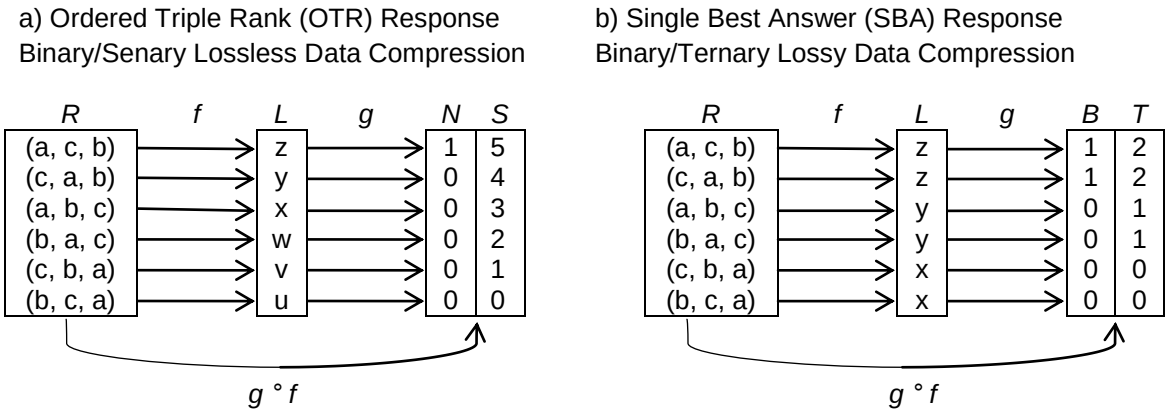


Figure 40. a) Function composition representing lossless and b) lossy data compression for item 6 with correct answer (a, c, b). *R* = response; *L* = letter score; *B* = binary score; *S* = senary score; *T* = ternary score; *f* = function *f*; *g* = function *g*; *g* ° *f* = composition of functions *f* and *g*.

Figure 41 a) displays the ordered triple rank response scores for ordered triple response lossless data compression and Figure 41 b) single best answer proxy response lossy data compression for item 4 with a correct answer (b, a, c). The composition of functions *f* and *g* is notated by *g* ° *f* (Makinson, 2012). In Figure 41 a), each element of domain set *R*, function *f* aligns with an element of the codomain set *L*. Then, for an element of domain set *L*, function *g* aligns with an element of codomain set *B S*. Alternatively, for an element of domain set *R*, the composition of functions *f* and *g*, that is, *g* ° *f*, aligns with an element of codomain set *B S*. In other words, the ordered triple rank response dataset can be compressed into a Winsteps software rectangular control file ready format. Figure 41 b) displays parallel compositions for single best answer (SBA) proxy responses via lossy data compression.

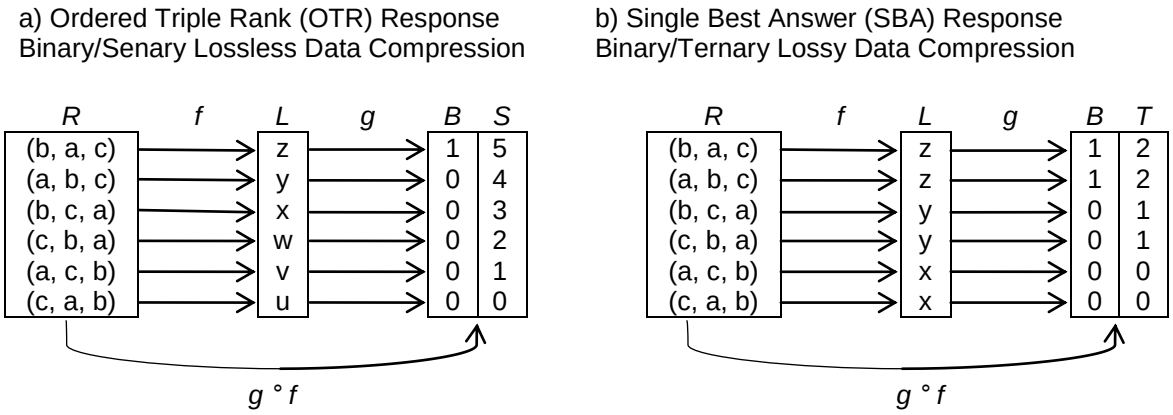


Figure 41. a) Function composition representing lossless and b) lossy data compression for item 4 with correct answer (b, a, c). *R* = response; *L* = letter score; *B* = binary score; *S* = senary score; *T* = ternary score; *f* = function *f*; *g* = function *g*; *g* ° *f* = composition of functions *f* and *g*.

Figure 42 a) displays the ordered triple rank response scores for ordered triple response lossless data compression and Figure 42 b) displays single best answer proxy response lossy data compression for item 2 with a correct answer (a, b, c). The composition of functions *f* and *g* is notated by *g* ° *f* (Makinson, 2012). In Figure 42 a), each element of domain set *R*, function *f* aligns with an element of the codomain set *L*. Then, for an element of domain set *L*, function *g* aligns with an element of codomain set *B S*. Alternatively, for an element of domain set *R*, the composition of functions *f* and *g*, that is, *g* ° *f*, aligns with an element of codomain set *B S*. In other words, the ordered triple rank response dataset can be compressed into a Winsteps software rectangular control file ready format. Figure 42 b) displays parallel compositions for single best answer (SBA) proxy responses via lossy data compression.

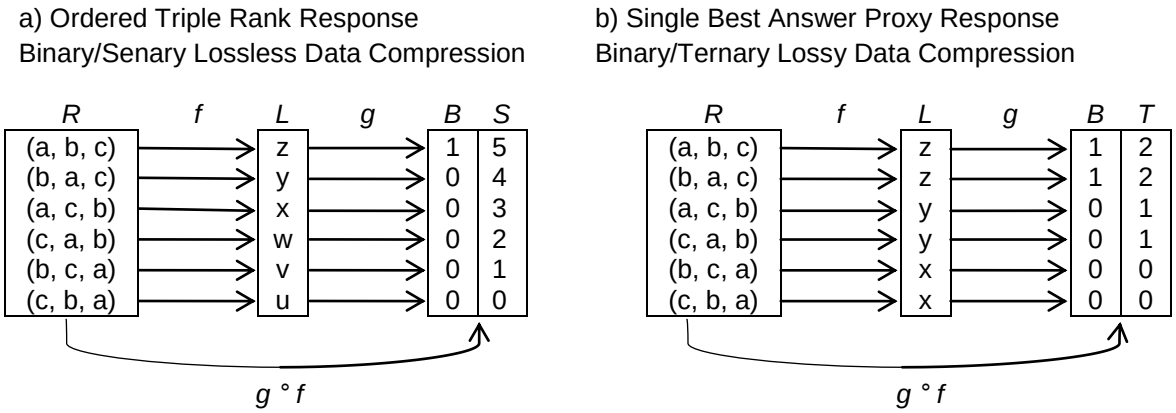


Figure 42. a) Function composition representing lossless and b) lossy data compression for item 2 with correct answer (a, b, c). *R* = response; *L* = letter score; *B* = binary score; *S* = senary score; *T* = ternary score; *f* = function *f*; *g* = function *g*; *g* ° *f* = composition of functions *f* and *g*.

Alternative visual representations of lossless and lossy data compression for ordered triple rank (OTR) and single best answer (SBA) proxy responses are displayed in Figures 43 to 54. Note that Figures 37 to 42 display composition of only two functions (f and g). In contrast, Figures 43 to 54 display composition of both two and three functions (from f, g, and h) (Makinson, 2012).

Figure 43 displays the composition of three functions, f, g, and h, starting from ordered triple rank (OTR) responses to item 5 with the correct answer (c, b, a). Lossless data compression, that is, with no loss of information when data is compressed and reconstructed, is applied to the responses that are represented by a three-way scoring guide using letter (L), binary (B) and senary (S) scales. These three scoring guide scales represent possible lossless data compression transformations to the dataset so that it can be used in a Winsteps software control file.

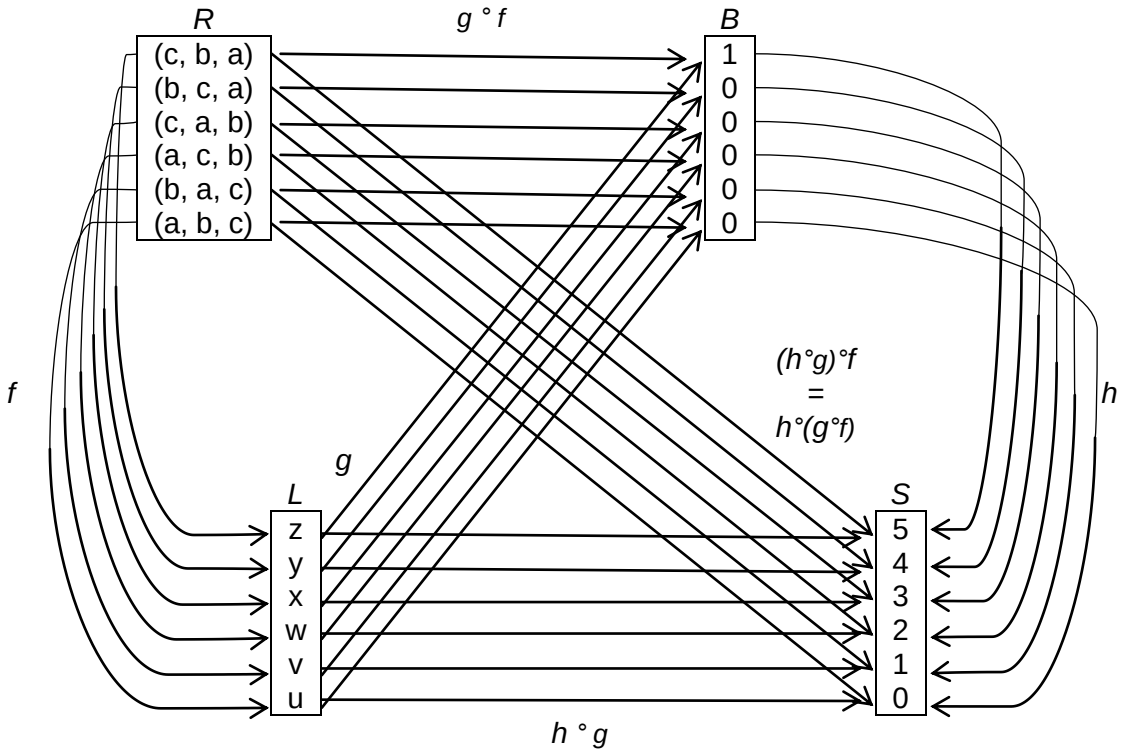


Figure 43. Three function composition (f, g and h) representing ordered triple rank responses via lossless data compression for item 5 with correct answer (c, b, a) when binary and senary scored. R = Response; L = Letter; B = Binary; S = Senary; *f*, *g, h* = functions; *g ° f*, *h ° g*, *(h°g)°f*, *h°(g°f)* = compositions of functions.

Figure 44 displays the composition of three functions, f, g and h, starting from proxy single best answer (SBA) responses to item 5 with the correct answer (c, b, a). Lossy data compression, that is, with some loss of information when data is compressed and reconstructed, is applied to the responses that are represented by a three-way scoring guide using letter (L), binary (B) and ternary (T) scales. These three scoring guide scales represent possible lossy data compression transformations to the dataset so that it can be used in a Winsteps software control file.

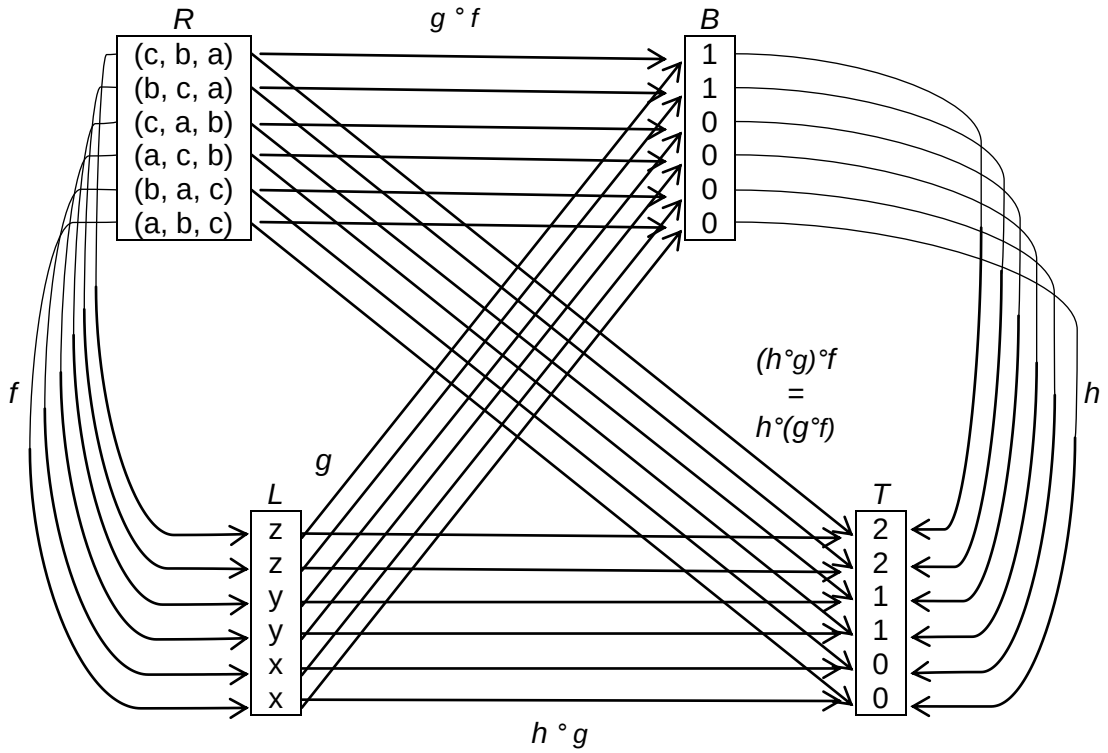


Figure 44. Three function composition (f, g, and h) representing proxy single best answer responses via lossy data compression for item 5 with correct answer (c, b, a) when binary and ternary scored. R = Response; L = Letter; B = Binary; T = Ternary; *f*, *g, h* = functions; *g ° f*, *h ° g*, *(h°g)°f*, *h°(g°f)* = compositions of functions.

Figure 45 displays the composition of three functions, f, g and h, starting from ordered triple rank (OTR) responses to item 1 with the correct answer (b, c, a). Lossless data compression, that is, with no loss of information when data is compressed and reconstructed, is applied to the responses that are represented by a three way scoring guide using letter (L), binary (B) and senary (S) scales. These three scoring guide scales represent possible lossless data compression transformations to the dataset so that it can be used in a Winsteps software control file.

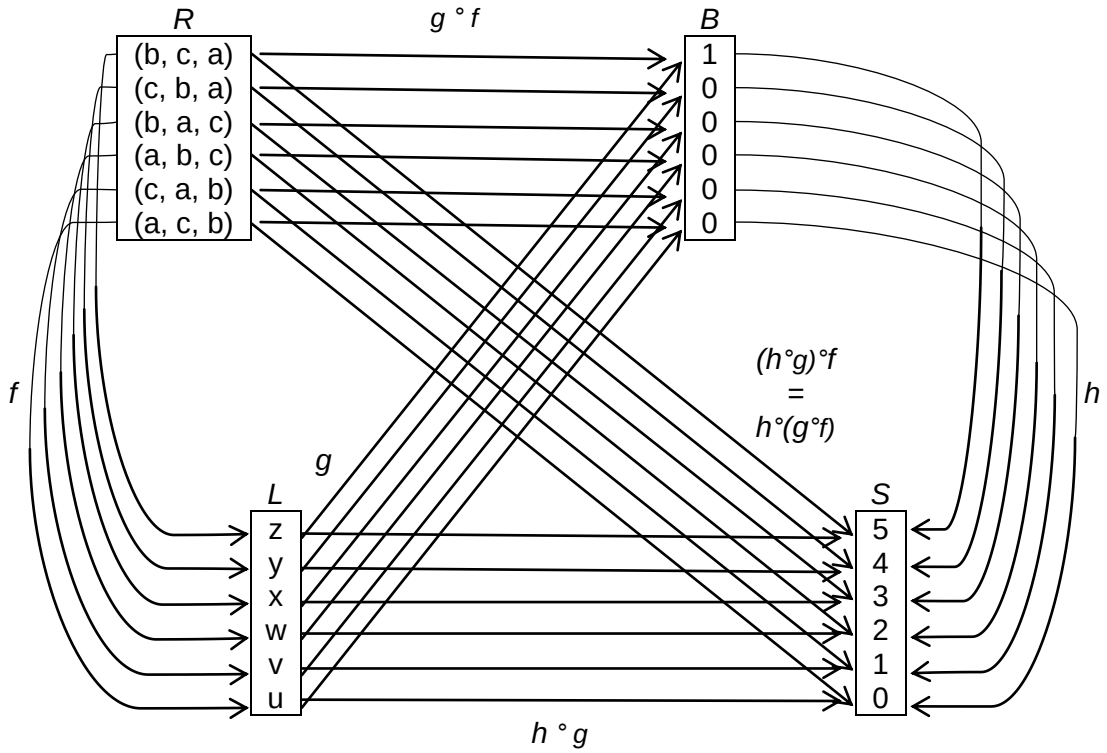


Figure 45. Three function composition (f, g and h) representing ordered triple rank responses via lossless data compression for item 1 with correct answer (b, c, a) when binary and senary scored. R = Response; L = Letter; B = Binary; S = Senary; *f*, *g, h* = functions; *g ° f*, *h ° g*, *(h°g)°f*, *h°(g°f)* = compositions of functions.

Figure 46 displays the composition of three functions, f, g and h, starting from proxy single best answer (SBA) responses to item 1 with the correct answer (b, c, a). Lossy data compression, that is, with some loss of information when data is compressed and reconstructed, is applied to the responses that are represented by a three way scoring guide using letter (L), binary (B) and ternary (T) scales. These three scoring guide scales represent possible lossy data compression transformations to the dataset so that it can be used in a Winsteps software control file.

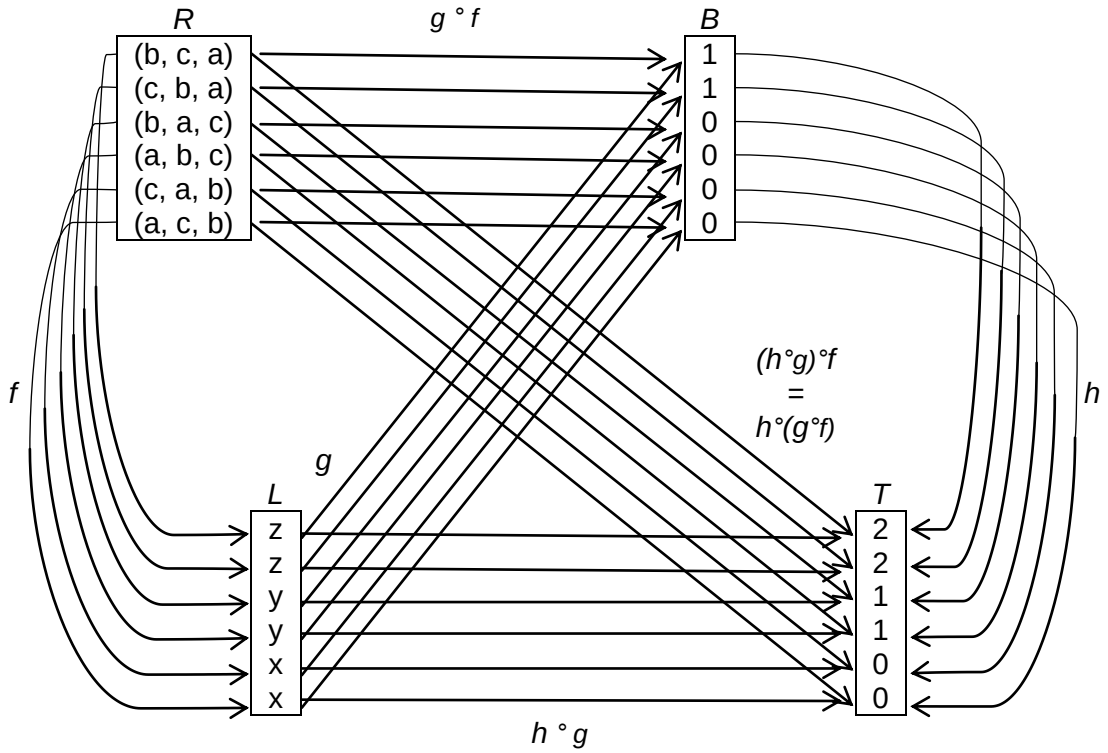


Figure 46. Three function composition (f, g and h) representing proxy single best answer responses via lossy data compression for item 1 with correct answer (b, c, a) when binary and ternary scored. R = Response; L = Letter; B = Binary; T = Ternary; *f*, *g, h* = functions; *g ° f*, *h ° g*, *(h°g)°f*, *h°(g°f)* = compositions of functions.

Figure 47 displays the composition of three functions, f, g and h, starting from ordered triple rank (OTR) responses to item 3 with the correct answer (c, a, b). Lossless data compression, that is, with no loss of information when data is compressed and reconstructed, is applied to the responses that are represented by a three way scoring guide using letter (L), binary (B), and senary (S) scales. These three scoring guide scales represent possible lossless data compression transformations to the dataset so that it can be used in a Winsteps software control file.

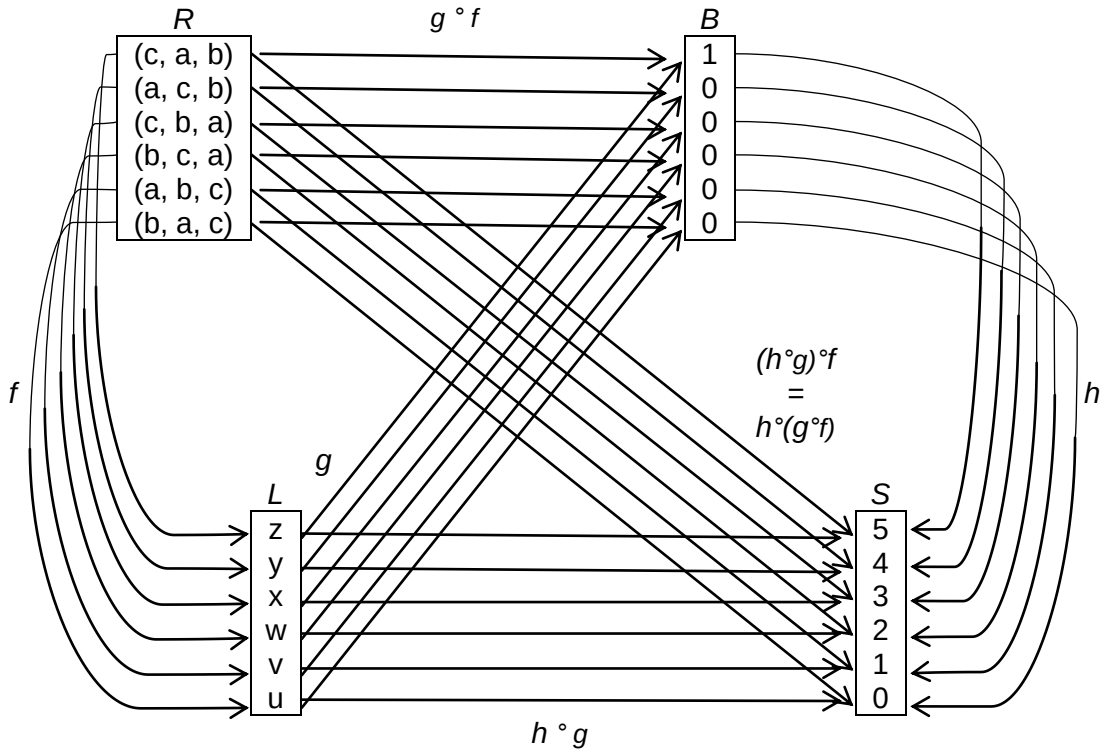


Figure 47. Three function composition (f, g and h) representing ordered triple rank responses via lossless data compression for item 3 with correct answer (c, a, b) when binary and senary scored. R = Response; L = Letter; B = Binary; S = Senary; *f*, *g, h* = functions; *g ° f*, *h ° g*, *(h°g)°f*, *h°(g°f)* = compositions of functions.

Figure 48 displays the composition of three functions, f, g and h, starting from proxy single best answer (SBA) responses to item 3 with the correct answer (c, a, b). Lossy data compression, that is, with some loss of information when data is compressed and reconstructed, is applied to the responses that are represented by a three way scoring guide using letter (L), binary (B) and ternary (T) scales. These three scoring guide scales represent possible lossy data compression transformations to the dataset so that it can be used in a Winsteps software control file.

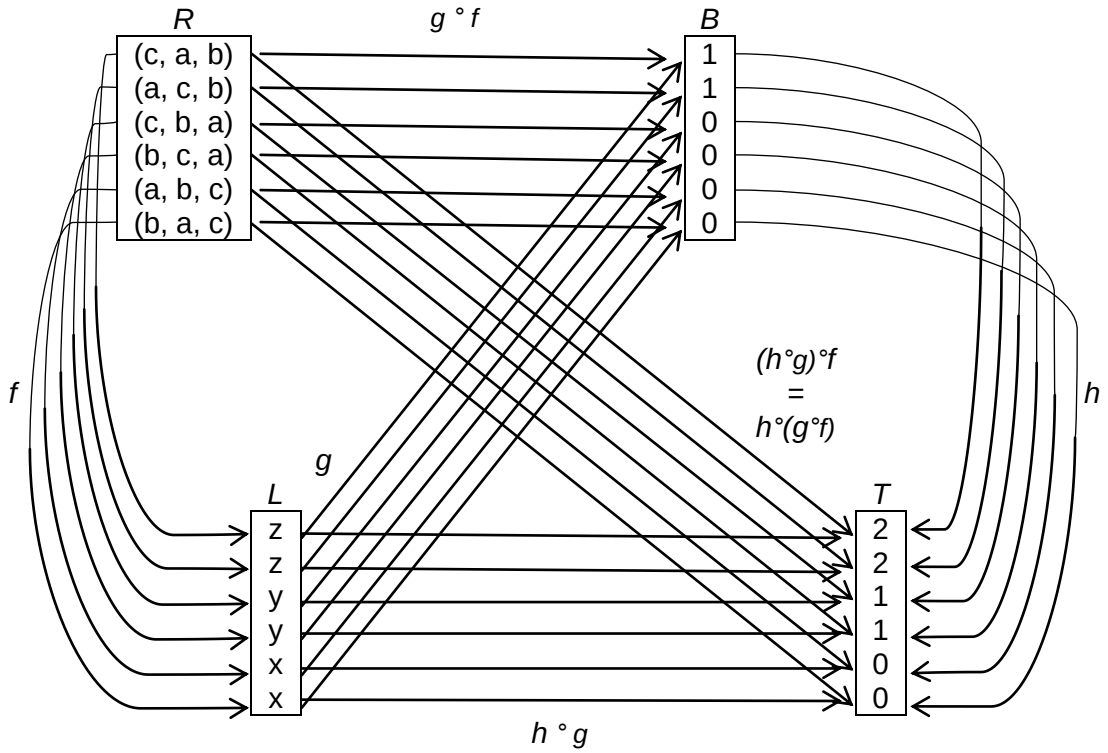


Figure 48. Three function composition (f, g, and h) representing proxy single best answer responses via lossy data compression for item 3 with correct answer (c, a, b) when binary and ternary scored. R = Response; L = Letter; B = Binary; T = Ternary; *f*, *g, h* = functions; *g ° f*, *h ° g*, *(h°g)°f*, *h°(g°f)* = compositions of functions.

Figure 49 displays the composition of three functions, f, g and h, starting from ordered triple rank (OTR) responses to item 6 with the correct answer (a, c, b). Lossless data compression, that is, with no loss of information when data is compressed and reconstructed, is applied to the responses that are represented by a three way scoring guide using letter (L), binary (B), and senary (S) scales. These three scoring guide scales represent possible lossless data compression transformations to the dataset so that it can be used in a Winsteps software control file.

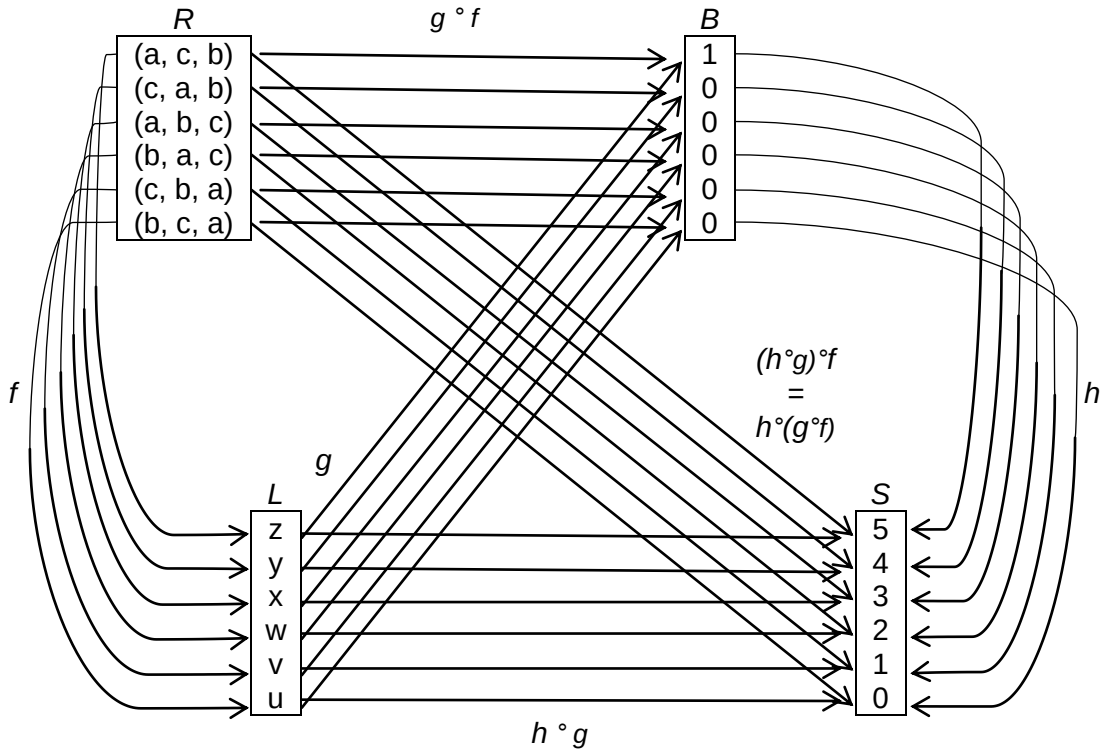


Figure 49. Three function composition (f, g and h) representing ordered triple rank responses via lossless data compression for item 6 with correct answer (a, c, b) when binary and senary scored. R = Response; L = Letter; B = Binary; S = Senary; *f*, *g, h* = functions; *g ° f*, *h ° g*, *(h°g)°f*, *h°(g°f)* = compositions of functions.

Figure 50 displays the composition of three functions, f, g and h, starting from proxy single best answer (SBA) responses to item 6 with the correct answer (a, c, b). Lossy data compression, that is, with some loss of information when data is compressed and reconstructed, is applied to the responses that are represented by a three way scoring guide using letter (L), binary (B), and ternary (T) scales. These three scoring guide scales represent possible lossy data compression transformations to the dataset so that it can be used in a Winsteps software control file.

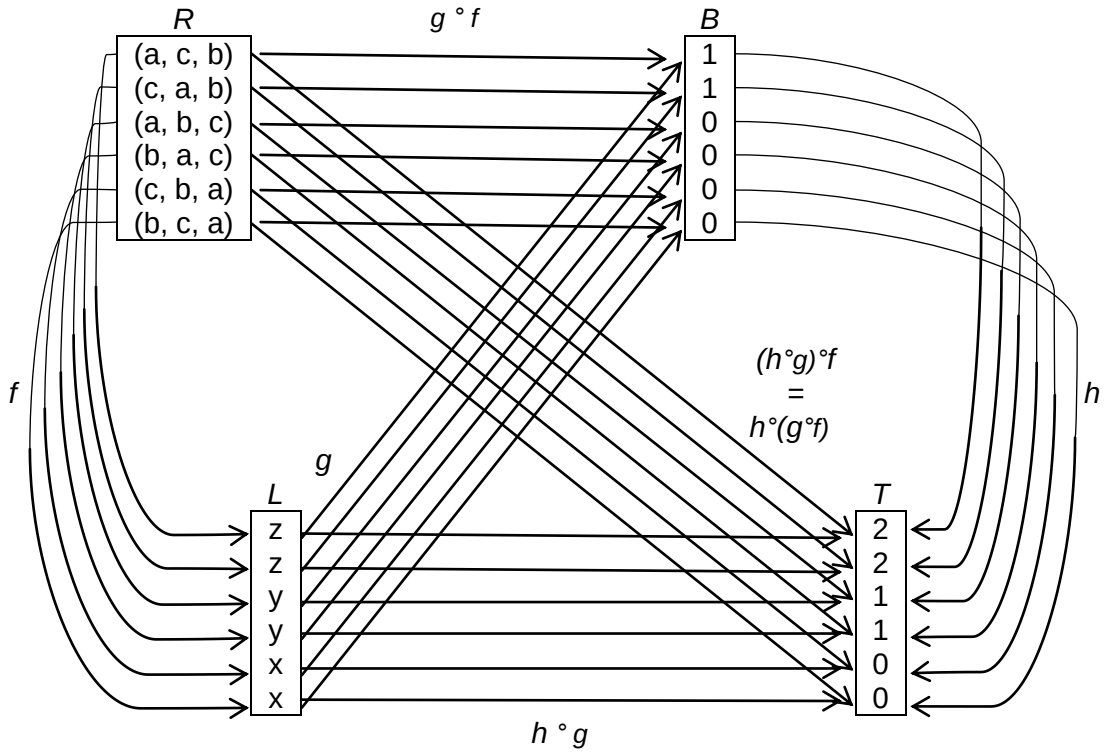


Figure 50. Three function composition (f, g and h) representing proxy single best answer responses via lossy data compression for item 6 with correct answer (a, c, b) when binary and ternary scored. R = Response; L = Letter; B = Binary; T = Ternary; *f*, *g, h* = functions; *g ° f*, *h ° g*, *(h°g)°f*, *h°(g°f)* = compositions of functions.

Figure 51 displays the composition of three functions, f, g and h, starting from ordered triple rank (OTR) responses to item 4 with the correct answer (b, a, c). Lossless data compression, that is, with no loss of information when data is compressed and reconstructed, is applied to the responses that are represented by a three way scoring guide using letter (L), binary (B), and senary (S) scales. These three scoring guide scales represent possible lossless data compression transformations to the dataset so that it can be used in a Winsteps software control file.

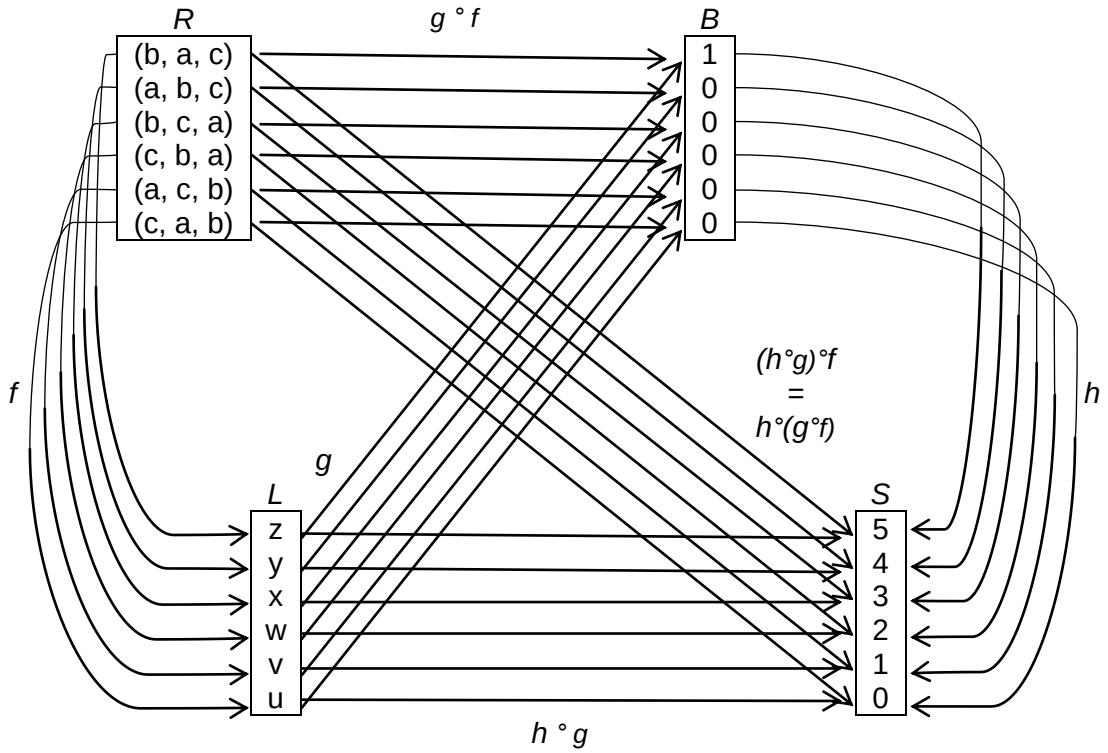


Figure 51. Three function composition (f, g and h) representing ordered triple rank responses via lossless data compression for item 4 with correct answer (b, a, c) when binary and senary scored. R = Response; L = Letter; B = Binary; S = Senary; *f*, *g, h* = functions; *g ° f*, *h ° g*, *(h°g)°f*, *h°(g°f)* = compositions of functions.

Figure 52 displays the composition of three functions, f, g and h, starting from proxy single best answer (SBA) responses to item 4 with the correct answer (b, a, c). Lossy data compression, that is, with some loss of information when data is compressed and reconstructed, is applied to the responses that are represented by a three way scoring guide using letter (L), binary (B), and ternary (T) scales. These three scoring guide scales represent possible lossy data compression transformations to the dataset so that it can be used in a Winsteps software control file.

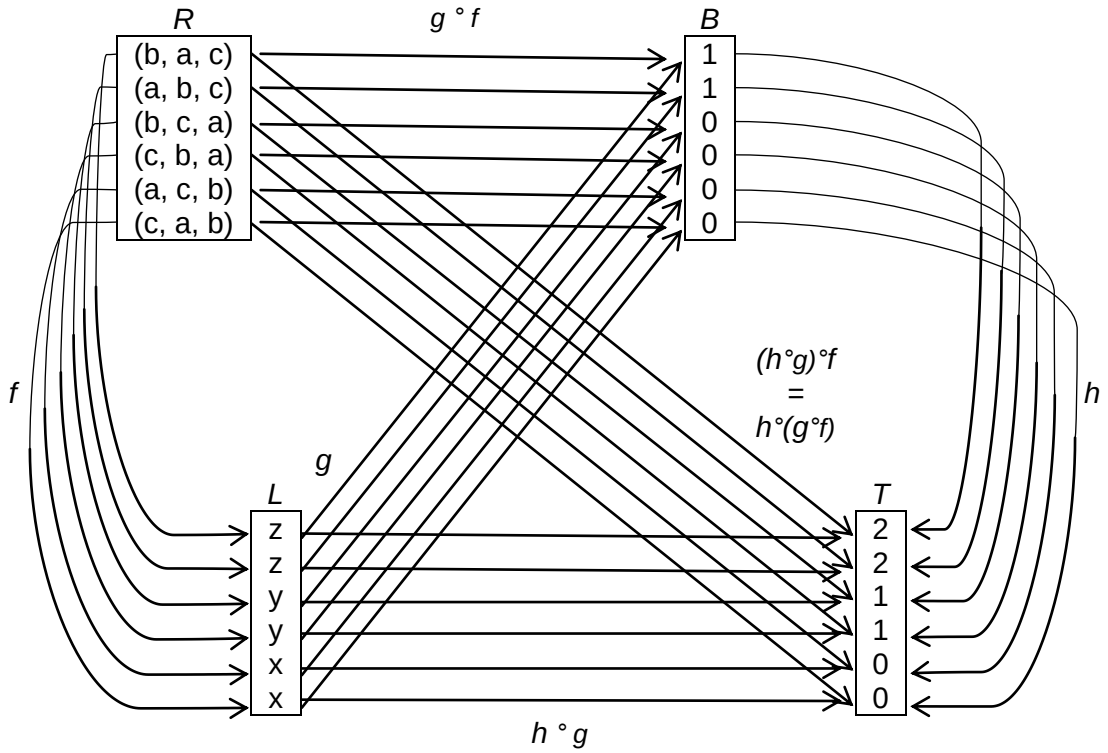


Figure 52. Three function composition (f, g, and h) representing proxy single best answer responses via lossy data compression for item 4 with correct answer (b, a, c) when binary and ternary scored. R = Response; L = Letter; B = Binary; T = Ternary; *f*, *g, h* = functions; *g ° f*, *h ° g*, *(h°g)°f*, *h°(g°f)* = compositions of functions.

Figure 53 displays the composition of three functions, f, g and h, starting from ordered triple rank (OTR) responses to item 2 with the correct answer (a, b, c). Lossless data compression, that is, with no loss of information when data is compressed and reconstructed, is applied to the responses that are represented by a three way scoring guide using letter (L), binary (B), and senary (S) scales. These three scoring guide scales represent possible lossless data compression transformations to the dataset so that it can be used in a Winsteps software control file.

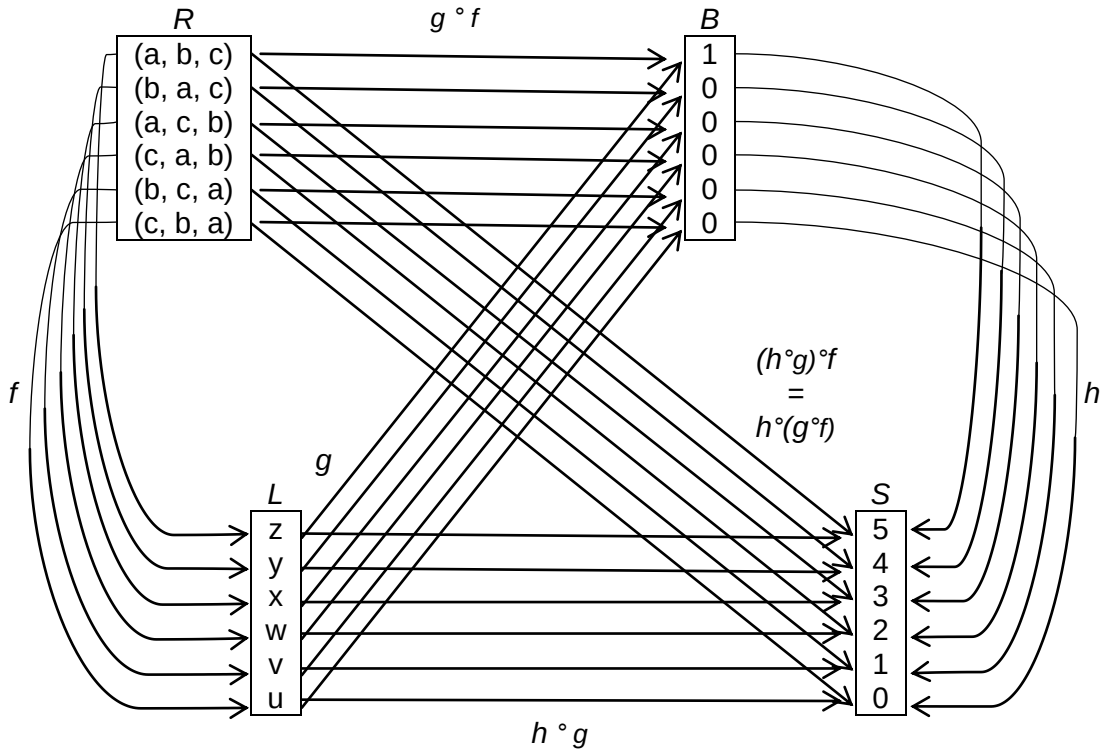


Figure 53. Three function composition (f, g, and h) representing ordered triple rank responses via lossless data compression for item 2 with correct answer (a, b, c) when binary and senary scored. R = Response; L = Letter; B = Binary; S = Senary; *f*, *g, h* = functions; *g ° f*, *h ° g*, *(h°g)°f*, *h°(g°f)* = compositions of functions.

Figure 54 displays the composition of three functions, f, g and h, starting from proxy single best answer (SBA) responses to item 2 with the correct answer (a, b, c). Lossy data compression, that is, with some loss of information when data is compressed and reconstructed, is applied to the responses that are represented by a three way scoring guide using letter (L), binary (B), and ternary (T) scales. These three scoring guide scales represent possible lossy data compression transformations to the dataset so that it can be used in a Winsteps software control file.

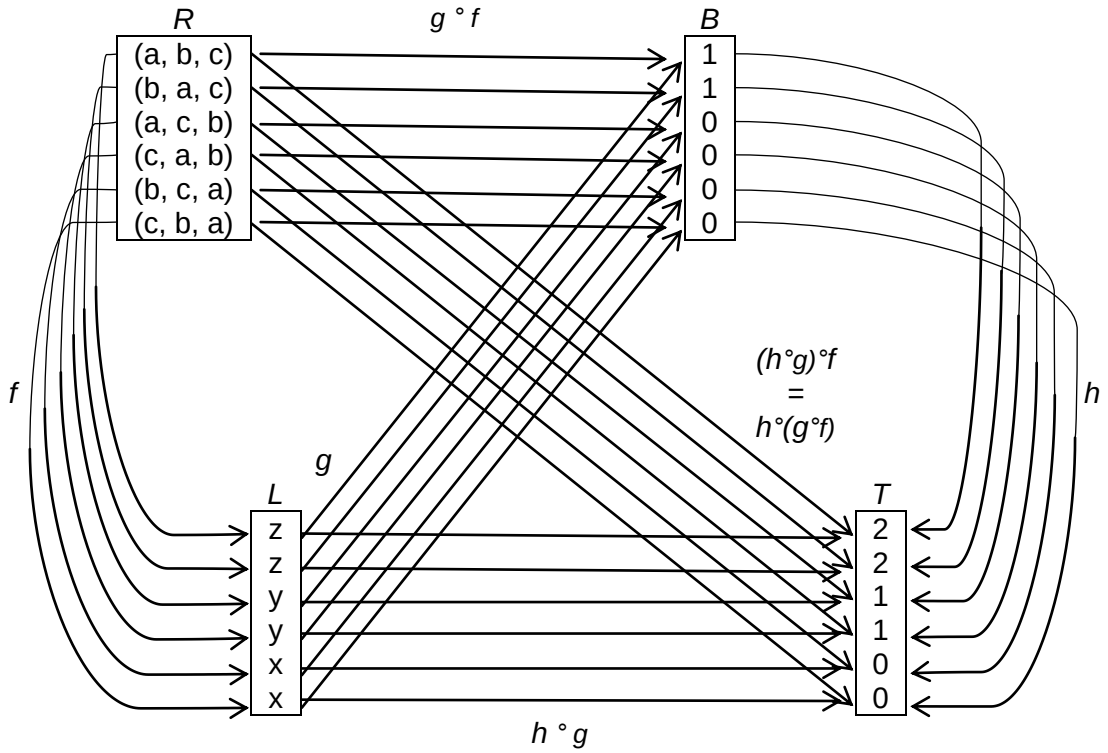


Figure 54. Three function composition (f, g, and h) representing proxy single best answer responses via lossy data compression for item 2 with correct answer (a, b, c) when binary and senary scored. R = Response; L = Letter; B = Binary; T = Ternary; *f*, *g, h* = functions; *g ° f*, *h ° g*, *(h°g)°f*, *h°(g°f)* = compositions of functions.

Data compression of ordered triple rank (OTR) responses from the first six people in the dataset is examined more closely in order to clearly illustrate the data transformation process. The ordered triple rank responses from the first six persons are displayed in Table 36. One example of correct answer permutation, for example, (a, b, c) is chosen from the beginning, items 1 and 2, the middle, items 16 and 17, and the end, items 35 and 36, of the test. These three groups of items are hypothesized to represent easy, medium and difficult items on a scale of item difficulty.

Table 36

Ordered Triple Rank (OTR) Responses from First Six Persons for Items 1, 2, 16, 17, 35, and 36

	1	2	16	17	35	36
Person	(b, c, a)	(a, b, c)	(c, b, a)	(c, a, b)	(b, a, c)	(a, c, b)
A	(b, c, a)	(b, a, c)	(c, b, a)	(a, c, b)	(b, a, c)	(c, b, a)
B	(c, b, a)	(a, b, c)	(c, b, a)	(c, a, b)	(b, a, c)	(b, c, a)
C	(b, c, a)	(a, b, c)	(c, a, b)	(c, a, b)	(b, a, c)	(b, a, c)
D	(b, c, a)	(a, b, c)	(a, c, b)	(a, c, b)	(b, a, c)	(b, a, c)
E	(c, b, a)	(a, b, c)	(b, c, a)	(c, a, b)	(a, b, c)	(a, c, b)
F	(b, c, a)	(a, b, c)	(b, c, a)	(a, c, b)	(b, a, c)	(a, b, c)

After lossless data compression using senary scoring where u = 0, v = 1, w = 2, x = 3, y = 4 and z = 5, the transformed dataset is displayed in Table 37.

Table 37

Ordered Triple Rank (OTR) Responses from First Six Persons via Lossless Data Compression Senary Scoring (u = 0, v = 1, w = 2, x = 3, y = 4, z = 5) for the Rating Scale Model Dataset in Winsteps

	1	2	16	17	35	36
Person	(b, c, a)	(a, b, c)	(c, b, a)	(c, a, b)	(b, a, c)	(a, c, b)
A	5	4	5	4	5	1
B	4	5	5	5	5	0
C	5	5	3	5	5	2
D	5	5	2	4	5	2
E	4	5	4	5	4	5
F	5	5	4	4	5	3

After lossless data compression using binary scoring where u=0, v = 0, w = 0, x = 0, y = 0, and z = 1, the transformed dataset is displayed in Table 38.

Table 38

Ordered Triple Rank (OTR) Responses from First Six Persons via Lossless Data *Compression Binary Scoring (u = 0, v = 0, w = 0, x = 0, y = 0, z = 1) for the Dichotomous Rasch Model Dataset in Winsteps*

Person	1 (b, c, a)	2 (a, b, c)	16 (c, b, a)	17 (c, a, b)	35 (b, a, c)	36 (a, c, b)
A	1	0	1	0	1	0
B	0	1	1	1	1	0
C	1	1	0	1	1	0
D	1	1	0	0	1	0
E	0	1	0	1	0	1
F	1	1	0	0	1	0

The values in the cells of Tables 37 and 38 can be compared to those of Tables 39 and 40 to conceptualize the differences between lossless and lossy data compression for ordered triple rank (OTR) and single best answer (SBA) proxy responses.

After lossy data compression using ternary scoring where x = 0, y = 1, and z = 2, the transformed dataset is displayed in Table 39.

Table 39

Single Best Answer (SBA) Proxy Response from First Six Persons via Lossy Data Compression Ternary Scoring (x = 0, y = 1, z = 2) for the Rating Scale Model Dataset in Winsteps

Person	1 (b, c, a)	2 (a, b, c)	16 (c, b, a)	17 (c, a, b)	35 (b, a, c)	36 (a, c, b)
A	2	2	2	2	2	0
B	2	2	2	2	2	0
C	2	2	1	2	2	1
D	2	2	1	2	2	1
E	2	2	2	2	2	2
F	2	2	2	2	2	1

After lossy data compression using binary scoring where x = 0, y = 0, and z = 1, the transformed dataset is displayed in Table 40.

Table 40

Single Best Answer (SBA) Proxy Response from First Six Persons via Lossy Data Compression Binary Scoring (x = 0, y = 0, z = 1) for the Dichotomous Rasch Model Dataset in Winsteps

Person	1 (b, c, a)	2 (a, b, c)	16 (c, b, a)	17 (c, a, b)	35 (b, a, c)	36 (a, c, b)
A	1	1	1	1	1	0
B	1	1	1	1	1	0
C	1	1	0	1	1	0
D	1	1	0	1	1	0
E	1	1	1	1	1	1
F	1	1	1	1	1	0

The shift from lossless to lossy data compression, that is, from ordered triple rank (OTR) to single best answer (SBA) proxy responses, involves potential loss of measurement information.

Table 41 displays the consolidation of data compression information from Tables 36, 37, 38, 39, and 40. Ordered triple rank (OTR) responses undergo lossless data compression via senary and binary scales and lossy data compression via ternary and binary scales. As one of the goals of Rasch paradigm measurement is to spread person ability measures over a wide a range as possible along the latent construct, lossless data compression is hypothesized as the preferred method of constructing person ability measures. The rationale for preferring lossless data compression is the retention of an optimal amount of measurement information in the raw scores before the Rasch model transformation of

Table 41

Ordered Triple Rank (OTR) Responses and Respective Data Compression Scores from First Six Persons for Items 1, 2, 16, 17, 35, and 36

				Correct	Answer		
		1	2	16	17	35	36
	Person	(b, c, a)	(a, b, c)	(c, b, a)	(c, a, b)	(b, a, c)	(a, c, b)
Response		OTR					
	A	(b, c, a)	(b, a, c)	(c, b, a)	(a, c, b)	(b, a, c)	(c, b, a)
	B	(c, b, a)	(a, b, c)	(c, b, a)	(c, a, b)	(b, a, c)	(b, c, a)
	C	(b, c, a)	(a, b, c)	(c, a, b)	(c, a, b)	(b, a, c)	(b, a, c)
	D	(b, c, a)	(a, b, c)	(a, c, b)	(a, c, b)	(b, a, c)	(b, a, c)
	E	(c, b, a)	(a, b, c)	(b, c, a)	(c, a, b)	(a, b, c)	(a, c, b)
	F	(b, c, a)	(a, b, c)	(b, c, a)	(a, c, b)	(b, a, c)	(a, b, c)
Lossless		Senary					
	A	5	4	5	4	5	1
	B	4	5	5	5	5	0
	C	5	5	3	5	5	2
	D	5	5	2	4	5	2
	E	4	5	4	5	4	5
	F	5	5	4	4	5	3
Lossless		Binary					
	A	1	0	1	0	1	0
	B	0	1	1	1	1	0
	C	1	1	0	1	1	0
	D	1	1	0	0	1	0
	E	0	1	0	1	0	1
	F	1	1	0	0	1	0
Lossy		Ternary					
	A	2	2	2	2	2	0
	B	2	2	2	2	2	0
	C	2	2	1	2	2	1
	D	2	2	1	2	2	1
	E	2	2	2	2	2	2
	F	2	2	2	2	2	1
Lossy		Binary					
	A	1	1	1	1	1	0
	B	1	1	1	1	1	0
	C	1	1	0	1	1	0
	D	1	1	0	1	1	0
	E	1	1	1	1	1	1
	F	1	1	1	1	1	0

ordinal level raw scores into interval level item difficulty calibrations and person ability measures.

This concludes the protocol for the consolidated building blocks of Items Design and Outcome Space. The next building block is the Measurement Model.

Measurement Model

The Rasch model measurement family is utilized in this study. Extensively, the polytomously scored Rasch-Andrich Rating Scale Model developed by Andrich (1978) is applied. However, there is also some use of the dichotomously scored Rasch Model developed by Rasch (1960).

As the purpose of this study is to construct a polysemous academic vocabulary extent test measuring both aspects of vocabulary size (breadth) and strength (depth) where all 36 polytomously-scored items are designed with one underlying theory of synonymy (O'Connor, 1940), the Rating Scale Model is suitable.

In the initial analyses, the vocabulary strength aspect is the primary focus as this has been problematic in the research literature. In fact, Smith (1987) is the only researcher to attempt measuring polytomously scored vocabulary strength by applying the Rating Scale Model of the Rasch family. However, Smith's (1987) attempt to measure vocabulary strength was hampered when disordered thresholds violated the results.

Referring back to the beginning of this chapter and the first of the four building blocks, one is reminded that the construct map should provide an a priori expectation of how the measurement model will be manifested. In this polytomously-scored case, the

measurement model is visualized via Winsteps software (Linacre, 2015b) Table 3.2 summary of category structure and the related category characteristic response curves.

In an attempt to reduce the possibility of disordered thresholds that would violate results, a series of three pilot studies were conducted during the 2012 and 2013 university academic years before the main study, which was conducted over the 2013 and 2014 academic years.

The first pilot study examined 39 Japanese tertiary level students with 36 items as displayed in Table 42. The responses were scored using a 012345 scale and the Rasch-Andrich rating scale model was applied for analyses. As displayed in Table 42, the percentage of responses that succeeded on the most difficult category, Synonymy Upper, was 49%, causing a ceiling effect. Therefore, this first pilot test was very easy for this sample of persons as almost half of the responses were fully correct at the Synonymy Upper level. These high ability responses did not leave much room for the middle categories to display pronounced unimodal distributions in the Category Characteristic

Table 42

Pilot Study 1 Summary of Category Structure

Cat lab	Category name	Observed count (%)	Observed average	Sample expect	Infit MNSQ	Outfit MNSQ	Andrich threshold	Category measure
0	S/L Like Lower	68 (5)	.14	.14	1.01	1.07	NONE	(-1.68)
1	S/L Like Upper	105 (7)	.31	.25	1.08	1.14	-0.24	-0.63
2	Context Lower	69 (5)	.31*	.37	0.89	0.92	0.73	-0.15
3	Context Upper	282 (20)	.54	.50	1.08	0.94	-0.98	0.21
4	Synonym Lower	185 (13)	.49*	.68	1.65	0.89	1.01	0.67
5	Synonym Upper	691 (49)	.98	.95	.95	0.97	-0.51	(1.57)

Note. Cat lab = category label; Sample expect = sample expectation.

Curve map. Both the observed averages and the Andrich thresholds in Table 42 are disordered.

The second pilot study examined 38 Japanese tertiary level students with 51 items as displayed in Table 43. The responses were scored using a 012345 scale and the Rasch-Andrich rating scale model was applied for analyses. As displayed in Table 43, the percentage of responses that succeeded on the most difficult category, Synonymy Upper, was 52%, causing a ceiling effect. Therefore, this second pilot test was very easy for this sample of persons as over half of the responses were fully correct at the Synonymy Upper level. These high ability responses did not leave much room for the middle categories to display pronounced unimodal distributions in the Category Characteristic Curve map. Both the observed averages and the Andrich thresholds in Table 43 are disordered.

Table 43

Pilot Study 2 Summary of Category Structure

Cat lab	Category name	Observed count (%)	Observed average	Sample expect	Infit MNSQ	Outfit MNSQ	Andrich threshold	Category measure
0	S/L Like Lower	95 (5)	.32	.14	1.23	2.84	NONE	(-1.57)
1	S/L Like Upper	132 (7)	.18*	.23	0.89	0.82	-0.15	-0.57
2	Context Lower	65 (4)	.35	.34	1.02	1.08	0.99	-0.15
3	Context Upper	251 (14)	.47	.48	0.96	0.96	-0.95	0.18
4	Synonym Lower	308 (17)	.52	.69	1.21	0.59	0.37	.60
5	Synonym Upper	935 (52)	1.10	1.05	0.87	0.92	-0.26	(1.51)

Note. Cat lab = category label; Sample expect = sample expectation.

The third pilot study examined 36 Japanese tertiary level students with 48 items as displayed in Table 44. The responses were scored using a 012345 scale and the Rasch-Andrich rating scale model was applied for analyses. As displayed in Table 44, the

percentage of responses that succeeded on the most difficult category, Synonymy Upper, was 42%, causing a ceiling effect. Therefore, this third pilot test was very easy for this sample of persons as about two-fifths of the responses were fully correct at the Synonymy Upper level. These high ability responses did not leave much room for the middle categories to display pronounced unimodal distributions in the Category Characteristic Curve map. Although the observed averages are ordered, the Andrich thresholds in Table 44 are disordered.

Table 44

Pilot Study 3 Summary of Category Structure

Cat lab	Category name	Observed count (%)	Observed average	Sample expect	Infit MNSQ	Outfit MNSQ	Andrich threshold	Category measure
0	S/L Like Lower	137 (8)	.05	.03	1.05	1.14	NONE	(-1.54)
1	SL/LL Upper	129 (7)	.15	.12	1.05	1.15	0.13	-0.62
2	Context Lower	159 (9)	.15	.22	0.80	0.63	-0.04	-0.18
3	Context Upper	224 (13)	.39	.35	1.08	1.46	-0.06	0.16
4	Synonym Lower	357 (21)	.51	.54	1.06	0.87	-0.03	0.61
5	Synonym Upper	722 (42)	.91	.91	0.97	0.98	0.00	(1.60)

Note. Cat lab = category label; Sample expect = sample expectation.

After the third pilot study involving Japanese non-native speakers of English, a group of 17 anonymous native speakers of English provided some additional and insightful feedback. Think-aloud protocol interviews enabled 14 of the 17 native speakers of English to orally comment in real time on their written responses to the individual item options on various test versions. As 3 of the 17 native speakers of English lived outside of Tokyo, Japan, geographical area, exit protocol interviews enabled them to orally comment via a telephone call after completing their written responses to the individual item options on various test versions. Think-aloud and exit protocol interviews are often used in qualitative

research as well as product design and development. However, in this study's case this valuable feedback was used to guide quantitative test items design and outcome space development. Consolidation of these native speakers of English comments assisted in the final decision of which key words and options would be included in the main study.

Before the main study administration, the participants were given 3 weekly quizzes in the preceding month. Each quiz consisted of 9 relatively easy bolded key word items containing 3 nouns, 3 verbs and 3 adjectives. Each subsequent administration after Quiz 1 included feedback of the correct answers of the previous quiz. The purpose of the quizzes was to get the participants familiar with the atypical nature of the new outcome space format of the test. This unorthodox format has a printed 3-by-3 matrix within which the participants must select an ordered triple rank response, that is, an ordered a, b, c response without repetition (see Appendix D). In contrast to usual testing experiences, the participants are familiar with the typical outcome space format of multiple-choice tests where there is a single row of options and they must select a single best answer.

In the main study, the test was administered to 403 test-takers (24 of whom were non-Japanese). As the underlying theories of the test were consolidated to investigate Japanese tertiary institution student's use of English loanwords, word frequency and synonymy acquisition, the 24 non-Japanese persons were eliminated from the study leaving 379 Japanese persons.

Quality Control of Baseline Data Fit to Model

An initial investigation of the main study's item and person misfit analyses using Winsteps (Linacre, 2015b) software output was conducted. Misfit analyses are helpful in the quality control process because they "indicate how accurately or predictably data fit the model" (Linacre, 2002, p. 878). Linacre (2015b) recommends when reporting misfit diagnosis that only Outfit should be documented. Misfit statistics targeting Infit values are not the primary focus as "these patterns can be hard to diagnose and remedy" (Linacre, 2002, 878).

However, as an additional criterion in this study Infit misfit statistics do provide a secondary and stringent data quality control check. Initially, the criterion of Outfit MNSQ from 0.5 to 1.5 was mandated as acceptable (Linacre, 2002). As overfitting items "rarely distort the measures enough to have any practical consequences," (Linacre, 2000, 755) misfit below 0.5 was not a concern. Therefore, only underfit, that is, the noisy off-dimensional misfit greater than Outfit and Infit MNSQ of 1.50, was targeted as the primarily important focus of the quality control process (Boone, Staver & Yale, 2014). Thus, only Outfit and Infit MNSQ underfit greater than 1.5 are reported. First items and then persons were examined for misfit.

A quality control procedure that diagnoses and corrects underfitting misfit items in the data is outlined by Boone, Staver, and Yale (2014). Following this procedure, two items, item 5 (Outfit MNSQ of 1.67) and item 3 (Oufit MNSQ of 1.57) were diagnosed as displaying underfitting misfit. All 27 person misfitting responses for item 5 and all 27 person misfitting responses for item 3 were identified as having negative *z*-residual values

of more than -2.00 in Winsteps (2015b) Table 11.1 (Table of Poorly Fitting Items). Therefore, these responses were in lower rating scale categories than those expected by the Rasch model (Linacre, 2015a). In other words, these responses were a result of carelessness. In contrast, if the responses had a positive z-residual of 2.00 or more they would have an inverse diagnosis. That is, they would be in higher rating scale categories than those expected by the Rasch model (Linacre 2015a). In other words, responses with highly positive *z*-residuals would be identified as being due to guessing. Indeed, returning to the issue of careless responses, item 3 was the easiest item and item 5 was the second easiest item of the 36 items in the test. As a way of reconfirmation, these 2 items were the two easiest in this test, therefore, a test analysis would be expected to mark these misfitting responses as indications of carelessness rather than guessing. The 27 person misfitting responses to item 5 and the 27 person misfitting responses to item 3 were replaced by missing data responses using an asterisk symbol in a new data file and run through Winsteps again. This time, none of the 36 items displayed underfitting Oufit MNSQ values.

Next, a diagnosis of the person ability measure misfit was analyzed. The diagnosis and quality control of the person ability measure misfit encompassed a sequence of seven stages. In Stage 1, 28 person responses (that is, 28 out of 379 persons or 7.39%) displayed underfitting Outfit MNSQ. These responses were changed to missing data asterisk symbols in a new data file and run through Winsteps. In Stage 2, the control file was run again with the updated data and 16 persons (i.e., 16 out of 379 persons or 4.22%) displayed underfitting Outfit MNSQ. These responses were substituted to missing data symbols. In Stage 3, a rerun of the control file indicated that nine persons (that is, 9 out of 379 persons

or 2.37%) displayed underfitting Outfit MNSQ. These responses were substituted for missing data. In Stage 4, the output from the new control file indicated that 8 persons (that is, 8 out of 379 or 2.11%) displayed underfitting Outfit MNSQ. These responses were adjusted accordingly. In Stage 5, the next control file's output indicated that three persons (3 out of 379 persons or 0.79%) displayed underfitting Outfit MNSQ. These were changed again. In Stage 6, the subsequent control file's output indicated that two persons (that is, 2 out of 379 persons or 0.53%) displayed underfitting Outfit MNSQ. Finally, in Stage 7 the last control file's output indicated that 0 persons (i.e., 0 out of 379 persons or 0.00%) displayed underfitting Outfit MNSQ. As McNamara (1996) asserted, person misfit is a very important test development issue and person misfit should be under 2% of test-takers if tests are to be workable for measurement.

Subsequently, the Infit MNSQ quality control was checked. All 36 items were within the acceptable Infit MNSQ range (under 1.5). Five persons, that is, F401 (1.70), A216 (1.62), C331 (1.59), B036 (1.58) and C198 (1.52) displayed Infit MNSQ misfit. From Winsteps Table 7.1, the responses displaying misfit were identified and replaced with missing data symbols in the dataset. The dataset was rerun through Winsteps software. All 36 items were within the acceptable Infit MNSQ range (under 1.5). Two persons, that is, F401 (1.62) and C196 (1.51) displayed Infit MNSQ misfit. From Winsteps Table 7.1, the responses displaying misfit were identified and replaced with missing data symbols in the dataset. The dataset was rerun through Winsteps software. All 36 items and 379 persons were within the acceptable range (under 1.5). Table 45 displays the summary of category structure for the initial main study.

As indicated in Table 45, the initial main study analysis indicated that 37% of responses were in the most difficult category, Synonymy Upper, giving an initial indication that the main study was more difficult than the three preceding pilot studies. The 37% fully correct responses also caused a ceiling effect. There are six ordered categories (0.05, 0.10, 0.17, 0.38, 0.60, and 0.82) as displayed in the Observed Average column. However, Table 45's Andrich Threshold column indicates that the five thresholds were disordered (-0.11, 0.14, -0.36, 0.10, and 0.24). Andrich (2013a) stated that threshold ordering is essential for statistical sufficiency and therefore meaningful measurement.

Table 45

Initial Main Study Summary of Category Structure

Cat lab	Category name	Observed count (%)	Observed average	Sample expect	Infit MNSQ	Outfit MNSQ	Andrich threshold	Category measure
0	S/L Like Lower	881 (7)	0.05	0.02	1.06	1.04	NONE	(-1.67)
1	SL/LL Upper	1049 (8)	0.10	0.11	0.97	0.93	-0.11	-0.68
2	Context Lower	1077 (8)	0.17	0.23	0.84	0.75	0.14	-0.21
3	Context Upper	2079 (16)	0.38	0.37	1.03	0.94	-0.36	0.16
4	Synonym Lower	2997 (23)	0.60	0.56	0.95	1.01	0.10	0.67
5	Synonym Upper	4707 (37)	0.82	0.84	1.03	1.02	0.24	(1.77)
	Missing	96 (1)						

Note. Cat lab = category label; Sample expect = sample expectation.

Therefore, to attain order of the five thresholds the initial main study's categories were collapsed, that is, the item responses are rescored. The four partially correct categories from the initial main study had the item responses rescored into two incrementally ordered groups. In other words, categories 1 and 2 were rescored to form a new category 1, Partial Non-Synonym. Categories 3 and 4 were also rescored to form a new category, 2, Partial Synonym. Category 0 (incorrect response) and 5 (correct response) were not rescored but

were renamed, Non-Synonym and Synonym, respectively. The result was the initial main study's category scale of 012345 was rescored into a secondary main study category scale of 011223.

The quality control of the RSM011223 scale was checked starting with Outfit MNSQ. All 36 items were within the acceptable Outfit MNSQ range (under 1.5). Four persons, that is, C151 (1.57), B273 (1.53), C302 (1.53) and C301(1.51) displayed Oufit MNSQ misfit. From Winsteps Table 7.1, the responses displaying misfit were identified and replaced with missing data symbols in the dataset. The dataset was rerun through Winsteps software. All 36 items and 379 persons were within the acceptable Outfit MNSQ range (under 1.5). Further, all 36 items and 379 persons were within the acceptable Infit MNSQ range (under 1.5). Table 46 displays the summary of category structure for the secondary main study.

Table 46

Secondary Main Study Summary of Category Structure

Cat lab	Category name	Observed count (%)	Observed average	Sample expect	Infit MNSQ	Outfit MNSQ	Andrich threshold	Category measure
0	Non synonym	875 (7)	0.06	0.00	1.06	1.05	NONE	(-2.18)
1	Partial non-syn	2122 (16)	0.21	0.31	0.86	0.79	-0.74	-0.69
2	Partial Synonym	5074 (39)	0.79	0.74	1.01	1.04	-0.36	0.58
3	Synonym	4707 (37)	1.32	1.34	1.04	1.03	1.10	(2.35)
	Missing	108 (1)						

Note. Partial non-syn = Partial non-synonym; Cat lab = category label; Sample expect = sample expectation.

As displayed in Table 46, the secondary main study scored 011223 indicated both four ordered categories (0.06, 0.21, 0.79 and 1.32) and three ordered thresholds (-0.74,

-0.36 and 1.10). As both categories and thresholds are ordered this analysis is considered viable for subsequent analyses.

Table 47 is a summary of the three pilot studies and one main study that were conducted during this period of time.

Table 47

Pilot and Main Studies Summary

		k	*n*	RSM scale	Cat. #5 %	Category order	Threshold order
Pilot	1	36	39	012345	49	No	No
	2	51	38	012345	52	No	No
	3	48	36	012345	42	Yes	No
Main	4 (a)	36	379	012345	37	Yes	No
	4 (b)	36	379	011223	37	Yes	Yes

Four Building Blocks Summary

In summary, the item specification development process has been defined as following Wilson's (2005) construct modeling approach. Under this approach there are four building blocks that are sequenced in a cyclical process: construct maps, items design, outcome space and measurement model. A summary of the polytomously-scored main study 4 (a) and (b), that is, the recoded main study scale, is outlined in terms of Wilson's (2005) construct modeling approach in Table 48.

Table 48

Construct Modeling Approach to the Item Specification and Construction Process

Construct maps	Item design		Outcome space		Measurement model
	Column	Options	Response	Scoring	
SCL/CSL/SLC/LSC/CLS/LCS	1,2,3	6	OTR	012345	RSM
SCL/CSL+SLC/LSC+CLS/LCS	1,2,3	6	OTR	011223	RSM

Note. S = Synonym; C = Context; L = Sound-/Look-Like; OTR = Ordered Triple Rank; RSM = Rating Scale Model.

Therefore, in this study the Rating Scale Model using the recoded scaling code of 011223 is employed as the baseline for all subsequent analyses.

Research questions 5 and 6, refer to 4 types of analyses. These 4 types of analyses are displayed in Table 47. In Table 49, the initial main study analyses of the Rating Scale Model 012345 scaling displayed disordered thresholds therefore the categories were collapsed whereby item responses were rescored into the Rating Scale Model 011223 scaling. Thus, the Rating Scale Model 011223 scaling formed the new baseline, that is, the first of the four types of analyses. The other three analyses are proxies of the Rating Scale

Table 49

Four Types of Analyses

Response	Categories	Response column	Score scale	Model	Status
SBA	Binary (2)	#3 only	Dichotomous 000011	DRM	Proxy
SBA	Ternary (3)	#3 only	Polytomous 001122	RSM	Proxy
OTR	Binary (2)	#1, #2, #3	Dichotomous 000001	DRM	Proxy
OTR	Quaternary (4)	#1, #2, #3	Polytomous 011223	RSM	Res baseline
OTR	Senary (6)	#1, #2, #3	Polytomous 012345	RSM	Dis threshold

Note. SBA = Single Best Answer; OTR = Ordered Triple Rank; DRM = Dichotomous Rasch Model; RSM = Rating Scale Model; Res Baseline = Rescored Baseline; Dis Threshold = Disordered thresholds.

Model 011223: the Dichotomous Rasch Model 000001 scaling, the Rating Scale Model 012 scaling and the Dichotomous Rasch Model 001 scaling.

In Table 50, the initial main study analyses of the Rating Scale Model 012345 scaling displayed disordered thresholds therefore the categories were collapsed whereby item responses were rescored into the Rating Scale Model 011223 scaling. Thus, the Rating Scale Model 011223 scaling formed the new baseline, that is, the first of the four types of analyses. The other three analyses are proxies of the Rating Scale Model 011223: the Dichotomous Rasch Model 000001 scaling, the Rating Scale Model 012 scaling and the Dichotomous Rasch Model 001 scaling. The four types of analyses can alternatively be displayed as in Table 50. Table 50 additionally displays the relationship that each of the four analyses has to vocabulary acquisition, awareness, size, strength and extent.

Table 50

Four Types of Analyses in a Matrix Design

Vocabulary	Dichotomous	Polytomous
Lossy	Size	Extent
Acquisition	DRM 001 (DRM 000011)	RSM 012 (RSM 001122)
Lossless	Size	Extent
Awareness	DRM 000001	RSM 012345
		RSM 011223 (Rescored Baseline)

Note. DRM = Dichotomous Rasch Model; RSM = Rating Scale Model. 001, 000011, 012, 001122, 000001, 012345 and 011223 = scoring codes.

An alternative way of visualizing the categorization of the measurement information is displayed in Figure 55. Viewed from left to right, the first three columns are representations of ordered triple rank (OTR) responses of vocabulary awareness with

lossless data compression in senary, collapsed quaternary and binary scaling modes. Columns four and five are representations of proxy single best answer (SBA) responses of vocabulary acquisition with lossy data compression in ternary and binary scaling modes.

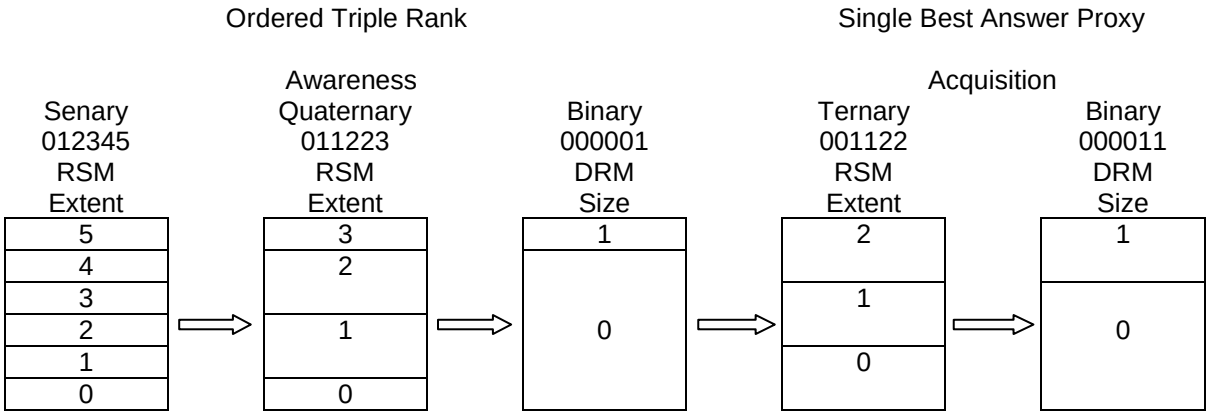


Figure 55. Categorization of available measurement information.

Quality Control of Collapsed Data Fit to Model

Previously in this chapter under the section Quality Control of Baseline Data Fit to the Model, the vocabulary awareness senary and quaternary scales data cleaning procedures were detailed. Following the sequence displayed in Figure 55, the data quality control for the vocabulary awareness binary scale and vocabulary acquisition ternary and binary scales are checked and cleaned.

The quality control of collapsed data fit to the Rasch model was checked for the vocabulary awareness binary scale of 000001. All 36 item values were within the misfit criteria of 1.50 for Outfit and Infit MNSQ. However, 19 persons displayed Outfit and/or Infit underfit exceeding 1.5. Three iterations of data cleaning were conducted, that is, these

persons misfitting responses were replaced with missing data symbols. Finally, all 379 persons indicated underfit within the acceptable levels.

The quality control of collapsed data fit to the Rasch model was checked for the vocabulary acquisition ternary scale of 001122. Two item values and 28 person values were outside the misfit criteria of Outfit and/or Infit underfit exceeding 1.5. Eight iterations of data cleaning were conducted, that is, these persons misfitting responses were replaced with missing data symbols. Finally, all 36 items and 379 persons indicated underfit within the acceptable levels.

The quality control of collapsed data fit to the Rasch model was checked for the vocabulary acquisition binary scale of 000011. All 36 item values were within the misfit criteria of 1.50 for Outfit and Infit MNSQ. However, 23 persons displayed Outfit and/or Infit underfit exceeding 1.5. Nine iterations of data cleaning were conducted, that is, these persons misfitting responses were replaced with missing data symbols. Finally, all 379 persons indicated underfit within the acceptable levels.

Quality Control of Additional Data Fit to Model

In addition to the Polysemous Academic Extent Test (PAVET), the test-takers completed a version of the Vocabulary Size Test (VST). The rationale underlying the use of the Vocabulary Size Test was to provide a preliminary check of the concurrent validity of inferences and interpretations to be made from the Polysemous Academic Vocabulary Extent Test. The version of the Vocabulary Size Test was amended from the 2000, 3000, 4000, 5000, 6000, 7000 and 8000 word levels and therefore contained 70 items. In other

words, as in the original Vocabulary Size Test, each word frequency level contained 10 items. In contrast to the Polysemous Academic Vocabulary Extent Test with 3 options per item, the Vocabulary Size Test has four options per item.

The quality control of data fit to the Rasch model was checked for the Vocabulary Size Test binary scale of 0001. All 36 item values were within the misfit criteria of 1.50 for Outfit and Infit MNSQ. However, 27 persons displayed Outfit and/or Infit underfit exceeding 1.5. Ten iterations of data cleaning were conducted, that is, these persons misfitting responses were replaced with missing data symbols. Finally, all 379 persons indicated underfit within the acceptable levels.

Examples of the Polytomous Academic Vocabulary Extent Test (PAVET) and optical mark recognition (OMR) mark sheet are in Appendices C and D. This concludes the chapter on the item specification and construction process.

CHAPTER 5

RESULTS

In this Results chapter, the research questions related to the polysemous academic vocabulary extent test (PAVET) are addressed in sequential order.

Research Question 1: Rasch Analysis Requirements

Research question 1, Does the test meet the unidimensional Rasch model requirements?, is addressed using the RSM 011223 scale baseline via the Rasch-Andrich rating scale model. The steps outlined by Iramaneerat, Smith, and Smith (2008) are investigated sequentially.

Step 1 involved checking the internal consistency reliability. The Rasch person ability measure separation value was 1.24 and the related real reliability was .61. However, the main focus of this study was the items. The Rasch item difficulty calibration separation value was 10.18 and the related real reliability was .99. Table 51 displays a comparison between the four analyses types. The baseline Rasch model scale, RSM 011223, shows the greatest separation for both persons, 1.24, with a corresponding reliability of .61, and items, 10.18, with a corresponding reliability (precision) of .99.

In Step 2 the characteristics of the two Wright maps in Figures 56 and 57, in which the distributions of person ability measures and item difficulty calibrations are located on a common latent construct variable scale, were investigated for item order hierarchy and targeting. In Figure 56, the person ability measure distribution is on the left side while the

Table 51

Person and Item Separation Statistics for the Four Types of Analyses

Rasch Model Scale	Person Sep	PS Reliability	Item Sep	IS Reliability
DRM 000011	1.02	.51	7.62	.98
RSM 001122	1.05	.52	7.59	.98
DRM 000001	1.12	.56	8.10	.98
RSM 011223	1.24	.61	10.18	.99

Note. DRM = Dichotomous Rasch Model; RSM = Rating Scale Model; 001, 000011, 012, 001122, 000001, 012345 and 011223 = scoring codes; Person Sep = Person Separation; PS Reliability = Person Separation Reliability; Item Sep = Item Separation; IS Reliability = Item Separation Reliability.

item difficulty calibration distribution is on the right side. Persons with higher ability are located at the top left of Figure 56 and persons with lower ability are located at the bottom left. Items with higher difficulty are located at the top right of Figure 56 and items with lower difficulty are located at the bottom right of the figure. In the dichotomous case (with only one threshold), if a person and an item are aligned at the same location on the logit scale, then that person has a 50% probability of responding to that item correctly. However, in this study's polytomous analysis with three thresholds, if a person and an item are aligned at the same location on the central logit scale in Figure 56, then that person has a 50% probability of responding to that item's second threshold correctly. In Figure 56 the location of the person distribution appears slightly higher than the item difficulty calibration distribution due to the item difficulty calibration distribution being the second threshold difficulty distribution.

The targeting of the persons and items in Figure 56 appears to be mismatched; however, as this study deals with polytomous scaling, the three distributions of item difficulty calibration distributions displayed in Figure 57 are more illustrative of the

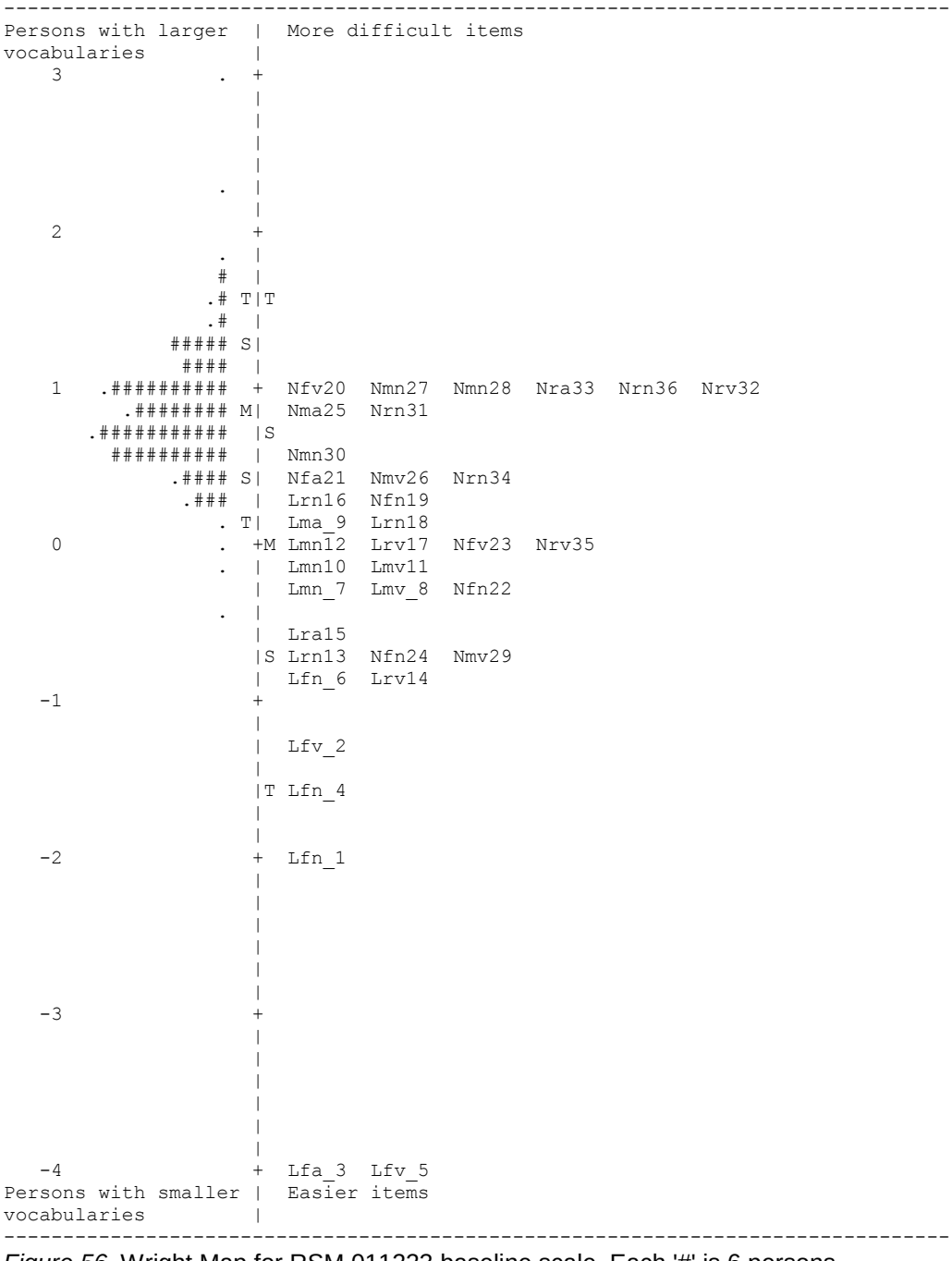


Figure 56. Wright Map for RSM 011223 baseline scale. Each '#' is 6 persons.

relationship between the person ability and item difficulty estimates. The left item difficulty calibration distribution displays the 50% probability of success of responding to that item's

first threshold correctly (i.e., the threshold between categories 0 and 1) correctly. The middle item difficulty calibration distribution displays the 50% probability of responding to that item's second threshold (i.e., the threshold between categories 1 and 2) correctly. The middle item difficulty calibration distribution displays the 50% probability of success of responding to that item's second threshold (i.e., the threshold between categories 2 and 3) correctly. Therefore, in Figure 57 the distribution of the person ability measures and the distribution of the item difficulty calibrations are targeted to a better degree than in the more basic distribution shown in Figure 56.

Figure 57 enables the display of the threshold difficulty parameter information via the Rasch-Andrich rating scale model. Despite this better targeting, one test-taker, person A216, represented by a dot at 3.05 logits in Figure 57, is situated at the highest point of the person ability measure distribution and exceeds the highest threshold difficulty location. Following the Rasch model paradigm targeting principles, this person has not been adequately measured with this test as this person's high location was outside the targeting of the threshold difficulty parameter distribution in Figure 57. However, standard error for the person ability measure should also be taken into account. Person A216's ability measure and standard error were 3.05 and .47, respectively. Comparing person A216's ability measure (3.05) and standard error (.47) to the next highest ability person A027 with a person ability measure (2.25) and standard error (.33) reveals no difference. Therefore, the targeting is not a problem, in other words, 379 of the 379 persons measured were adequately targeted by the test.

Step 3 involves checking for item fit to the Rasch model. None of the 36 items underfit the Rasch model, as no Outfit or Infit MNSQ statistics were > 1.50. The item

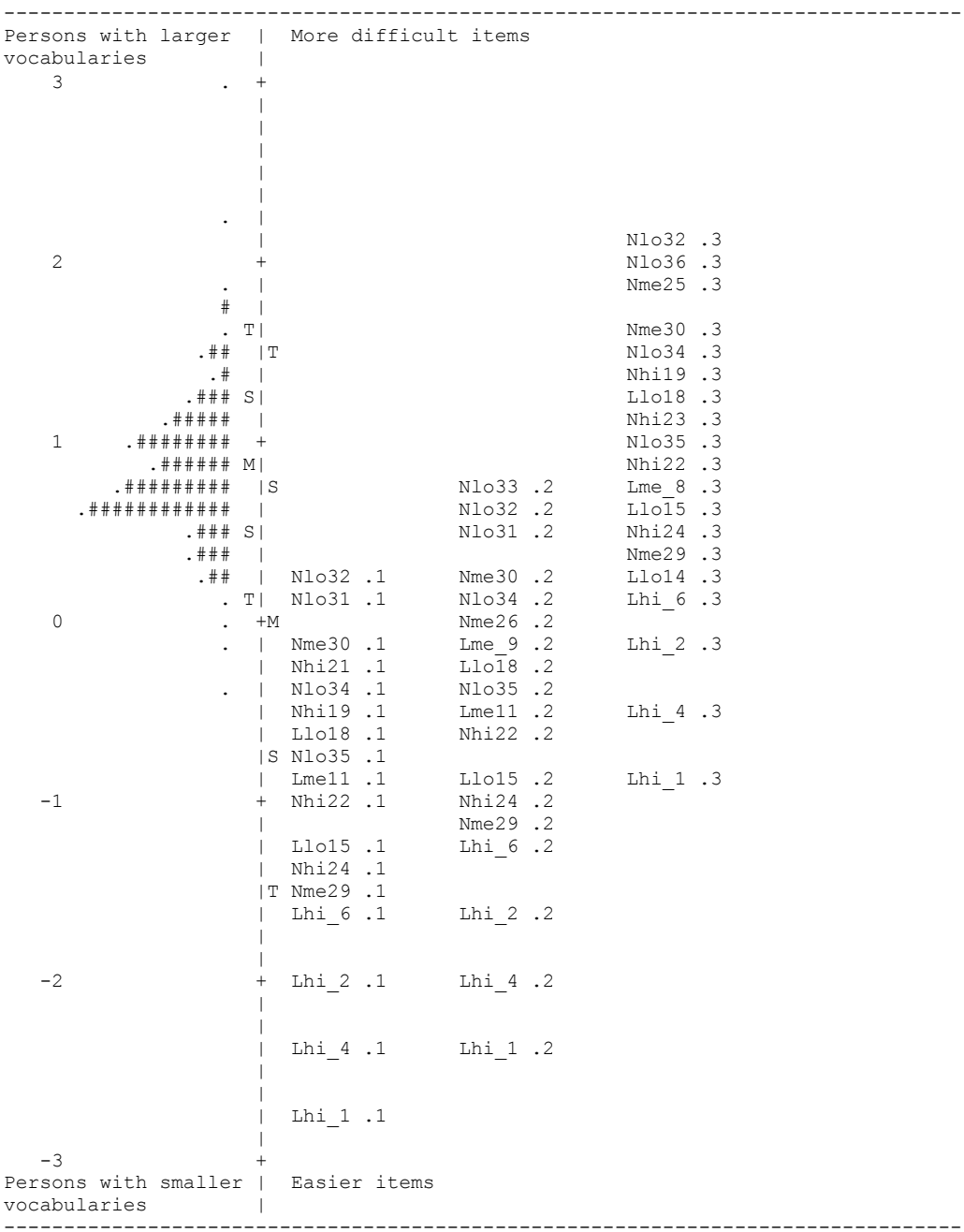


Figure 57. Polysemous Wright Map for baseline quaternary RSM 011223 baseline scale. Each '#' in the Person column is 6 persons. Each '.' is 1 to 5 persons.

indicating the highest underfit item to model values was item Nm30 (keyword: evaluation) with an Outfit MNSQ value of 1.38 and an Infit MNSQ value of 1.39. Two items, Items Lf_5 (keyword: relaxed) and Lf_3 (keyword: positive), displayed extreme (maximum) scores; thus, these two items were very easy for this sample. Although Items Lf_3 and Lf_5 had extreme scores, Winsteps calculated item difficulty calibrations for each; however, no misfit statistics were calculated because as Linacre (*n.d.*) details in an explanation about diagnosing misfit, "Extreme scores always fit the Rasch model exactly, so they are omitted from the computation of fit statistics." This information is displayed in Table 52 under the columns labeled Infit MNSQ and Outfit MNSQ.

Step 4 was an investigation of person fit. None of the 379 persons underfit the Rasch model by displaying Outfit or Infit MNSQ statistics of > 1.5. Person C323 had the highest underfit person to model values with an Outfit MNSQ value of 1.41 and an Infit MNSQ value of 1.49. Step 5 involved an examination of whether all the items were oriented with the latent trait variable. As displayed in Table 52, no items displayed negative point biserial correlation coefficient values.

Steps 6 to 11 concerned checking six requirements laid out by Linacre (2004) concerning category functioning. The information for these requirements is presented in Table 52. Step 6 involved verifying that there were at least 10 observations in each rating category. As shown in the second column of Table 52, the minimum number of observations was in category zero with 875 observations. Step 7 concerns checking that observations are regularly distributed across all rating categories. In Table 52, the numbers of the observed counts frequency distribution is unimodal and indicates an increase from

Table 52

Item Difficulty Calibration Hierarchy

Item	Loan word status	Freq band	Measure	*SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt-measure correlation
33	non	low	1.05	0.06	0.91	-1.6	0.92	-1.4	.22
20	non	high	1.00	0.06	1.22	3.6	1.22	3.5	.29
27	non	med	0.99	0.06	0.95	-0.8	0.97	-0.5	.30
28	non	med	0.97	0.06	0.92	-1.4	0.92	-1.4	.30
32	non	low	0.96	0.06	1.10	1.7	1.10	1.7	.28
36	non	low	0.96	0.06	0.93	-1.2	0.93	-1.2	.35
31	non	low	0.87	0.06	0.98	-0.3	0.99	-0.2	.18
25	non	med	0.81	0.06	1.03	0.6	1.04	0.7	.17
30	non	med	0.56	0.06	1.39	5.5	1.38	5.3	.32
21	non	high	0.49	0.06	1.01	0.2	1.01	0.2	.29
34	non	low	0.43	0.06	1.18	2.6	1.17	2.5	.28
26	non	med	0.36	0.06	1.12	1.7	1.11	1.5	.22
16	loan	low	0.32	0.06	1.10	1.4	1.10	1.5	.27
19	non	high	0.30	0.06	1.10	1.5	1.08	1.2	.36
9	loan	med	0.18	0.06	0.97	-0.4	0.96	-0.6	.31
18	loan	low	0.10	0.07	1.08	1.1	1.10	1.4	.08
17	loan	low	0.04	0.07	1.01	0.2	0.99	-0.1	.35
23	non	high	0.02	0.07	1.17	2.2	1.15	2.0	.30
12	loan	med	-0.01	0.07	0.49	-8.8	0.49	-8.5	.28
35	non	low	-0.05	0.07	1.03	0.4	1.00	0.0	.35
11	loan	med	-0.10	0.07	1.33	4.1	1.31	3.8	.29
10	loan	med	-0.20	0.07	0.55	-7.0	0.56	-6.9	.25
22	non	high	-0.22	0.07	0.83	-2.3	0.83	-2.3	.24
8	loan	med	-0.29	0.07	0.91	-1.2	0.90	-1.3	.37
7	loan	med	-0.30	0.07	0.65	-5.1	0.65	-5.0	.23
15	loan	low	-0.52	0.08	0.89	-1.4	0.89	-1.4	.11
24	non	high	-0.65	0.08	1.11	1.3	1.12	1.4	.14
29	non	med	-0.73	0.08	1.15	1.7	1.08	0.9	.37
13	loan	low	-0.76	0.08	1.06	0.8	1.04	0.4	.29
14	loan	low	-0.80	0.08	0.91	-1.0	0.93	-0.8	.14
6	loan	high	-0.92	0.09	0.87	-1.4	0.85	-1.7	.24
2	loan	high	-1.27	0.10	0.85	-1.6	0.82	-1.9	.24
4	loan	high	-1.62	0.12	0.83	-1.6	0.82	-1.6	.26
1	loan	high	-1.98	0.13	0.90	-0.8	0.93	-0.5	.11
3	loan	high	-7.35	1.83	Min	Min	Min	Min	Min
5	loan	high	-7.35	1.83	Min	Min	Min	Min	Min
M			-0.41	0.17	0.99	-0.2	0.98	-0.3	
SD			1.84	0.40	0.19	2.8	0.18	2.7	

875 to 2,122 (an ascent of 1,247), from 2,122 to 5,074 (an ascent of 2,952), and a decrease from 5,074 to 4,707 (a descent of 367). Therefore, there was a reasonably uniform

distribution across categories for the counted observations. Step 8 involved checking whether the measures advanced monotonically with category. As displayed in the Observed Average column, the logit measures increase (0.06, 0.21, 0.79, and 1.32) with the rating scale categories. Step 9 involved verifying that the Infit and Outfit mean square values were less than 2.0. As shown in Table 52, all Outfit and Infit MNSQ values are less than 2.0.

Step 10 concerned checking that the thresholds advanced with categories. All three Andrich thresholds positively advanced (-0.74, -0.36, and 1.10), so the thresholds were ordered as intended. In an update to Linacre's (2004) guideline recommending a minimum of 1.4 logits threshold advance for a three-category scale and 1.0 logits for a five-category scale, Linacre (2010) asserted that any positively advancing thresholds are acceptable, and even disordered thresholds can be useful. In this study, a positively advancing threshold criterion is adhered to as recommended by Andrich (1978, 2013a, 2013b). In Step 11, thresholds were confirmed to advance by no more than 5.0 logits (see Table 53).

Table 53

Summary of Category Structure

Metasynonymy awareness category	Count (%)	Observed average	Infit MNSQ	Outfit MNSQ	Andrich threshold	Category measure
0 = Null	875 (7)	0.06	1.06	1.05	NONE	(-2.18)
1 = Marginal	2,122 (16)	0.21	0.86	0.79	-0.74	-0.69
2 = Partial	5,074 (39)	0.79	1.01	1.04	-0.36	0.58
3 = Full	4,707 (37)	1.32	1.04	1.03	1.10	2.35

In Step 12 the Winsteps expected score map was examined. The distribution of 379 persons by ability measures is located at the bottom of Figure 58. The person with the lowest ability is located on the left hand side while the person with the highest ability is

located on the right hand side. The number of persons at a particular location can be discerned by reading the number from top to bottom, that is, $\frac{7}{2}$, indicates 72 persons at that

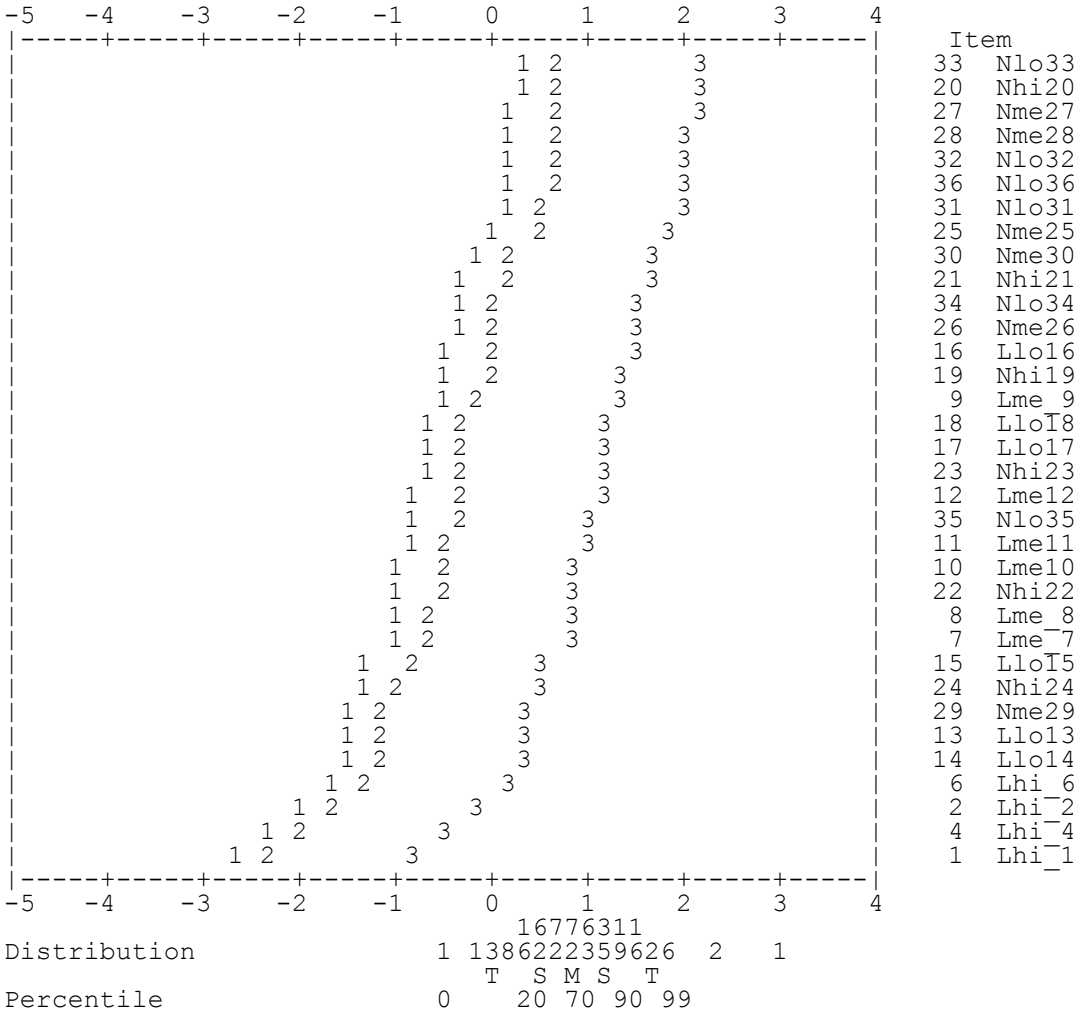


Figure 58. Three Rasch-Andrich thresholds of the baseline quaternary RSM 011223 baseline scale by category score.

location. The items are displayed on the right hand side of Figure 58 with the most difficult item at the top and the easiest item at the bottom. To determine the expected rating response of a person with a particular logit value, a vertical line can be visualized up from

the person's location to determine which rating zone the line passes through for each item. For example, the person with the highest person ability measure of about 3 logits indicated by a "1" at the far right of the person distribution would be expected to respond with a "3" correct response to all items. To the left of this person are 2 persons indicated by a "2" in the person distribution. For these two persons, for the easier items in the item difficulty hierarchy up until Nm30, they were expected to respond with a "3".

However, when these two persons encountered the more difficult items Nm25 and Nlo31, they were expected to have a 50% probability of successfully answering with a correct response (i.e., 3). Further, when these two persons encountered the very difficult items, Nlo36, Nlo32, Nm28, Nm27, Nhi20, and Nlo33 they were expected to respond with a partially correct response (i.e., 2).

An extra final step not included by Iramaneerat, Smith, and Smith (2008) involved checking the unidimensionality of the tested items. In Step 13, the Rasch item residuals were analyzed with a Rasch principal components analysis (PCA) of item residuals. The

Table 54

Standardized Item Residual Variance

	Eigenvalue	Observed % total variance	Observed % unexplained variance	Expected % total variance
Total raw variance in observations	46.71	100.0		100.0
Raw variance explained by measures	12.71	27.2		27.4
Raw variance explained by persons	1.86	4.0		4.0
Raw variance explained by items	10.85	23.2		23.4
Raw unexplained variance (total)	34.00	72.8	100.0	72.6
Unexplained variance in 1st contrast	1.66	3.6	4.9	

Note. Eigenvalue units = item information units.

purpose was to determine if there was more than one variance component in the data. The results are displayed in Table 54.

Residual components should ideally be at the random noise level. An eigenvalue of 1.66, that is, with the strength of 1.66 items of the unexplained variance in the first contrast does not exceed the critical value of 2.0 eigenvalues (Linacre, 2015a). An eigenvalue of greater than 2.0 is not expected by chance. Further, at least two items are needed to define a second dimension. A comparison was also made of the raw variance explained by the items (23.20%) with the unexplained variance in the first contrast (3.60%). This comparison indicated that the Rasch dimension is dominant, as it is 6.44 times larger than the secondary sub-dimension.

In summary, the test meets the all unidimensional Rasch model requirements that were outlined in the sequential steps in this section. These essential requirements are crucial checks to be addressed before further analyses can continue and be reassured of meaningful measurement. Therefore, it is considered appropriate to proceed with any subsequent analyses.

Research Question 2: Hypothesized Difficulty of Loanwords and Non-Loanwords

Research question 2 asked whether the hypothesized easiness of loanwords and the difficulty of non-loanwords are confirmed by the data. This research question was answered in four ways. First, an initial visualization of the right side of the Wright map indicating the item difficulty calibration hierarchy was used. Second, a numeric examination of the item difficulty calibration hierarchy in Table 55 was inspected. Third, a

visual examination of the interval measure data was conducted by exporting interval measured data to IBM SPSS (version 21.0) in order to graph a plot separating the data according to word frequency bands. Fourth, an independent samples *t*-test was conducted using IBM SPSS (version 21.0).

The Wright Map item difficulty calibration distribution has been separated into two classifications according to loanword status in Figure 59.

The first item classification, the left side distribution, shows item keywords that are categorized as loanwords, and the second classification, the right side distribution, displays item keywords that are categorized as non-loanwords. The item keywords in the non-loanword distribution are located relatively higher (i.e., they are more difficult) on the logit scale in comparison to the item keywords in the loanword distribution.

Table 55 displays the item difficulty calibration hierarchy categorized by loanword status, that is, by loanwords and non-loanwords. Keywords classified as non-loanwords are relatively more difficult than keywords classified as loanwords as displayed by the demarcation in the second and third columns.

In Figure 60, the mean difficulty of keywords classified as non-loanwords is higher than the mean difficulty of the keywords classified as loanwords. Despite some overlap of the whiskers of the boxplots, the stems of the item keywords classified as non-loanwords are higher than the mean difficulty of the item keywords classified as loanwords.
An independent samples *t*-test (one-tailed) was conducted to evaluate the directional alternative hypothesis that test keywords classified as loanwords (k = 18) were easier than non-loanwords (k = 18) are different among polysemous academic vocabulary extent test

Figure 59. Wright map according to loanword status (loanwords and non-loanwords). Each '#' is 6.

items (k = 36). Data screening utilizing a z-score transformation to the data indicated two initial outliers in the loanword group. After changing the values of these outliers a

Table 55

Item Difficulty Calibration Hierarchy by Loanword Status

Item	Loan word	Non-loan word	Measure	*SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt-measure correlation
33		√	1.05	0.06	0.91	-1.6	0.92	-1.4	.22
20		√	1.00	0.06	1.22	3.6	1.22	3.5	.29
27		√	0.99	0.06	0.95	-0.8	0.97	-0.5	.30
28		√	0.97	0.06	0.92	-1.4	0.92	-1.4	.30
32		√	0.96	0.06	1.10	1.7	1.10	1.7	.28
36		√	0.96	0.06	0.93	-1.2	0.93	-1.2	.35
31		√	0.87	0.06	0.98	-0.3	0.99	-0.2	.18
25		√	0.81	0.06	1.03	0.6	1.04	0.7	.17
30		√	0.56	0.06	1.39	5.5	1.38	5.3	.32
21		√	0.49	0.06	1.01	0.2	1.01	0.2	.29
34		√	0.43	0.06	1.18	2.6	1.17	2.5	.28
26		√	0.36	0.06	1.12	1.7	1.11	1.5	.22
16	√		0.32	0.06	1.10	1.4	1.10	1.5	.27
19		√	0.30	0.06	1.10	1.5	1.08	1.2	.36
9	√		0.18	0.06	0.97	-0.4	0.96	-0.6	.31
18	√		0.10	0.07	1.08	1.1	1.10	1.4	.08
17	√		0.04	0.07	1.01	0.2	0.99	-0.1	.35
23		√	0.02	0.07	1.17	2.2	1.15	2.0	.30
12	√		-0.01	0.07	0.49	-8.8	0.49	-8.5	.28
35		√	-0.05	0.07	1.03	0.4	1.00	0.0	.35
11	√		-0.10	0.07	1.33	4.1	1.31	3.8	.29
10	√		-0.20	0.07	0.55	-7.0	0.56	-6.9	.25
22		√	-0.22	0.07	0.83	-2.3	0.83	-2.3	.24
8	√		-0.29	0.07	0.91	-1.2	0.90	-1.3	.37
7	√		-0.30	0.07	0.65	-5.1	0.65	-5.0	.23
15	√		-0.52	0.08	0.89	-1.4	0.89	-1.4	.11
24		√	-0.65	0.08	1.11	1.3	1.12	1.4	.14
29		√	-0.73	0.08	1.15	1.7	1.08	0.9	.37
13	√		-0.76	0.08	1.06	0.8	1.04	0.4	.29
14	√		-0.80	0.08	0.91	-1.0	0.93	-0.8	.14
6	√		-0.92	0.09	0.87	-1.4	0.85	-1.7	.24
2	√		-1.27	0.10	0.85	-1.6	0.82	-1.9	.24
4	√		-1.62	0.12	0.83	-1.6	0.82	-1.6	.26
1	√		-1.98	0.13	0.90	-0.8	0.93	-0.5	.11
3	√		-7.35	1.83	Min	Min	Min	Min	Min
5	√		-7.35	1.83	Min	Min	Min	Min	Min
M			-0.41	0.17	0.99	-0.2	0.98	-0.3	
SD			1.84	0.40	0.19	2.8	0.18	2.7	

subsequent *z*-score transformation indicated three outliers. After changing these values, a third *z*-score transformation indicated that there were no outliers. The normal distribution assumption was checked by using the Kolmogorov-Smirnov test on the criterion variable.

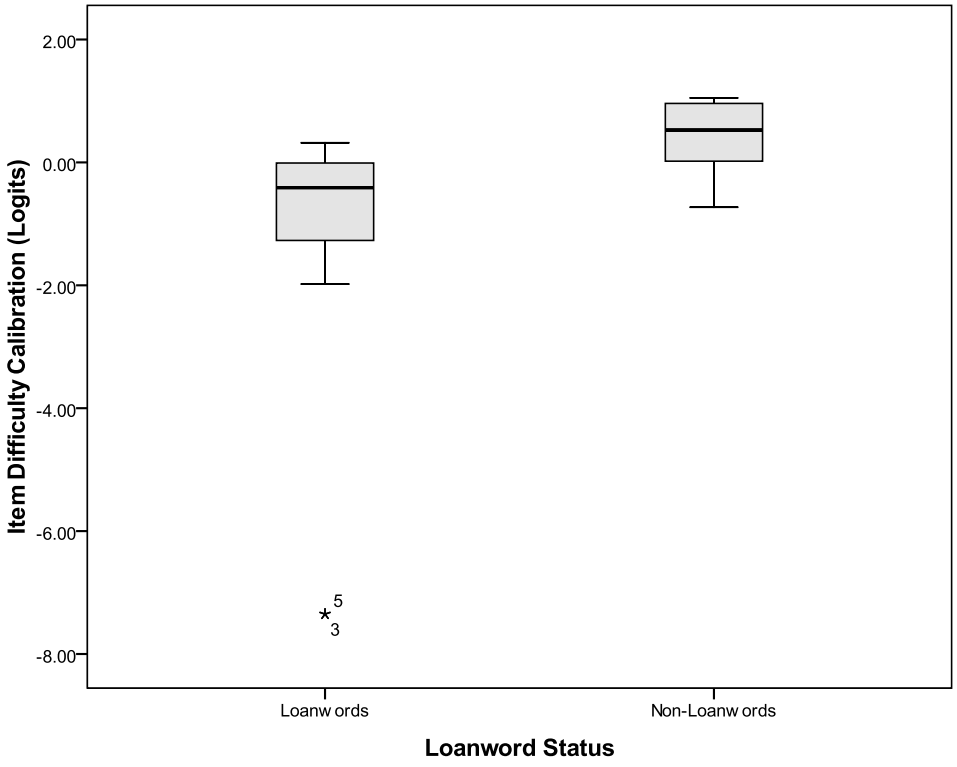


Figure 60. Boxplot of keyword loanword status means according to item difficulty calibrations.

The test statistic indicated that the loanword item difficulty calibrations $D(18) = 0.18$, $p = .15$ and the non-loanword item difficulty calibrations $D(18) = 0.18$, $p = .13$, both had significantly normal distributions. The homogeneity of variance assumption was

checked using Levene's test on the criterion variable. For the item difficulty calibrations, the variances were equal for loanwords and non-loanwords, $F(1, 34) = 1.73$, $p = .19$.

As predicted, on average the non-loanword item difficulty calibrations were found to be more difficult at a statistically significant level, $t(34) = -5.01$, $p < .001$, one-tailed. The mean item difficulty calibration for loanwords ($M = -0.62$, $SD = .71$) was lower than the mean item difficulty calibration for non-loanwords ($M = 0.45$, $SD = .57$). The difference represented a large effect, size $r = .65$. The 95% confidence interval around the difference between the sample means, -1.07, ranged from -1.51 to -0.64.

In summary, in terms of item difficulty calibrations, on average the keywords classified as loanwords were easier than the keywords classified as non-loanwords.

Research Question 3: Word Frequency Theory

Research question 3 asked whether Palmer's (1917) word frequency acquisition theory was confirmed by the data. This research question was answered in four ways. First, an initial visualization of the right side of the Wright map (Figure 61) was performed to inspect the item difficulty calibration hierarchy according to high, medium, and low frequency bands. Second, a numeric examination of the item difficulty calibration hierarchy in Table 56 was conducted. Third, a visual examination of the interval measure data was conducted by exporting the interval measured data to IBM SPSS in order to graph a plot separating the data according to word frequency bands. Fourth, a one-way independent one-tailed analysis of variance (ANOVA) and three planned non-orthogonal comparisons were conducted.

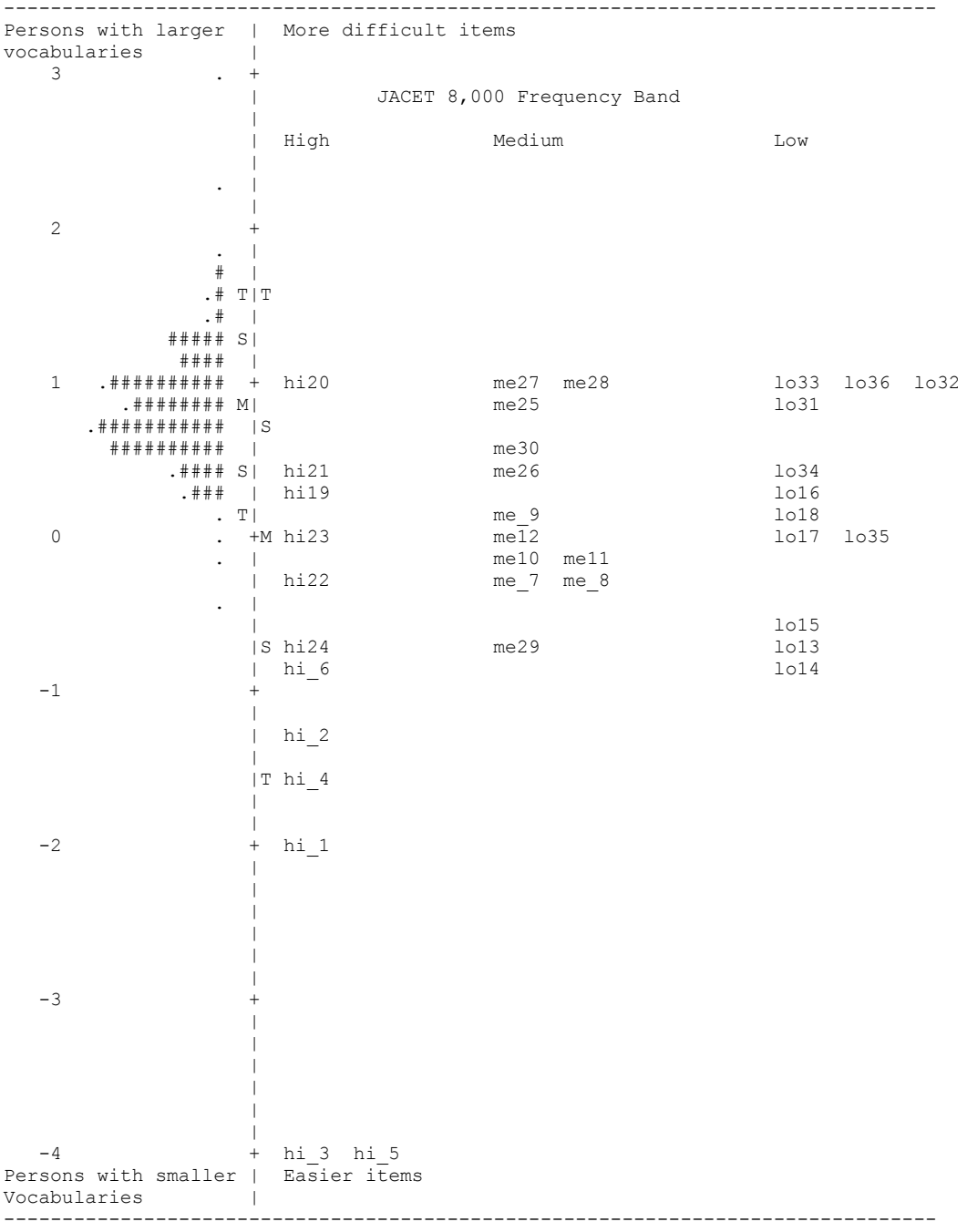


Figure 61. Rank-ordered option response 4-category RSM 011223 Wright map according to high, medium, and low frequency keyword frequency band levels of the JACET 8000 academic word list. Each '#' is 6.

In Figure 61, the Wright Map item difficulty calibration distribution was separated into three classifications using word frequency status. The first classification, in the left distribution, was defined as high frequency words from the 1,000, 2,000, and 3,000 JACET 8000 (Aizawa et al., 2008) word frequency bands.

The second classification, in the middle distribution, is made up of item keywords that are categorized as medium word frequency, defined as words from the 4,000 and 5,000 JACET 8000 word frequency bands. The third classification, in the right distribution, is made up of item keywords that are categorized as low word frequency, which was defined as items from the 6,000, 7,000, and 8,000 JACET 8000 word frequency bands. The keywords classified as high frequency are located relatively lower (i.e., easier items) on the logit scale in comparison to the medium and high word frequency distributions.

Table 56 displays the item difficulty calibration hierarchy categorized by word frequency band, that is, by high, medium, and low frequency. The items are arranged according to difficulty, with more difficult words at the top (e.g., item 33 is the most difficult item; logit measure = 1.05) and easier items toward the bottom (e.g., item 5 is the easiest item; logit measure = -7.35). Keywords classified as high frequency are relatively easier than keywords classified as medium and low frequency.

A visual examination of the boxplots in Figure 62 indicates that all the error bars overlap, giving an initial indication that there are no between-group differences; however, this visualization is only an approximation. The boxplot indicates a slight linear trend, in that, as keywords move from high to medium to low frequency band levels, there is a subsequent small increase in item difficulty calibration.

Table 56

Item Difficulty Calibration Hierarchy by Word Frequency Band

Item	High Freq Band	Medium Freq Band	Low Freq Band	Measure	*SE*	Infit MNSQ	Outfit MNSQ	Pt-Measure Correlation
33			√	1.05	0.06	0.91	0.92	.22
20	√			1.00	0.06	1.22	1.22	.29
27		√		0.99	0.06	0.95	0.97	.30
28		√		0.97	0.06	0.92	0.92	.30
32			√	0.96	0.06	1.10	1.10	.28
36			√	0.96	0.06	0.93	0.93	.35
31			√	0.87	0.06	0.98	0.99	.18
25		√		0.81	0.06	1.03	1.04	.17
30		√		0.56	0.06	1.39	1.38	.32
21	√			0.49	0.06	1.01	1.01	.29
34			√	0.43	0.06	1.18	1.17	.28
26		√		0.36	0.06	1.12	1.11	.22
16			√	0.32	0.06	1.10	1.10	.27
19	√			0.30	0.06	1.10	1.08	.36
9		√		0.18	0.06	0.97	0.96	.31
18			√	0.10	0.07	1.08	1.10	.08
17			√	0.04	0.07	1.01	0.99	.35
23	√			0.02	0.07	1.17	1.15	.30
12		√		-0.01	0.07	0.49	0.49	.28
35			√	-0.05	0.07	1.03	1.00	.35
11		√		-0.10	0.07	1.33	1.31	.29
10		√		-0.20	0.07	0.55	0.56	.25
22	√			-0.22	0.07	0.83	0.83	.24
8		√		-0.29	0.07	0.91	0.90	.37
7		√		-0.30	0.07	0.65	0.65	.23
15			√	-0.52	0.08	0.89	0.89	.11
24	√			-0.65	0.08	1.11	1.12	.14
29		√		-0.73	0.08	1.15	1.08	.37
13			√	-0.76	0.08	1.06	1.04	.29
14			√	-0.80	0.08	0.91	0.93	.14
6	√			-0.92	0.09	0.87	0.85	.24
2	√			-1.27	0.10	0.85	0.82	.24
4	√			-1.62	0.12	0.83	0.82	.26
1	√			-1.98	0.13	0.90	0.93	.11
3	√			-7.35	1.83	Min	Min	Min
5	√			-7.35	1.83	Min	Min	Min
M				-0.41	0.17	0.99	0.98	
SD				1.84	0.40	0.19	0.18	

Note. SE = Standard Error; Pt-measure correlation = point measure correlation; Min = minimum estimated value.

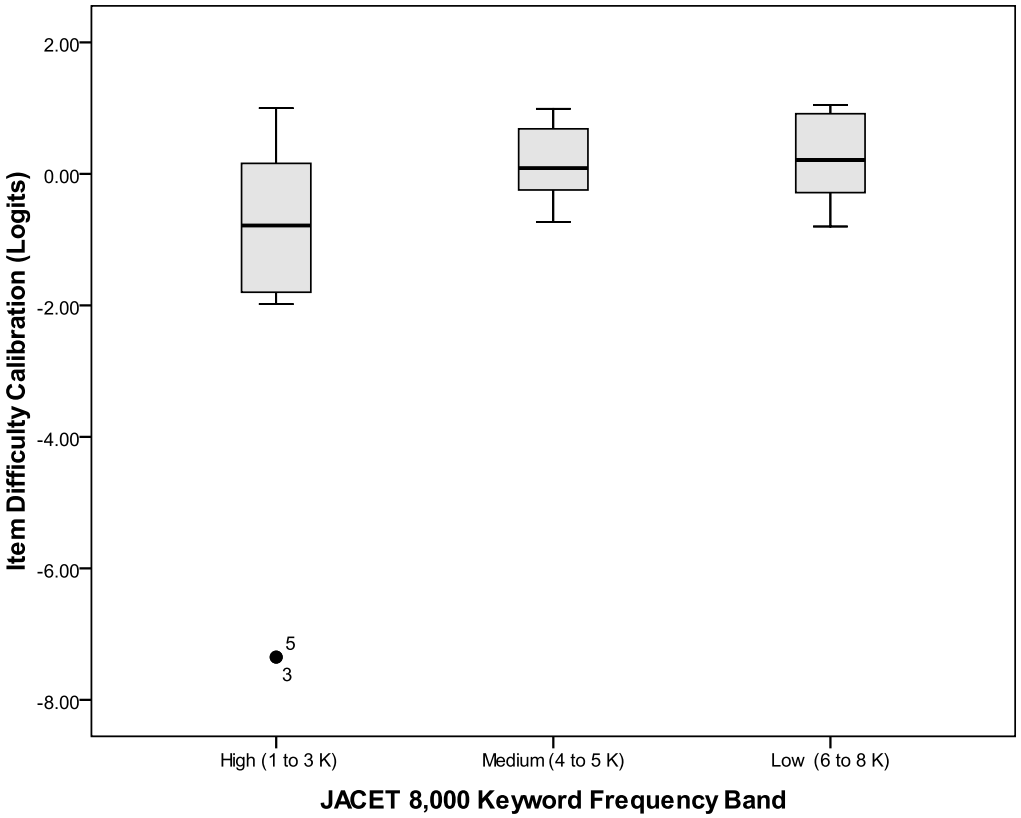


Figure 62. Boxplot of keyword frequency band means according to item difficulty calibrations.

A one-way independent analysis of variance (ANOVA) test (one-tailed) was conducted to evaluate the alternative hypothesis that at least two of the three group means, that is word frequency, high (k = 12), medium (k = 12), and low (k = 12), are different among polysemous academic vocabulary extent test items (k = 36). A one-tailed test was chosen because high word frequency band keywords are hypothesized a priori to be easier than medium word frequency band keywords, which in turn, are also hypothesized to be easier than low word frequency band keywords. Therefore, the expected outcome on the

criterion variable is directional. As the analysis of variance test is an omnibus test, planned contrasts were also conducted.

The three planned contrasts used three independent samples *t*-tests (one-tailed) to evaluate the directional alternative hypotheses that for test keywords classified according to word frequency, high (k = 12) would be easier than medium (k = 12), medium would be easier than low (k = 12) and high would be easier than low.

Data screening utilizing a *z*-score transformation from the previous research question analysis indicated that there were no outliers. The normal distribution assumption was checked by using the Kolmogorov-Smirnov test on the criterion variable. The test statistic indicated that the high word frequency band item difficulty calibrations, $D(12) = 0.17$, $p = .20$, the medium word frequency bad item difficulty calibrations, $D(12) = 0.14$, $p = .20$, and the low word frequency bad item difficulty calibrations, $D(12) = 0.17$, $p = .20$, all had significantly normal distributions. The homogeneity of variance assumption was checked using Levene's test on the criterion variable. For the item difficulty calibrations, the variances were equal for high, medium and low word frequency bands, $F(2, 33) = 1.73$, $p = .06$.

The overall *F* test for the one-way independent analysis of variance was statistically significant, $F(2, 33) = 5.29$, $p = .01$. This corresponded to a medium-sized effect of $\omega = .44$. Planned non-orthogonal contrasts familywise error was controlled by adopting a Bonferroni adjustment correction.

As three planned contrasts were conducted the criterion for significance, .05, was divided by 3, therefore, $p = .017$. As was predicted, the first planned contrast indicated that high word frequency band keywords ($M = -0.66$, $SD = .96$, 95% confidence interval (*CI*), -1.27 to -0.05) were easier than medium word frequency band keywords ($M = 0.19$, $SD = .56$, 95% *CI*, -0.17 to 0.54), $t(33) = 2.77$, $p = .0045$, one-tailed, with a medium-sized effect, $r = .43$. As was not predicted, the second planned contrast indicated that medium word frequency band keywords were not different to low word frequency band keywords ($M = 0.22$, $SD = .67$, 95% *CI*, -0.21 to 0.64), $t(33) = 0.10$, $p = .462$, one-tailed, with a small-sized effect, $r = .02$. As was predicted, the third planned contrast indicated that high word frequency band keywords were easier than low word frequency band keywords, $t(33) = 2.86$, $p = .0035$, one-tailed, with a medium-sized effect $r = .45$.

In summary, in terms of item difficulty calibrations, on average, high frequency keywords were easier than medium frequency keywords and high frequency keywords were easier than low frequency keywords, however, medium frequency keywords were not different to low frequency keywords.

Research Question 4: Metasynonymy Awareness Hypothesis

Research question 4 asked whether the metasynonymy awareness hypothesis based on an amendment to O'Connor's (1940) vocabulary acquisition theory was confirmed by the data. This research question was answered by analyzing the baseline ordered triple rank (OTR) response quaternary scale 011223 data output. The four categories crossover at three ordered thresholds (-0.74, -0.36, and 1.10) as indicated in the Andrich thresholds column in

Table 57. Therefore, the data output in Table 57 confirms that the metasynonymy awareness hypothesis based on an amendment to O'Connor's (1940) amended vocabulary acquisition theory displays ordered thresholds.

Table 57

Summary of Category Structure for RSM 011223 Scale Ordered Triple Rank

Category	Count (%)	Observed average	Infit MNSQ	Outfit MNSQ	Andrich thresholds	Category measure
0 = Incorrect	875 (7)	0.06	1.06	1.05	NONE	(-2.18)
1 = Partially incorrect	2,122 (16)	0.21	0.86	0.79	-0.74	-0.69
2 = Partially correct	5,074 (39)	0.79	1.01	1.04	-0.36	0.58
3 = Correct	4,707 (37)	1.32	1.04	1.03	1.10	2.35

Andrich (1978, 1996, 2010, 2013a, 2013b) has advocated the ordering of thresholds to attain sufficient statistics, invariance and meaningful measurement. Andrich advocates that the thresholds do not have to be separated by a specific amount; the thresholds are required to be ordered only. As these three thresholds are ordered (-0.74, -0.36, and 1.10) the analyses can proceed to attain sufficient statistics, invariance and meaningful measurement.

Threshold order is further illustrated by the ordered thresholds in the category probability curves displayed in Figure 63. The crossover point between the category curves 0 and 1 is located on the left side, forming the first threshold, T_1, at -0.74 logits, the crossover point between the category curves 1 and 2 is located next on the right side of the first threshold, forming the second threshold, T_2, at -0.36 logits and the crossover point between the category curves 1 and 2 is located on the far right, forming the third threshold,

T_3, at 1.10 logits. Therefore, the three rescored thresholds in this study are empirically ordered.

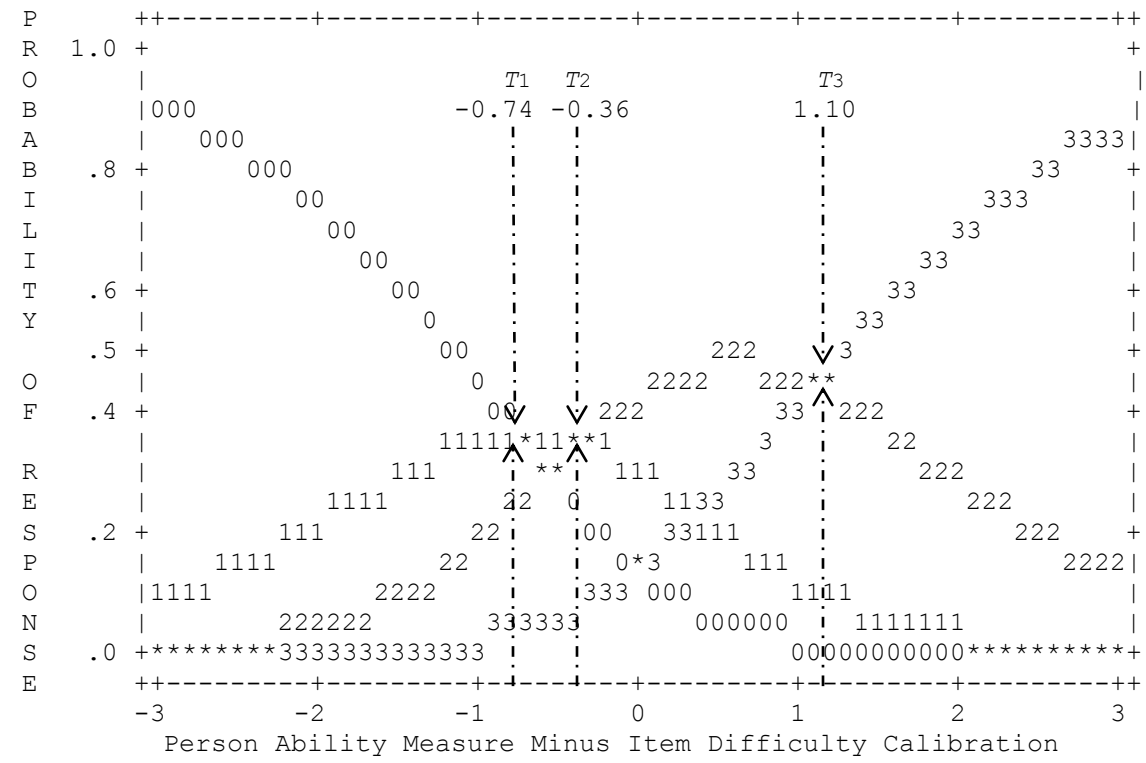


Figure 63. Category probability curves for 3 categories of the RSM 011223 scale. T_1 (-0.74) indicates the crossover point, Threshold 1, at -0.74 logits; T_2 (-0.36) indicates Threshold 2 at -0.36 logits; T_3 indicates Threshold 3 at 1.10 logits.

In summary, the amended and collapsed baseline RSM 011223 scoring guide was confirmed by the data because the three thresholds were ordered. Therefore, as the thresholds were ordered, sufficient statistics and meaningful measurement can be assumed and subsequent statistical analyses might be conducted.

Research Question 5: Vocabulary Knowledge Estimates and the Type of Analysis

Research question 5 asked how incremental vocabulary knowledge estimates change depending on the type of analysis utilized. There are four types of analyses of the person ability measures from four ordered triple rank (OTR) response scales as displayed in Table 58: (a) baseline quaternary RSM 011223 compared to (b) binary DRM 000001; (c) ternary RSM 001122, which is equivalent to a single best answer (SBA) RSM 012, and; (d) binary DRM 000011, which is equivalent to a single best answer (SBA) DRM 001.

Table 58

Description of Four Rasch Model Scale Analyses

Vocabulary	Dichotomously scored	Polytomously scored
Acquisition (Data Column 3 Only)	Size DRM 000011 (DRM 001)	Extent (Size and Strength) RSM 001122 (RSM 012)
Awareness (Data Columns 1, 2 and 3)	Size DRM 000001	Extent (Size and Strength) RSM 011223

Initially, the assumptions of the parametric *t*-test for two dependent samples were checked. The normal distribution assumption in each of the conditions was checked by using the Kolmogorov-Smirnov test. The test statistic indicated that the Rasch person ability measures generated from the four scoring methods, that is, RSM 011223, $D(379) = 0.08$, $p < .001$, DRM 000001, $D(379) = 0.10$, $p < .001$, RSM 001122, $D\ (379) = 0.08$, $p < .001$ and DRM 000011, $D(379) = 0.13$, $p < .001$, all had significantly non-normal distributions. The homogeneity of variance assumption was checked using Levene's test on the conditions. The test statistic indicated that for the person ability measures generated

from RSM 011223 and DRM 000001 scoring, $F(1, 756) = 47.12$, $p < .001$, RSM 011223 and RSM 001122 scoring, $F(1, 756) = 6.75$, $p < .001$ and RSM 011223 and DRM 000011 scoring, $F(1, 756) = 53.68$, $p < .001$, therefore, all variances were non-equal between the scoring methods. As both these parametric statistical assumptions of the *t*-test for two dependent samples were violated a nonparametric statistical procedure was selected.

Three nonparametric Wilcoxon matched-pairs signed-ranks tests (one-tailed) were conducted to evaluate the directional alternative hypotheses that for conditions 1 and 2, the median of the difference measures is a value less than zero, that is, a negative value among Japanese university students ($N = 379$) (Sheskin, 2007). The first Wilcoxon matched-pairs signed-ranks test evaluated the median of the person ability measure differences generated in condition 1 (RSM 011223 scoring) and condition 2 (DRM 000001 scoring). The second Wilcoxon matched-pairs signed-ranks test evaluated the median of the person ability measure differences generated in condition 1 (RSM 011223 scoring) and condition 2 (RSM 001122 scoring). The third Wilcoxon matched-pairs signed-ranks test evaluated the median of the person ability measure differences generated in condition 1 (RSM 011223 scoring) and condition 2 (DRM 000011 scoring). All three tests had directional alternative hypotheses as it was predicted that baseline RSM 011223 scoring would generate higher person ability measures compared to DRM 000001, RSM 001122 and DRM 000011 scoring. The directional alternative hypotheses were set a priori where H_1: $\Theta_D < 0$, that is, the median of the difference measures is a number less than zero. In other words, the directional hypothesis for the difference scores is that the sum of the negative ranks is greater than the sum of the positive ranks (Sheskin, 2007). Wilcoxon matched-pairs signed-

ranks test assumptions were checked. The assumption of measures being at least interval level was met. The difference scores symmetrical distribution about the median assumption was also met (Sheskin, 2007).

As predicted, the person ability measure median difference was significantly higher under RSM 011223 scoring (*Mdn* = 0.80) than under the DRM 000001 scoring (*Mdn* = -0.73), $z = -16.87$, $p < .001$ (1-tailed). Of the total rank sum of 72,010, the favorable sum of negative ranks totaled 72,010 (100%) and the unfavorable sum of the positive ranks totaled 0 (0%). The effect size, represented by the matched-pairs rank biserial correlation coefficient was $r_c = 1.00$.

As predicted, the person ability measure median difference was significantly higher under RSM 011223 scoring (*Mdn* = 0.80) than under RSM 001122 scoring (*Mdn* = 0.65), $z = -11.22$, $p < .001$ (1-tailed). Of the total rank sum of 70,500, the favorable sum of negative ranks totaled 58,813 (83%) and the unfavorable sum of the positive ranks totaled 11,687 (17%). The effect size, represented by the matched-pairs rank biserial correlation coefficient was $r_c = .66$.

As predicted, the person ability measure median difference was significantly higher under RSM 011223 scoring (*Mdn* = 0.80) than under DRM 000011 scoring (*Mdn* = 0.06), $z = -16.73$, $p < .001$ (1-tailed). Of the total rank sum of 71,631, the favorable sum of negative ranks totaled 71,368 (100%) and the sum of the positive ranks totaled 263 (0%). The effect size, represented by the matched-pairs rank biserial correlation coefficient was $r_c = 1.00$.

The precision, that is, the reliability, of the person ability measures decreased as scoring guides changed from the baseline RSM 011223 scoring guide. Person separation

indices follow: for the baseline RSM 011223 scoring guide, 1.24, the DRM 000001 scoring guide, 1.12, the RSM 001122 scoring guide, 1.05 and the DRM 000011 scoring guide, 1.02. Comparing the DRM 000001 and RSM 011223 scoring guides resulted in a person separation ratio of 1.12/1.24 = .90, indicating the DRM 000001scoring guide is 100(1.12/1.24) = 90% efficient of the baseline RSM 011223 scoring guide. Comparing the RSM 001122 and RSM 011223 scoring guides resulted in a person separation ratio of 1.05/1.24 = .85, indicating the RSM 001122 scoring guide is 100(1.05/1.24) = 85% efficient of the baseline RSM 011223 scoring guide. Comparing the DRM 000011 and RSM 011223 scoring guides resulted in a person separation ratio of 1.02/1.24 = .82, indicating the DRM 000001 scoring guide is 100(1.02/1.24) = 82% efficient of the baseline RSM 011223 scoring guide (Stone, 1998).

Additionally, changes in person ability measures after a shift constant was applied (Wright & Masters, 1982) were investigated. The purpose was to provide a descriptive statistical analysis of the person ability measures generated from the four scoring guides after adjusting for any differences in visual figures. An inferential statistical procedure was also used to examine any differences. The means of the person ability measure distributions were as follows: RSM 011223 scoring, M = 0.84, DRM 000001 scoring, M = -0.70, RSM 001122, M = 0.70 and DRM 000011, M = 0.00. Therefore, before comparisons could be made, numerical adjustments were made to the person ability measures (Wright & Masters, 1982). The shift constant adjustments were made as follows: DRM 000001 scored person ability measures were increased by 1.54 logits (i.e., 0.84 - -0.70 = 1.54), RSM 00112 scored person ability measures were increased by 0.14 logits (i.e., 0.84 - 0.70 = 0.14) and DRM

000011 scored person ability measures were increased by 0.84 logits (i.e., 0.84 - 0.00 = 0.84). Therefore, all analyses were centered on the baseline RSM 011223 scoring analysis mean value.

The changes in person ability measures due to scoring under the baseline RSM 011223 and DRM 000001 scoring guides is visualized in Figure 64. Due to deletion of non-

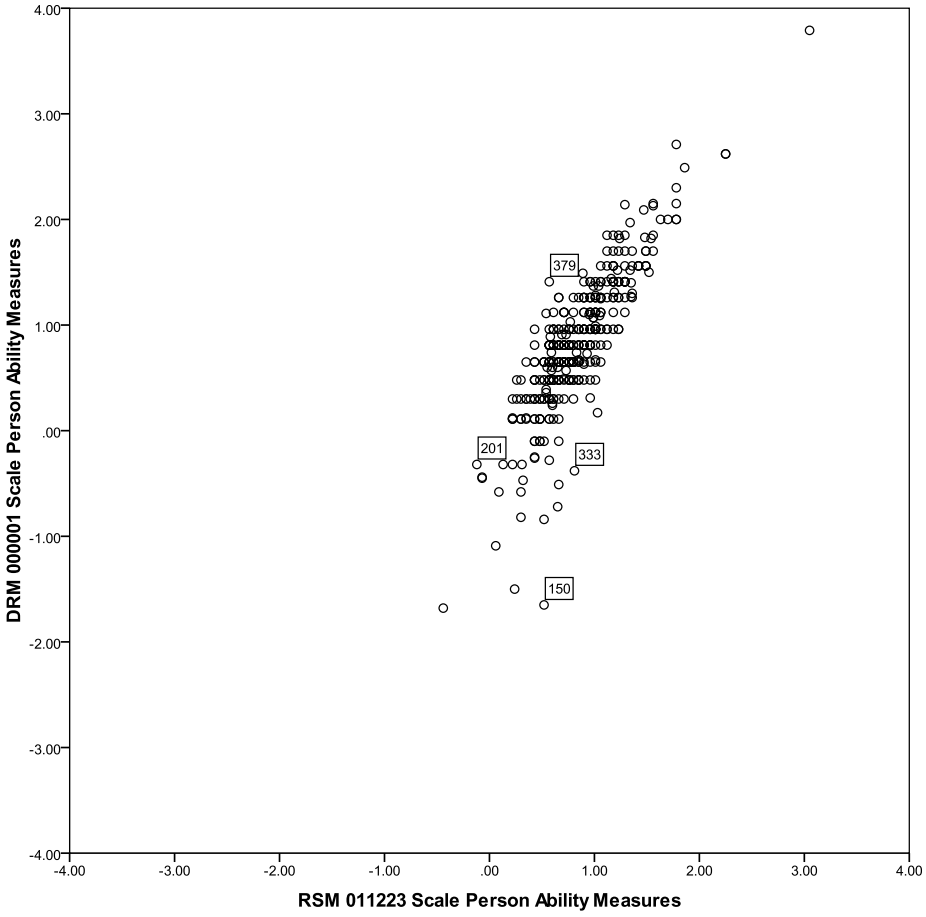


Figure 64. Difference between scoring RSM 011223 and DRM 000001 in person ability measures. IBM SPSS Persons 150, 201, 333 and 379 = C153, D209, D355 and F403, respectively.

Japanese persons in Winsteps dataset, IBM SPSS read the output as slightly different person numbers. These person numbers have been corrected inside parentheses following the IBM SPSS person number in the following analyses. As indicated by the flagged data point values in Figure 64, IBM SPSS person 150 (C153) (0.52 logits) and person 379 (F403) (0.57 logits) had similar (0.05 logits) person ability measures under the baseline RSM 011223 scoring guide. However, under proxy DRM 000001 scoring, person 379 (F403) (1.41 logits) and person 150 (C153) (-1.65 logits) had different (3.05 logits) person ability measures.

In contrast, IBM SPSS person 201 (D209) (-0.12 logits) and person 333 (D355) (0.81 logits) flagged in Figure 64 had different (0.93 logits) person ability measures under the baseline RSM 011223 scoring guide. However, under proxy DRM 000001 scoring, person 201 (D209) (-0.32 logits) and person 333 (D355) (-0.38 logits) had similar (0.06 logits) person ability measures.

The changes in person ability measures due to scoring under the baseline RSM 011223 and RSM 001122 scoring guides is visualized in Figure 65. As indicated by the flagged data point values in Figure 65, IBM SPSS person 336 (D358) (0.09 logits) and person 346 (E369) (0.06 logits) had similar (0.03 logits) person ability measures under the baseline RSM 011223 scoring guide. However, under proxy RSM 001122 scoring, person 336 (D358) (0.34 logits) and person 346 (E369) (-1.01 logits) had different (1.35 logits) person ability measures.

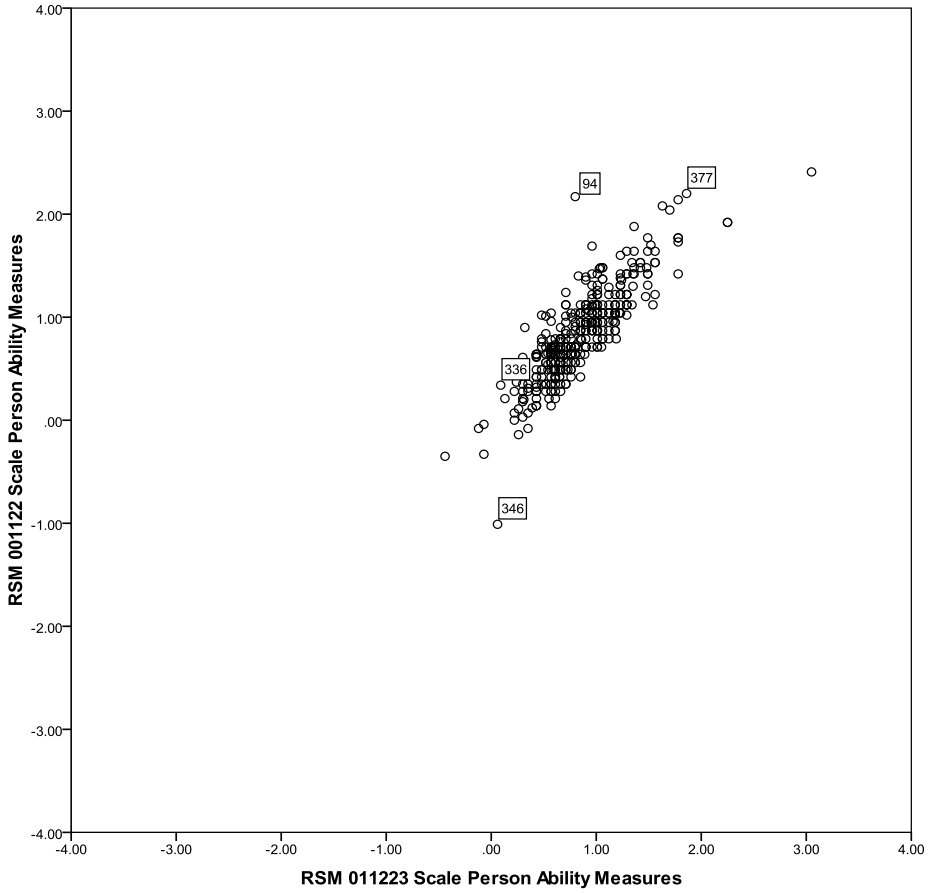


Figure 65. Difference between scoring RSM 011223 and RSM 001122 in person ability measures. IBM SPSS Persons 94, 336, 346 and 377 = C096, D358, E369 and F401, respectively.

In contrast, IBM SPSS person 94 (C096) (0.80 logits) and person 377 (F401) (1.86 logits) flagged in Figure 65 had different (1.06 logits) person ability measures under the baseline RSM 011223 scoring guide. However, under proxy RSM 001122 scoring, person 94 (C096) (2.17 logits) and person 377 (F401) (2.20 logits) had similar (0.03 logits) person ability measures.

The changes in person ability measures due to scoring under the baseline RSM 011223 and DRM 000011 scoring guides is visualized in Figure 66. As indicated by the flagged data point values in Figure 66, IBM SPSS person 331 (D353) (0.13 logits) and person 336 (D358) (0.09 logits) had similar (0.04 logits) person ability measures under the

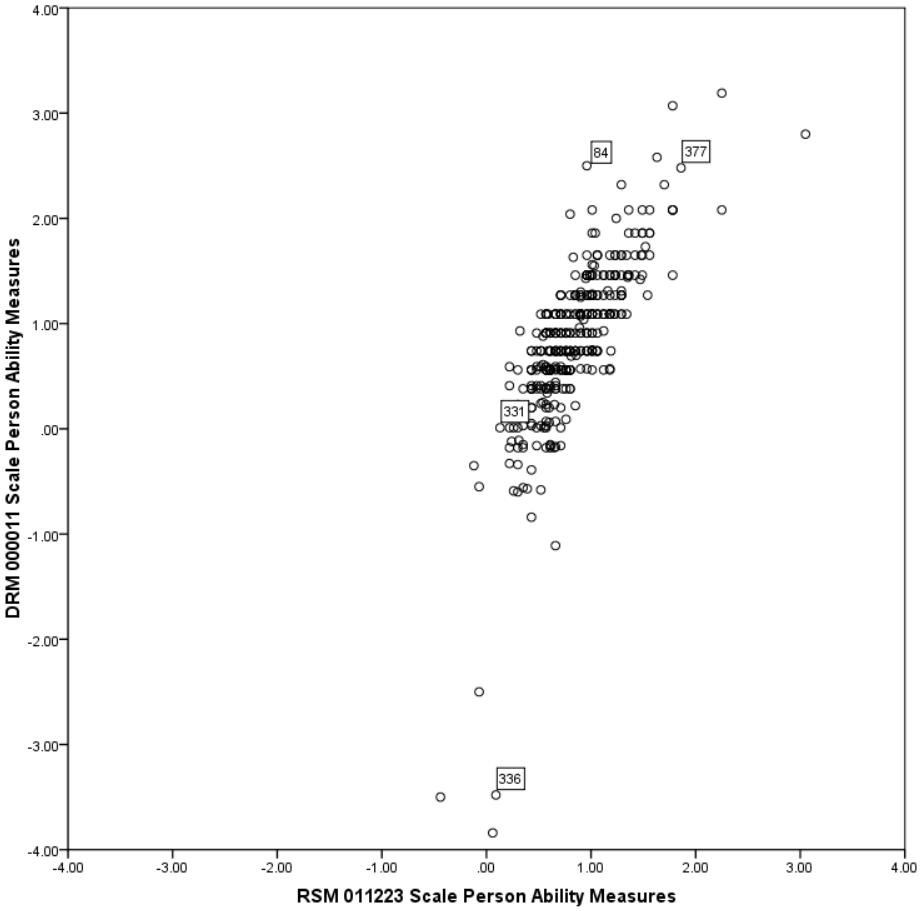


Figure 66. Difference between scoring RSM 011223 and DRM 000011 in person ability measures. IBM SPSS Persons 84, 331, 336 and 377 = C086, D353, E358 and F401, respectively.

baseline RSM 011223 scoring guide. However, under proxy DRM 000011 scoring, person 331 (D353) (0.01 logits) and person 336 (D358) (-3.48 logits) had different (3.49 logits) person ability measures.

In contrast, IBM SPSS person 84 (C086) (0.96 logits) and person 377 (F401) (1.86 logits) flagged in Figure 66 had different (1.06 logits) person ability measures under the baseline RSM 011223 scoring guide. However, under proxy DRM 000011 scoring, person 84 (C086) (2.50 logits) and person 377 (F401) (2.48 logits) had similar (0.02 logits) person ability measures.

Parametric statistical procedure assumptions of normally distributed data and homogeneity of variance were violated for each of the four scoring methods. Subsequently, the Wilcoxon matched-pairs signed-ranks test assumptions were checked. The assumption of measures being at least interval level was met. The difference scores symmetrical distribution assumption about the median was also met. The Wilcoxon matched-pairs signed-ranks test was conducted to assess if person ability measures changed when the scoring guide was changed from the baseline RSM 011223 scoring to DRM 000001 scoring, from the baseline RSM 011223 scoring to RSM 001122 scoring and from the baseline RSM 011223 scoring to DRM 000011 scoring. As a shift constant was applied to this data, changes in person ability measures due to different scoring methods were not anticipated to be significant. Therefore, the null hypothesis was set a priori where H_0: $\Theta_D = 0$, that is, the median of the difference measures is zero. In other words, the null hypothesis for the difference scores is that the sum of the negative ranks is equal to the sum of the positive ranks.

As predicted, the person ability measure median difference was not significantly different under RSM 011223 scoring (*Mdn* = 0.80) and DRM 000001 scoring (*Mdn* = 0.81), *z* = -1.29, *p* =.20 (2-tailed), *r* = -.05. The sum of negative ranks totaled 32,010 and the sum of the positive ranks totaled 37,368. The matched-pairs rank biserial correlation coefficient was 0.96. As predicted, the person ability measure median difference was not significantly different under RSM 011223 scoring (*Mdn* = 0.80) and RSM 001122 scoring (*Mdn* = 0.79), *z* = -1.16, *p* = .25 (2-tailed), *r* = -.04. The sum of negative ranks totaled 36,309.50 and the sum of the positive ranks totaled 31,586.50. The matched-pairs rank biserial correlation coefficient is 0.93. As predicted, the person ability measure median difference was not significantly different under RSM 011223 scoring (*Mdn* = 0.80) and DRM 000011 scoring (*Mdn* = 0.91), *z* = -1.89, *p* = .06 (2-tailed), *r* = -.07. The sum of negative ranks totaled 31,965.50 and the sum of the positive ranks totaled 40,044.50. The matched-pairs rank biserial correlation coefficient is 1.00.

In summary, as predicted the person ability measures changes occurred three times between the four analyses. First, the person ability measures were higher on average under RSM 011223 scoring than under DRM 000001 scoring. Second, the person ability measures were higher on average under RSM 011223 scoring than under RSM 001122 scoring. Third, the person ability measures were higher on average under RSM 011223 scoring than under DRM 000011 scoring. Overall, the baseline RSM 011223 scoring guide person ability measures were on average higher than the other three scoring guide analysis types. The efficiency of the person separation ratio decreased as the scoring guides changed from the baseline of RSM 011223. The DRM 000001 scoring guide is 90% efficient of the

baseline RSM 011223 scoring guide, the RSM 001122 scoring guide is 85% efficient of the baseline RSM 011223 scoring guide, and the DRM 000011 scoring guide is 82% efficient of the baseline RSM 011223 scoring guide (Stone, 1998). As predicted the person ability measures did not change after a shift constant was applied to the person ability measures.

Research Question 6: Guessing and Carelessness

Research question 6 asked to what degree guessing and carelessness are apparent in the type of analyses utilized. First, guessing is investigated and it is followed by an investigation of carelessness.

Item Guessability

First, lower asymptotes related to item guessability were examined. Table 59 displays the lower asymptote index initial approximations for the first 18 test items using keywords that are classified as loanwords. Substantial lower asymptote approximation values that exceed the critical values are bolded in Tables 59 and 60, which indicate items that display item guessability.

Table 59 and 60's first column displays the item, the second column displays quaternary 011223 scale lower asymptote values in a regular Rasch analysis and the third column displays quaternary 011223 scale lower asymptote values after cut elimination. The fourth column displays binary 000001 scale lower asymptote values in a regular Rasch analysis and the fifth column displays binary 000001 scale lower asymptote values after cut elimination. The sixth column displays ternary 001122 scale lower asymptote values in a

regular Rasch analysis and the seventh column displays ternary 001122 scale lower asymptote values after cut elimination. The eighth column displays binary 000011

Table 59

Lower Asymptote Index Initial Approximations Indicating Loanword Keywords Eliciting Item Guessability

Loan-word Item	Quatry 011223 RSM Extent *c* > .75 Regular	Cutlo	Binary 000001 DRM Size *c* > .25 Regular	Cutlo	Ternary 001122 RSM Extent *c* > .66 Regular	Cutlo	Binary 000011 DRM Size *c* > .33 Regular	Cutlo
L hi _1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
L hi _2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
L hi _3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
L hi _4	0.00	0.00	**0.26**	0.00	0.00	0.00	0.00	0.00
L hi _5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
L hi _6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
L m _7	0.00	0.00	0.03	0.07	0.00	0.00	0.00	0.00
L m _8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
L m _9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.07
L m 10	0.00	0.00	0.02	0.03	0.00	0.00	0.00	0.00
L m 11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
L m 12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
L lo 13	0.00	0.00	0.00	0.03	0.00	0.00	0.03	0.00
L lo 14	0.00	0.00	0.00	0.24	0.00	0.00	0.18	0.00
L lo 15	0.00	0.00	0.13	0.08	0.00	0.00	0.09	0.03
L lo 16	0.00	0.00	0.01	0.02	**0.75**	0.00	0.05	0.05
L lo 17	0.00	0.00	0.00	0.00	0.00	**0.75**	0.00	0.00
L lo 18	0.00	0.00	0.07	0.02	0.00	0.00	0.19	0.08

Note. Quatry = Quaternary; Regular = Regular analysis; Cutlo = Cut elimination analysis; bolded values = those exceeding the critical value, *c*.

scale lower asymptote values in a regular Rasch analysis and the ninth column displays binary 000011 scale lower asymptote values after cut low elimination.

Table 60 displays the lower asymptote index initial approximations for the next 18 test items using keywords that are classified as non-loanwords.

Consolidating the relevant information from Table 59 and 60 enables the summary in Table 61. Therefore, in Table 61 the lower asymptote index initial approximations that indicate item guessability from the keywords classified as loanwords and non-loanwords are consolidated into one table.

Table 60

Lower Asymptote Index Initial Approximations Indicating Non-Loanword Keywords Eliciting Item Guessability

Non Loan-word Item	Quatry 011223 RSM Extent *c* > .75		Binary 000001 DRM Size *c* > 25		Ternary 001122 RSM Extent *c* > .66		Binary 000011 DRM Size *c* > .33	
	Regular	Cutlo	Regular	Cutlo	Regular	Cutlo	Regular	Cutlo
N hi 19	0.00	0.00	0.00	0.04	**0.75**	0.00	0.01	0.00
N hi 20	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00
N hi 21	0.00	0.00	0.00	0.05	0.00	0.00	0.01	0.03
N hi 22	0.00	0.00	0.03	0.00	0.00	0.00	0.09	0.00
N hi 23	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00
N hi 24	0.00	0.00	0.19	0.11	0.00	0.00	0.13	0.16
N m 25	0.00	0.00	0.00	**0.27**	0.00	0.00	0.07	0.12
N m 26	0.00	0.00	0.02	0.05	0.00	0.00	0.02	0.00
N m 27	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00
N m 28	0.00	0.00	0.00	0.00	**0.67**	**2.00**	0.00	0.00
N m 29	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
N m 30	0.00	0.00	0.00	0.00	**0.77**	0.00	0.00	0.00
N lo 31	0.00	0.00	0.01	0.09	0.00	0.00	0.01	0.00
N lo 32	0.00	0.00	0.00	0.00	0.15	0.00	0.02	0.05
N lo 33	0.60	0.00	0.00	**0.35**	0.16	0.00	0.00	0.00
N lo 34	0.00	0.00	0.00	0.04	**0.79**	0.00	0.00	0.00
N lo 35	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00
N lo 36	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Note. Quatry = Quaternary; Regular = Regular analysis; Cutlo = Cutlo elimination analysis; bolded values = those exceeding the critical value, *c*.

In Table 61, for the binary DRM 000001 scale, three items were flagged for item guessability. The first flagged item, Lhi4 (*c* = .26) had a lower asymptote of *c* > .25.

Checking the Winsteps most expected responses and most misfitting response string tables for the initial analysis indicated that no persons were responsible for the item guessability. The second flagged item, Nm25 (c = .27) had a lower asymptote value of c > .25. Checking the Winsteps most expected responses and most misfitting response string tables in the cut lo analysis indicated that no persons were flagged. The third item, Nlo33 (c = .35) had a lower asymptote value of c > .25.

Checking the Winsteps most expected responses and most misfitting response string tables in the cut analysis indicated that no persons were flagged. After running the cutlo analysis, item Lhi4 did not exceed the lower asymptote critical value of c > .25.

For the ternary RSM 001122 scale, six items were flagged for eliciting lucky guesses. The first item, Llo16 (c = .75) had a lower asymptote value of c > .33. Checking

Table 61

Lower Asymptote Index Initial Approximations Indicating Loanwords and Non- Loanword Keywords Eliciting Item Guessability

Quaternary 011223 RSM Extent c > .75			Binary 000001 DRM Size c > .25			Ternary 001122 RSM Extent c > .33			Binary 000011 DRM Size c > .33		
Item	Reg	Cutlo	Item	Reg	Cutlo	Item	Reg	Cutlo	Item	Reg	Cutlo
—	—	—	Lhi _4	**0.26**	0.00	Llo16	**0.75**	0.00	—	—	—
			Nm25	0.00	**0.27**	Lrv17	0.00	**0.75**			
			Nlo33	0.00	**0.35**	Nhi19	**0.75**	0.00			
						Nm28	**0.67**	**2.00**			
						Nm30	**0.77**	0.00			
						Nlo34	**0.79**	0.00			

Note. Reg = regular analysis; Cutlo = cut low elimination analysis; c = lower asymptote initial approximation critical value.

the Winsteps most expected responses and most misfitting response string tables for the regular analysis indicated that only two persons, A017 and F400, both with a '0' scored incorrect response were responsible for item guessing. The second item, Lrv17 (c = .75) had a lower asymptote value of c > .33. Checking the Winsteps most expected responses and most misfitting response string tables for the cut analysis indicated that five persons, A023, E363, A213, C115 and C107, all with a '0' scored incorrect response were responsible for the item guessing. The third item, Nhi19 (c = .75) had a lower asymptote value of c > .33. Checking the Winsteps most expected responses and most misfitting response string tables for the initial analysis indicated that four persons, A007, C195, A023, and C175, all with a '0' scored incorrect response were responsible for item guessing. The fourth item, Nm28 (c = .67) had a lower asymptote value of c > .33. Checking the Winsteps most expected responses and most misfitting response string tables indicated that no persons were responsible for item guessing. The fifth item, Nm30 (c = .77) had a lower asymptote value of c > .33. Checking the Winsteps most expected responses and most misfitting response string tables indicated that two persons, C086 and C280, both with a '0' scored incorrect response, were responsible for item guessing. The sixth item, Nlo34 (c = .79) had a lower asymptote value of c > .33. Checking the Winsteps most expected responses and most misfitting response string tables indicated that only two persons, C173 and A016, both with a '0' scored incorrect response, were responsible for item guessing. After running the cut analysis, items Llo16, Nhi19, Nm30 and Nlo34 did not exceed the lower asymptote critical value; however, item Nm28 did exceed the lower asymptote critical value (c = 2.00).

In summary, for the binary DRM 000011 scale, no items were flagged for eliciting guessing. In summary of the four lower asymptote analyses, different methods of scoring in this study resulted in different items eliciting lucky guesses from different people. The ternary RSM 001122 scale had two items flagged for lucky guessing after the cut procedure. The binary DRM 000001 scale flagged three items eliciting lucky guesses after the cut procedure. The quaternary RSM 011223 scale and the binary DRM 000011 scales had no instances of items eliciting lucky guesses.

Item Carelessness

Second, upper asymptotes related to item carelessness were examined. Table 62 displays the upper asymptote index initial approximations for the first 18 test items using keywords that are classified as loanwords. Substantial upper asymptote approximation values are bolded and indicate items that display items that elicit carelessness. Table 62 and 63's first column displays the item, the second column displays quaternary RSM 011223 scale upper asymptote values for the regular Rasch analysis and the third column displays quaternary RSM 011223 scale upper asymptote values after cut elimination. The fourth column displays binary DRM 000001 scale upper asymptote values for the regular Rasch analysis and the fifth column displays binary DRM 000001 scale upper asymptote values after cut elimination. The sixth column displays ternary RSM 001122 scale upper asymptote values for the regular Rasch analysis and the seventh column displays ternary RSM 001122 scale upper asymptote values after cut elimination. The eighth column

displays binary DRM 000011 scale upper asymptote values for the regular Rasch analysis and the ninth column displays binary DRM 000011 scale upper asymptote values after cut elimination. The bolded values exceed the critical values.

Table 62

Upper Asymptote Index Initial Approximations Indicating Loanword Keywords Eliciting Item Carelessness

Loan-word Item	Quatry 011223 RSM Extent $c < 2.25$ Regular	Cuthi	Binary 000001 DRM Size $c < 0.75$ Regular	Cuthi	Ternary 001122 RSM Extent $c < 1.34$ Regular	Cuthi	Binary 000011 DRM Size $c < 0.67$ Regular	Cuthi
L hi _1	3.00	3.00	1.00	1.00	**1.00**	**1.00**	1.00	1.00
L hi _2	3.00	3.00	1.00	1.00	**1.00**	**1.00**	1.00	1.00
L hi _3	**1.00**	**1.00**	1.00	1.00	**1.00**	**1.00**	1.00	1.00
L hi _4	3.00	**0.00**	1.00	1.00	**1.00**	**1.00**	1.00	1.00
L hi _5	**1.00**	**1.00**	1.00	1.00	**1.00**	**1.00**	1.00	1.00
L hi _6	3.00	2.98	1.00	1.00	**1.00**	**1.00**	1.00	1.00
L m _7	2.98	3.00	0.92	0.94	**1.00**	**1.00**	1.00	1.00
L m _8	3.00	3.00	1.00	1.00	2.00	2.00	1.00	1.00
L m _9	3.00	3.00	1.00	1.00	2.00	2.00	1.00	1.00
L m 10	2.97	2.96	1.00	1.00	**1.00**	**1.00**	1.00	1.00
L m 11	3.00	3.00	1.00	1.00	2.00	2.00	1.00	1.00
L m 12	2.85	3.00	1.00	1.00	**1.00**	**1.00**	1.00	1.00
L lo 13	3.00	3.00	1.00	1.00	1.98	2.00	0.98	0.96
L lo 14	2.90	3.00	0.92	0.92	1.96	2.00	0.97	0.99
L lo 15	2.85	2.93	0.85	0.86	1.85	1.89	0.92	0.92
L lo 16	3.00	3.00	1.00	1.00	2.00	2.00	1.00	1.00
L lo 17	3.00	3.00	1.00	1.00	2.00	2.00	1.00	1.00
L lo 18	3.00	3.00	1.00	1.00	1.84	1.78	0.89	0.92

Note. Quatry = Quaternary; Regular = Regular analysis; Cut = Cut elimination analysis; bolded values = those exceeding the critical value, *c*.

Table 63 displays the upper asymptote index initial approximations for the next 18 test items using keywords that are classified as non-loanwords.

Table 63

Upper Asymptote Index Initial Approximations Indicating Non-Loanword Keywords Eliciting Item Carelessness

Non Loan-word Item	Quatry 011223 RSM Extent $c < 2.25$ Regular	Cuthi	Binary 000001 DRM Size $c < 0.75$ Regular	Cuthi	Ternary 001122 RSM Extent $c < 1.34$ Regular	Cuthi	Binary 000011 DRM Size $c < 0.67$ Regular	Cuthi
N hi 19	3.00	3.00	1.00	1.00	2.00	2.00	1.00	1.00
N hi 20	3.00	**0.00**	1.00	1.00	2.00	2.00	1.00	1.00
N hi 21	3.00	3.00	1.00	1.00	1.80	1.65	0.93	1.00
N hi 22	3.00	3.00	1.00	1.00	1.94	1.92	0.98	1.00
N hi 23	3.00	3.00	1.00	1.00	1.95	1.89	0.98	1.00
N hi 24	2.88	2.85	0.90	0.96	1.91	1.98	0.95	0.95
N m 25	**0.00**	**0.00**	**0.00**	1.00	2.00	2.00	0.96	0.97
N m 26	3.00	3.00	1.00	1.00	2.00	2.00	1.00	1.00
N m 27	3.00	**0.00**	1.00	1.00	2.00	2.00	1.00	1.00
N m 28	3.00	**0.00**	1.00	1.00	2.00	2.00	1.00	1.00
N m 29	3.00	3.00	1.00	1.00	2.00	2.00	1.00	1.00
N m 30	3.00	3.00	1.00	1.00	2.00	2.00	1.00	1.00
N lo 31	3.00	**0.00**	1.00	1.00	2.00	2.00	1.00	1.00
N lo 32	3.00	**0.00**	1.00	1.00	2.00	2.00	0.92	0.94
N lo 33	3.00	**0.00**	1.00	1.00	2.00	2.00	1.00	1.00
N lo 34	3.00	3.00	1.00	1.00	2.00	2.00	1.00	1.00
N lo 35	3.00	3.00	1.00	1.00	2.00	2.00	1.00	1.00
N lo 36	3.00	**0.00**	1.00	1.00	2.00	2.00	1.00	1.00

Note. Quatry = Quaternary; First = First analysis; Cut = Cut elimination analysis; bolded values = those exceeding the critical value, *c*.

Consolidating the relevant information from Table 62 and 63 forms the summary in Table 64. Therefore, in Table 64 the upper asymptote index initial approximations that indicate item carelessness from the keywords classified as loanwords and non-loanwords are consolidated into one table. Items Lhi_3, Lhi_4 and Lhi_5 are anomalies, as they are flagged in both the baseline quaternary RSM 011223 and ternary DRM 001122 scales.

Table 64

Upper Asymptote Index Initial Approximations Indicating Loanwords and Non- Loanword Keywords Eliciting Item Carelessness

Quaternary 011223 RSM Extent *c* < 2.25			Binary 000001 DRM Size *c* < 0.83			Ternary 001122 RSM Extent *c* < 1.34			Binary 000011 DRM Size *c* < 0.67		
Item	Reg	Cuthi	Item	Reg	Cuthi	Item	Reg	Cuthi	Item	Reg	Cuthi
Lhi_3	**1.00**	**1.00**	Nm25	0.00	1.00	Lhi_1	**1.00**	**1.00**	—	—	—
Lhi_4	3.00	**0.00**				Lhi_2	**1.00**	**1.00**			
Lhi_5	**1.00**	**1.00**				Lhi_3	**1.00**	**1.00**			
Nhi20	3.00	**0.00**				Lhi_4	**1.00**	**1.00**			
Nm25	**0.00**	**0.00**				Lhi_5	**1.00**	**1.00**			
Nm27	3.00	**0.00**				Lhi_6	**1.00**	**1.00**			
Nm28	3.00	**0.00**				Lhi_7	**1.00**	**1.00**			
Nlo31	3.00	**0.00**				Lm10	**1.00**	**1.00**			
Nlo32	3.00	**0.00**				Lm12	**1.00**	**1.00**			
Nlo33	3.00	**0.00**									
Nlo36	3.00	**0.00**									

Note. Reg = regular analysis, Cut = cut elimination analysis, *c* = upper asymptote initial approximation critical value, *c*.

For the quaternary RSM 011223 baseline scale, eleven items were flagged as eliciting careless responses in the cut analysis. The first flagged item, Lhi_3 (*c* = 1.00), had a higher asymptote value of *c* < 2.25; however, checking the Winsteps most expected responses and most misfitting response string tables for item Lhi_3 in the regular analysis indicated that no persons responded carelessly. Moreover, in the subsequent cut analysis, no persons were flagged as eliciting careless responses. The second flagged item, Lhi_4 (*c* = 0.00), had a higher asymptote value of *c* < 2.25 in the cut analysis. However, checking the Winsteps most expected responses and most misfitting response string tables for item Lhi_4 in the cut high analysis indicated that no persons responded carelessly. The third flagged item, Lhi_5 (*c* = 1.00), had a higher asymptote value of *c* < 2.25. However,

checking the Winsteps most expected responses and most misfitting response string tables for item Lhi_5 in the first analysis indicated that no persons responded carelessly. Also, in the subsequent cut high analysis, no persons responded carelessly. The fourth flagged item, Nhi20 (c = 0.00), had a higher asymptote value of $c < 2.25$ in the cut high analysis. However, checking the Winsteps most expected responses and most misfitting response string tables for item Nhi20 in the cut high analysis indicated that no persons responded carelessly. The fifth flagged item, Nm25 (c = 0.00), had a higher asymptote value of $c <$ 2.25 in both the first and cut high analyses. However, checking the Winsteps most expected responses and most misfitting response string tables for item Nm25 in both first and cut high analyses indicated that the same person (B050 with a '0' scored incorrect response in both the first and cut high analyses) responded carelessly. The sixth to eleventh upper asymptotes were flagged only in the cut high analysis. The sixth flagged item, Nm27 (c = 0.00), had a higher asymptote value of $c < 2.25$ in the cut high analysis. However, checking the Winsteps most expected responses and most misfitting response string tables for item Nm27 in the cut high analysis indicated that only one person (C189 with a '3' scored correct response) responded carelessly. The seventh flagged item, Nm28 (c = 0.00), had a higher asymptote value of $c < 2.25$ in the cut high analysis. However, checking the Winsteps most expected responses and most misfitting response string tables for item Nm28 in the cut high analysis indicated that no persons responded carelessly. The eighth flagged item, Nlo31 (c = 0.00), had a higher asymptote value of $c < 2.25$ in the cut high analysis. However, checking the Winsteps most expected responses and most misfitting response string tables for item Nlo31 in the cut high analysis indicated that two persons

(B264 with a ‘0’ scored incorrect response and C138 with a ‘3’ scored correct response) responded carelessly. The ninth flagged item, Nlo32 (c = 0.00), had a higher asymptote value of $c < 2.25$ in the cut high analysis. However, checking the Winsteps most expected responses and most misfitting response string tables for item Nlo32 in the cut high analysis indicated that only one person (A011 with a ‘0’ scored incorrect response) responded carelessly. The tenth flagged item, Nlo33 (c = 0.00), had a higher asymptote value of $c <$ 2.25 in the cut high analysis. However, checking the Winsteps most expected responses and most misfitting response string tables for item Nlo33 in the cut high analysis indicated that two persons (C276 and C189 both with ‘3’ scored correct responses) responded carelessly. The eleventh flagged item, Nlo36 (c = 0.00), had a higher asymptote value of $c < 2.25$ in the cut high analysis. However, checking the Winsteps most expected responses and most misfitting response string tables for item Nlo36 in the cut high analysis indicated that only one person (C138 with a ‘3’ scored correct response) responded carelessly.

For the binary DRM 000001 scale, one item elicited careless responses in the regular analysis. Checking the Winsteps most expected responses and most misfitting in the response string tables for the regular analysis indicated that seven persons, C074, C092, C134, C147, A222, B244 and F399 all with a response of “1” were responsible for careless responses.

For the ternary RSM 001122 scale, nine items elicited careless responses in both the regular and cut analyses. All nine items, Lhi_1, Lhi_2, Lhi_3, Lhi_4, Lhi_5, Lhi_6, Lhi_7, Lm10 and Lm12 had the same critical value (c = 1.00) indicating a substantial lower asymptote value of $c < 1.34$. Checking the Winsteps most expected responses and most

misfitting in the response string tables for the regular analysis indicated that no persons were responsible for careless responses.

In summary of all four upper asymptote item carelessness analyses, the quaternary RSM 011223 scale had eleven items that elicited instances of carelessness after the cut high elimination procedure. The binary DRM 000001 scale indicated that one item elicited careless responses. The ternary RSM 001122 scale had nine items that elicited careless responses after the cut high elimination procedure. The binary DRM 000011 scale indicated that no items elicited careless responses.

CHAPTER 6

DISCUSSION

The results for the six research questions are discussed sequentially.

Research Question 1: Rasch Analysis Requirements

Research question 1 asked whether the test met the unidimensional Rasch model requirements. Overall, the test indicated it appropriately matched the requirements of the Rasch model measurement family. As the requirements were met, it is appropriate to use the item difficulty calibrations, person ability measures and threshold difficulty measures to conduct parametric and nonparametric statistical procedures to gain further insight into the test's usefulness. In particular, quality control of the test indicated that the test measured just one unidimensional construct.

In a preliminary validation of the Vocabulary Size Test, Beglar (2010) indicated that the test met the requirements of the dichotomous Rasch model. The quality control standardized infit statistic criterion was met by most of the 140 items in the dataset. Substantive construct validity was checked by confirming the hypothesis that progressively lower word frequency levels were aligned using descriptive statistics with increased item difficulty calibrations. Substantive predictive validity was assessed by confirming the hypothesis that respectively low, medium, high non-native English speakers and native English speakers were aligned using descriptive statistics with increased person ability measures of vocabulary size. The tests of unidimensionality indicated that the vocabulary

size construct was the sole dimension measured. Invariance of the measures was investigated in four ways to document generalizability validity evidence.

Research Question 2: The Difficulty of Loanwords and Non-Loanwords

Research question 2 asked whether the hypothesized easiness of loan words and difficulty of non-loan words was confirmed by the data (Palmer, 1917, amended). The a priori hypothesized easiness of loanwords and difficulty of non-loanwords was confirmed by the empirical results. The empirical results indicated that the item stem keywords that were a priori classified as loanwords according to dictionary reference were easier to answer correctly compared to keywords that were classified as non-loanwords. Figures 59, 60 and Table 55 in the Results chapter display these findings.

Palmer (1917) proposed that frequently encountered words are easier to learn than rarely encountered words. In this study, an amendment to Palmer's (1917) word frequency theory was proposed whereby English to Japanese loanwords are more frequently encountered than non-loanwords in the Japanese English as a foreign language context. Therefore, as loanwords are more frequently utilized, English to Japanese loanwords are easier to acquire for Japanese learners of English. However, this is not always the case as often loanwords into Japanese are written with katakana characters, so, if the learner reads a loanword written in English they might be unaware of its loanword status. The findings indicate that the amendment to Palmer's (1917) word frequency theory is confirmed.

In a study constructing and assessing two bilingual (English-Russian) versions of the Vocabulary Size Test, Elgort (2012) investigated the cognate (English-Russian)

keyword status advantage effect. Two 70-item bilingual Vocabulary Size Tests versions were paired with two monolingual English Vocabulary Size Test versions administered in a counterbalanced design to 121 native speakers of Russian. Elgort hypothesized that the probability of correctly responding to cognate (English-Russian) keywords would be higher than non-cognates. Further, Elgort hypothesized that there is a larger effect of cognate advantage for lower frequency words than for higher frequency words. Elgort's results indicated that for both the monolingual (English) and bilingual (English-Russian) Vocabulary Size Test versions, items with cognate keywords were easier than non-cognates, especially at higher word frequency levels. However, the cognate keyword advantage gradually disappeared as keywords progressed from higher to lower word frequency. Elgort (2012) suggests that this occurs because intermediate learners like those in her study have solid knowledge of high frequency words but marginal knowledge of low frequency words unless the word was a loanword.

The findings of the present study are similar to Elgort's (2012). The results of the present study indicate that English-Japanese loanwords were easier to answer successfully than non-loanwords, thus supporting the Elgort's finding. As shown in Figure 67, there is a relatively large difference between the item difficulty calibrations of loanwords in the high frequency band (first box and whiskers from the left in Figure 67) and non-loanwords in the high frequency band (fourth box and whiskers from the left). There is a relatively small difference between the item difficulty calibrations of loanwords in the medium frequency band (second box and whiskers from the left in Figure 67) and non-loanwords in the medium frequency band (fifth box and whiskers from the left in Figure 67). However, there

is a relatively medium difference between the item difficulty calibrations of loanwords in the low frequency band (third box and whiskers from the left in Figure 67) and non-loanwords in the low frequency band (sixth box and whiskers from the left in Figure 67).

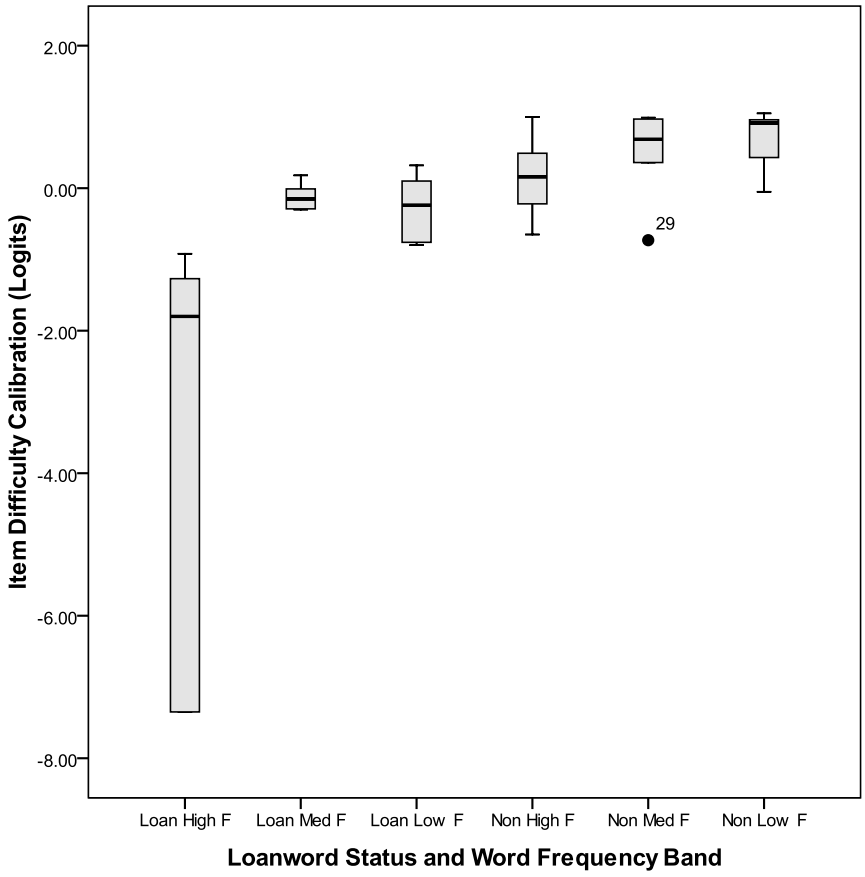


Figure 67. Item stem keyword difficulty calibrations classified according to loanword status and word frequency band.

These findings from the present study suggest that there is a substantive difference leading to a loanword advantage for assisting persons to successfully answer with a correct response. However, once the loanword status factor is combined with the word frequency

factor, there is a relatively large difference when comparing the loanwords and non-loanwords at the high word frequency band. Further, there is a relatively small difference when comparing loanwords and non-loanwords at the medium frequency band and a medium difference when comparing loanwords and non-loanwords at the low frequency band. The loss of the loanword advantage is speculated to occur because of a plateau effect, similar to a diminishing returns effect, of word frequency bands in the present study. In other words, as learners develop, word frequency has a large effect from high to medium and medium word frequency bands, however, this word frequency advantage levels off between medium and low word frequency bands. This issue is expanded upon in the next section, the discussion of research question three's results.

Another point of interest from Elgort's (2012) study is the notion of the proportion of cognates in the Vocabulary Size Test in comparison with the British National Corpus (BNC) (Leech et al., 2001) base lists. The Vocabulary Size Test versions used in Elgort's (2012) study had a 34% average proportion of cognates while the target language, that is, English, as represented by the British National Corpus (BNC) base lists had a 27% average proportion of cognates. Therefore, the receptive size estimates produced by these two versions of the Vocabulary Size Test might overestimate the participants' actual vocabulary sizes. The average vocabulary size of the participants in Elgort's study was 6,260 word-families, however because of over representation of cognates, that is, 34% in the English-Russian bilingual Vocabulary Size Test instead of 27% in the British National Corpus, Elgort (2012) reported that the average vocabulary size can be overestimated.

In the present study, loanwords composed 50% (18 of the 36 items) of the test. The rationale underlying this high representation was the a priori hypothesis that the loanword advantage might help illustrate the construct validity in the form of the item difficulty calibration order hierarchy that is produced a posteriori with Rasch model analyses. If typical English to Japanese loanword representation in the test-takers' first language, Japanese, was adhered to, only about 10% loanword representation (Daulton, 2008; Tomoda, 2009) would be allowable. However, if this criterion was implemented, out of a total of 36 items in the test, only four items would be included in the test. Statistically, comparing the means of 4 loanword keyword items versus 32 non-loanword keyword items is not feasible because of the under representation of loanwords. Even if Elgort's (2012) 34% tested representation was used, this would only allow 12 loanword keyword items versus 24 non-loanword keyword items. Therefore, a balanced design utilizing 18 loanword and 18 non-loanword keyword items was used in this study to allow for equal representation for parametric statistical analysis. However, as pointed out by Elgort (2012), the over representation of loanwords leads to easy test items and an overestimation of the measures produced. The persons in this sample seemed to find the items easy as the mean of the person ability measures distribution was located higher than the mean of the item difficulty calibration distribution in the Wright Map.

In order to assess how well test-takers recognized the keyword loanword status, a yes/no checklist survey was administered to a sub-sample of 173 persons from the original 379 person PAVET sample. The original PAVET was arranged in two blocks with an item keyword sequence in which items 1 to 18 were loanwords and items 19 to 36 were non-

loanwords. However, this sequence was not suitable for the loanword status recognition yes/no checklist survey. Therefore, the sequence of the loanword status recognition yes/no checklist survey keywords was scrambled in a random-like order, with for example, a loanword, a non-loanword, a loanword, a non-loanword, and a non-loanword. The test-takers were instructed to indicate with a yes or no if the keyword was a loanword or not. To ease clarity of analysis and mirror the PAVET item numbers, the loanword status recognition yes/no checklist survey items were renumbered into the PAVET item numbers.

As displayed in Table 65, the most difficult loanword status recognition keywords were the ten loanwords: involve, mature, depression, deduction, distortion, sustainable, authority, refine, export and portion. The easiest loanword recognition keywords were the five loanwords: relax, positive, volume, link and monitor. The non-loanwords were in between the two loanword extremes of this loanword status recognizability scale.

In summary, ten of the loanwords were difficult to recognize, five of the loanwords were easy to recognize and the non-loanwords were of medium recognizability in regards to loanword status recognition.

Table 65

Loanword Status Recognition Yes/No Checklist Survey by Item Difficulty Calibration Hierarchy

Item	Keyword	Loan word	Non-loan word	Measure	*SE*	Infit MNSQ	Outfit MNSQ	Pt-measure correlation
8	involve	√		5.20	0.46	1.06	3.08	-.28
9	mature	√		3.87	0.26	1.05	1.84	-.03
12	depression	√		3.87	0.26	1.10	2.59	-.22
16	deduction	√		3.81	0.25	1.12	2.13	-.21
13	distortion	√		3.75	0.25	1.03	1.92	-.04
15	sustainable	√		3.63	0.24	1.03	1.50	.07
6	authority	√		2.71	0.18	0.94	1.12	.28
17	refine	√		2.47	0.17	1.11	1.49	-.05
2	export	√		1.79	0.16	0.99	1.05	.25
10	portion	√		1.64	0.16	1.03	1.07	.21
22	attitude		√	0.38	0.18	0.99	0.99	.26
30	evaluation		√	0.18	0.19	0.97	0.98	.28
20	affect		√	0.14	0.19	0.97	1.04	.25
35	unify		√	-0.22	0.21	0.89	0.84	.40
18	simulation	√		-0.27	0.22	1.04	1.07	.15
36	contradiction		√	-0.42	0.23	1.00	1.07	.20
26	commence		√	-0.47	0.23	0.88	0.75	.42
28	thesis		√	-0.58	0.24	0.98	0.83	.28
27	consistent		√	-0.89	0.27	0.90	0.81	.35
29	derive		√	-0.89	0.27	0.95	0.83	.29
34	deviation		√	-1.13	0.29	0.90	0.67	.38
23	survey		√	-1.32	0.32	1.00	1.25	.10
11	pose	√		-1.32	0.32	1.07	1.26	.00
24	equipment		√	-1.54	0.35	0.89	0.64	.36
32	invoke		√	-1.54	0.35	0.92	0.83	.31
21	significant		√	-1.66	0.37	0.85	0.53	.43
19	consequence		√	-1.97	0.42	0.89	0.86	.30
31	inhibition		√	-1.97	0.42	0.92	0.58	.31
33	incompatible		√	-1.97	0.42	0.90	0.45	.37
7	license	√		-2.16	0.46	1.00	0.71	.18
25	exploitation		√	-2.40	0.51	0.87	0.38	.39
14	monitor	√		-3.11	0.71	0.93	0.39	.27
1	link	√		-3.81	1.00	1.00	0.87	.04
4	volume	√		-3.81	1.00	0.93	0.13	.29
3	positive	√		-5.02	1.83	Min	Min	.00
5	relax	√		-5.02	1.83	Min	Min	,00
M				-0.28	0.42	0.97	1.07	
SD				2.59	0.39	0.07	0.62	

Research Question 3: Word Frequency Acquisition Theory

Research question 3 asked whether Palmer's (1917) word frequency acquisition theory was confirmed by the data. Palmer's (1917) word frequency acquisition theory was partly confirmed by the data in that keywords selected from the high word frequency band were easier to answer successfully than those from the medium word frequency band. However, the keywords selected from the medium word frequency band were not easier or more difficult to answer successfully than those from the low word frequency band.

These findings are not consistent with previous findings. Elgort's (2012) preliminary analysis investigated British National Corpus (BNC) (Leech, et al, 2001) word frequency by English monolingual and English-Russian bilingual Vocabulary Size Test raw scores over 14 frequency levels, that is, the 1,000 to 14,000 word frequency levels. Overall, the raw score mean number of correct responses decreased as keywords progressed from the 1,000 through to the 14,000 word frequency levels. However, as these mixed-effects regression analysis results were based on data from Classical Test Theory (True Score Model) ordinal-level raw scores, the results are non-linear and not invariant.

Beglar (2010) used a linear Rasch measurement approach to investigate this word frequency hypothesis via an English monolingual version Vocabulary Size Test. Item difficulty calibrations were interval-level measured and conflated into means for each of the 14 British National Corpus (BNC) word frequency levels. Despite plateaus especially between the 2,000, 3,000, and 4,000 word frequency levels and the 11,000, 12,000, 13,000 and 14,000 word frequency levels, overall, the first 1,000 word frequency level was the

easiest and the 14,000 word frequency level was the most difficult. These descriptive statistics findings provided support for Palmer's (1917) word frequency theory.

However, a caveat should also be kept in mind that, "frequency level is not a perfect indicator of which words are likely to be known" (Nation & Beglar, 2007, p. 11). Therefore, word frequency level is not the be all and end all of predicting vocabulary difficulty. Indeed, "there are other factors involved in knowing a word and frequency counts can give differing results depending on the size and nature of the corpus used" (Nation & Beglar, 2007, p. 11), and "Frequency of occurrence however is only one factor, although a very important one, affecting order of acquisition," (Nation & Beglar, 2007, p. 11). Thus, differing results found in this study might have occurred because of the word corpuses utilized. In this study the initial word corpus source of test keywords was the Academic Word List (Coxhead, 2000). Words were classified by loanword status and JACET 8000 academic word list frequency bands (high word frequency, 1,000 to 3,000 word lists, medium word frequency, 4,000 to 5,000 word lists, and low frequency, 6,000 to 8,000 word lists) to match the Japanese English as a foreign language context. In effect, this probably reduced the number of difficult words on the test; thus, this characteristic of the test contributed to the overall ease of the test items.

In the present study, there was a lack of item difficulty calibration difference between medium and low frequency key words. This might be due to the non-random sampling of items and/or persons in this study. Another contributing factor might be the use of ordered triple responses and the rating scale model, for example, RSM 011223 scale in this study, which contributed to a consolidation of item difficulty calibration and person

ability measure distributions. This is in contrast to the use of single best answer responses and the dichotomous Rasch model, for example DRM 000011 in this study and Beglar's (2010) study, where the stricter scoring has the effect of separating item difficulty calibration and person ability measure distributions into wider spread scales. Beglar's (2010) study indicated differences between high, medium and low frequency key word item difficulties at an even greater micro scale. Therefore, as the present study did not detect a difference between medium and low frequency key words, this indicates a problem in the sampling of the items and/or the persons.

Research Question 4: Metasynonymy Awareness Hypothesis

Research question 4 asked whether the metasynonymy awareness hypothesis based on an amendment to O'Connor's (1940) synonymy acquisition theory was confirmed by the data. The metasynonymy awareness hypothesis was partly confirmed by the data. The metasynonymy awareness hypothesis proposed in this study predicted that there are six categories developmentally ordered from null to full, respectively:

(a) synonym/decontextualized/null, (b) decontextualized/synonym/null,

(c) synonym/null/decontextualized, (d) null/synonym/decontextualized,

(e) decontextualized/null/synonym, and (e) null/decontextualized/synonym.

However, due to disordered thresholds, these six categories were collapsed and rescored into four categories developmentally ordered from null to full, respectively:

(a) synonym/decontextualized/null, (b) decontextualized/synonym/null, and synonym/null/decontextualized, (c) null/synonym/decontextualized, and decontextualized/null/synonym, and (d) null/decontextualized/synonym.

When the categories were collapsed and rescored into the four categories as displayed in Table 66, the metasynonymy awareness hypothesis based on O'Connor's (1940) amended synonymy acquisition theory was confirmed. The three thresholds between the four categories were empirically ordered, with estimated values of -0.74 for threshold one, -0.36 for threshold two and 1.10 for threshold three. The findings in this study are different from Smith's (1987) study in which the data indicated disordered thresholds.

Table 66

Metasynonymy Awareness Hypothesis: Comparing A Priori Prediction and A Posteriori Realization Scoring Guide

A priori hypothesized six categories	Test-taker ordered triple rank (OTR) response	Collapsed and rescored four categories
5 Full	(null, decontextualized, synonym)	3 Full
4 Partial a)	(decontextualized, null, synonym)	2 Partial
3 Partial b)	(null, synonym, decontextualized)	2 Partial
2 Marginal a)	(synonym, null, decontextualized)	1 Marginal
1 Marginal b)	(decontextualized, synonym, null)	1 Marginal
0 Null	(synonym, decontextualized, null)	0 Null

Smith (1987) attempted to construct a receptive multiple-choice vocabulary test in which English as a first language test-takers were instructed to respond with single best answer (SBA) responses. The test consisted of 50 multiple-choice items each with five options. A subset of 16 items was determined from the 50 items based on data fit to the

model and correlations between various vocabulary raw scores. The five options were based on O'Connor's (1940) five-stage theory of vocabulary acquisition. The data were analyzed using the Rasch-Andrich Rating Scale Model (Andrich, 1978). One of the reasons that Smith (1987) chose to use the Rating Scale Model over the Partial Credit Model was that it:

> is more in keeping with the vocabulary learning theory developed by O'Connor, which has five ordered learning stages. These stages have the same structure progressing from low knowledge to high knowledge, and the order should be independent of the stimulus word and the difficulty of the item. (p. 221)

However, despite this rationale, the four thresholds between the five categories were disordered. The first threshold's estimated value was 0.33, the second threshold was 0.23, the third threshold was -0.06 and the fourth threshold was -0.50. Not only were the thresholds disordered, but, they were also in reversed order, that is, 0.33, 0.23, -0.06 and -0.50. Reversed and disordered thresholds are not congruent with statistical statistics that are a crucial requirement for polytomous Rasch models. Despite the disordered and reversed thresholds, Smith (1987) stated that the "results also indicate moderate support for O'Connor's theory of vocabulary acquisition" (p. 217). It would have been useful if Smith (1987) had collapsed and rescored the categories and then run the analysis again. If the subsequent analysis had displayed ordered thresholds, Smith (1987) would have had an opportunity to amend O'Connor's substantive vocabulary acquisition theory. As Smith (1987) constructed a test with a single best answer instruction, on the next iteration of test administration, the number of options could be reduced from five to four or three options.

In fact, Andrich (2013b) recommended administering an updated instrument, for example, with collapsed categories, that is, a reduced number of options, to a new sample. However, in most cases Andrich recommends this to instrument developers who are constructing self-report attitudinal surveys that utilize Likert scales. Therefore, in the case of the present study's test design consisting of a fixed three-by-three matrix for each item response outcome space, administering the test to a new sample with a reduced number of options is not feasible. Therefore, in the present study the constructed outcome space design was retained and categories were collapsed and rescored for analyses.

Despite Smith's (1987) adherence to disordered thresholds, empirically ordered thresholds are what Andrich (1978, 2013a) deems mandatory, especially if researchers aim to develop and interpret polytomous Rasch models in the spirit of Ronald A. Fisher's legacy (Andrich, 2013b). This includes three of Fisher's legacies: first, the creation of sufficient statistics, second, acknowledging that ordered thresholds need to be an empirical characteristic of the data, and third, the consolidation of empirical study design and data statistical analyses. In other words, the presence of disordered thresholds in the data is problematic for polytomous Rasch measurement model analyses. When disordered categories are in the data, sufficient statistics cannot be constructed as the data is not invariant. Therefore, ordered thresholds should be an a priori characteristic of the data that can be tested empirically. Furthermore, during data analysis if anomalies are present in the data, for example, disordered thresholds, these categories can be then rescored and reanalyzed. Subsequent test administrations would ideally take this information into consideration when amending future test designs.

Figure 68 displays the three Andrich thresholds in relation to three persons, D357, the lowest ability person, E396, the mean ability person, and A216, the highest ability person. The location where the thresholds cross the particular person's vertical dotted line is the location where that person has approximately a 50% probability of crossing the

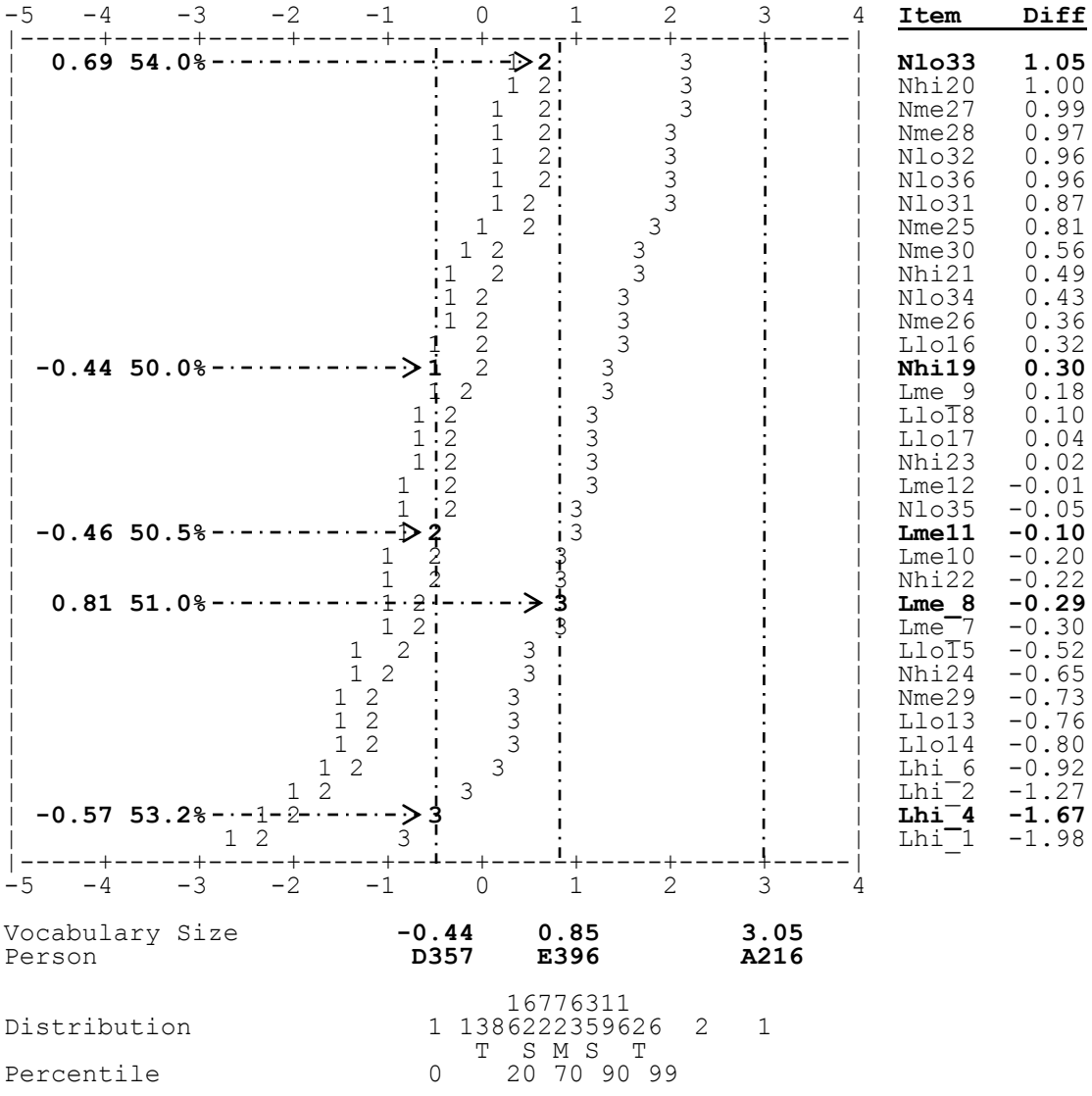


Figure 68. Andrich thresholds in relation to person ability and item difficulty.

threshold into the higher category. For example, for person D357 answering the easy item Lhi_4, they have a 53.2% probability of successfully crossing its third threshold.

However, for person D357 responding to the more difficult item Nhi19, they have a 50% probability of only crossing its first (i.e., easiest) threshold. In between these two extremes, for person D357 when answering item Lme11, they have a 50.5% probability of crossing the second (i.e. middle) threshold. For person E396 at the mean of the person ability distribution, when answering item Lme_8, they have a 51.0% probability of crossing the third threshold. However, when person E396 answers the most difficult item, Nlo33, they have a 54.0% probability of crossing the second threshold. In the case of the highest ability person, A216, they have surpassed the thresholds for the items used in this test, therefore, they have a stronger than 50% probability of crossing all thresholds for all items.

Alternatively, this information and more is displayed in Tables 67 and 68. Comparisons can be made between the easiest item, Lhi_1, the mean item, Lme12, and the most difficult item, Nlo33. Also comparisons can be made between the lowest ability person, D357, the mean ability person, E396, and the highest ability person, A216. The three individual person's probabilities on the three thresholds of each of the three items can be compared and contrasted. Rasch (1977/2006) and Stone and Stenner (2014) have emphasized the importance of comparisons for objectivity when making scientific statements. Tables 67 and 68 both enable comparisons in a 3-by-3 equaling 9 values for these two success probability matrices.

Lexical space is a concept Daller, Milton and Treffers-Daller (2007) used to describe the three aspects of knowing a word. The three proposed aspects were vocabulary

size, strength and fluency. In terms of this vocabulary extent study using a polytomous Rasch-Andrich rating scale model (Andrich, 1978), the fluency aspect can be indicated by the person's probability of success at a particular item threshold. In Tables 67 and 68 the comparisons between the lowest, mean and highest ability persons at the easiest, mean and most difficult items at the first, second and third thresholds might give some insight to Daller et al.'s (2007) lexical space hypothesis. Tables 67 and 68 display the same values rearranged for clarity of comparison.

Table 67

Success Probabilities for the Lowest, Mean and Highest Difficulty Items for the Easiest, Mean and Highest Ability Persons at Thresholds 1, 2, and 3

		Threshold 1		Threshold 2		Threshold 3	
Item	Person	Item difficulty plus *T*1 (-0.74)	Success probability at *T*1 %	Item difficulty plus *T*2 (-0.36)	Success probability at *T*2 %	Item difficulty plus *T*3 (1.10)	Success probability at *T*3 %
Nlo33	A216	0.31	93.9	0.69	91.4	2.15	71.1
Nlo33	E396	0.31	63.2	0.69	54.0	2.15	21.4
Nlo33	D357	0.31	32.1	0.69	24.4	2.15	7.0
Lme12	A216	-0.75	97.8	-0.37	96.8	1.09	87.7
Lme12	E396	-0.75	83.2	-0.37	77.2	1.09	44.0
Lme12	D357	-0.75	57.7	-0.37	48.3	1.09	17.8
Lhi_1	A216	-2.72	99.7	-2.34	99.5	-0.88	98.1
Lhi_1	E396	-2.72	97.3	-2.34	96.0	-0.88	85.0
Lhi_1	D357	-2.72	90.7	-2.34	87.0	-0.88	60.8

Note. T_1 = Threshold 1; T_2 = Threshold 2; T_3 = Threshold 3. Lowest ability person is D357,mean ability person is E396 and highest ability person is A216. The lowest in difficulty item is Lhi_1, the mean difficulty item is Lme12 and the most difficult item is Nlo33.

Table 68

Success Probabilities of Lowest, Mean and Highest Ability Persons for the Easiest, Mean and Most Difficult Items at Thresholds 1, 2 and 3

		Threshold 1		Threshold 2		Threshold 3	
Person	Item	Item calibration plus *T*1 (-0.74)	Success Probability at *T*1 %	Item calibration plus *T*2 (-0.36)	Success probability at *T*2 %	Item calibration plus *T*3 (1.10)	Success probability at *T*3 %
A216	Nlo33	0.31	93.9	0.69	91.4	2.15	71.1
A216	Lme12	-0.75	97.8	-0.37	96.8	1.09	87.7
A216	Lhi_1	-2.72	99.7	-2.34	99.5	-0.88	98.1
E396	Nlo33	0.31	63.2	0.69	54.0	2.15	21.4
E396	Lme12	-0.75	83.2	-0.37	77.2	1.09	44.0
E396	Lhi_1	-2.72	97.3	-2.34	96.0	-0.88	85.0
D357	Nlo33	0.31	32.1	0.69	24.4	2.15	7.0
D357	Lme12	-0.75	57.7	-0.37	48.3	1.09	17.8
D357	Lhi_1	-2.72	90.7	-2.34	87.0	-0.88	60.8

Note. T_1 = Threshold 1; T_2 = Threshold 2; T_3 = Threshold 3. Lowest ability person is D357, mean ability person is E396 and highest ability person is A216. The lowest in difficulty item is Lhi_1, the mean difficulty item is Lme12 and the most difficult item is Nlo33.

Another alternative to visualizing vocabulary extent is displayed. The lexical space (Daller, Milton & Treffers-Daller, 2007) extent of four persons can be visualized in Figure 69, in which the scale has been transformed from a logit into a CHIPs scale. A CHIPs scale changes all values into positive values and the mean of items is centered at a mean of 50.0. In Figure 69, the horizontal axis represents the person's vocabulary size and the vertical axis represents the person's vocabulary strength. The person's vocabulary size horizontal axis represents the same values as read from the person analysis report or from the person distribution on the left hand side of a typical Wright map. The person's vocabulary extent is represented by spatial area going up the vertical axis to the intercept point where the person vocabulary size dotted line crosses the third (last) threshold, indicated by a "3" number is aligned to the corresponding item in the item difficulty hierarchy. By subtracting the value

of the easiest item's (Lhi_1) third threshold, 46.0, the strength value of a person can be determined.

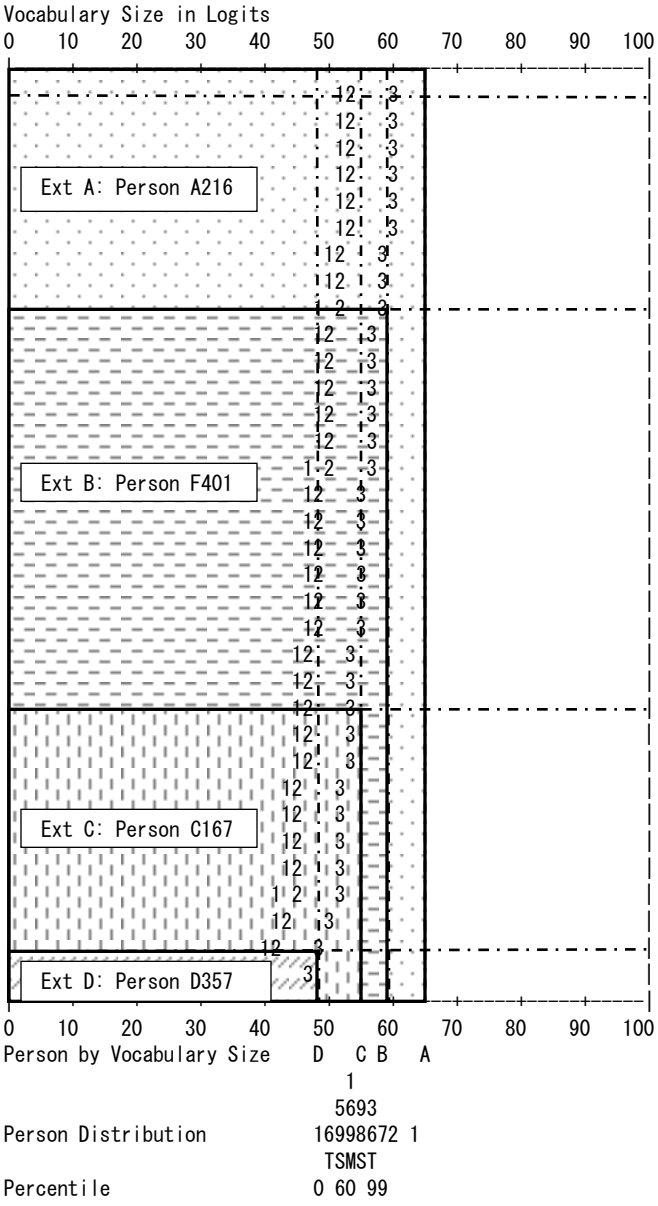
Tentatively, this might be a way of assessing an individual person's size, strength and extent. Table 69 displays the tentative estimated values for four persons. It should be noted that these estimates are based on the third (i.e., most difficult) threshold only. These extent estimates are displayed in Figure 69.

Table 69

Tentative Estimates of Vocabulary Extent

Person ability rank	Person	Person's vocabulary size	Item difficulty rank	Item	Item's third threshold	Person's vocabulary strength	Person's vocabulary extent
1	A216	63.9	1	Nlo33	59.8	13.8	882
3	F401	58.5	6	Nme30	57.6	11.6	679
32	C167	53.8	21	Lme_8	53.7	7.7	414
71	D357	48.0	30	Lhi_4	47.7	1.7	82

There is a possibility that there might be a cognitive linguistic threshold at which point learners can visualize or imagine an alternative sentence context for an alternative sense of a keyword. This means that test-takers below a certain proficiency level probably cannot successfully complete a traditional item designed to elicit single best answer, while those test-takers above a certain proficiency level probably can successfully complete a traditional item with a single best answer. This might affect the results in this present study as test-takers are administered items designed to elicit non-traditional ordered triple rank responses that are scored on an incremental polytomous scale. Therefore, instead of only using a dichotomous scoring guide, polytomous scoring guides were also used. The use of



Item Rank	Item Name	Third Threshold	Standard Error
1	**Nlo33**	**59.8**	.1
2	Nhi20	59.5	.1
2	Nme27	59.5	.1
3	Nme28	59.4	.1
3	Nlo32	59.4	.1
3	Nlo36	59.4	.1
4	Nlo31	59.0	.1
5	Nme25	58.7	.1
6	**Nme30**	**57.6**	.1
7	Nhi21	57.3	.1
8	Nlo34	57.0	.1
9	Nme26	56.6	.1
10	Llo16	56.5	.1
11	Nhi19	56.4	.1
12	Lme_9	55.8	.1
13	Llo18	55.5	.1
14	Llo17	55.2	.1
15	Nhi23	55.1	.1
16	Lme12	55.0	.1
17	Nlo35	54.8	.1
18	Lme11	54.6	.1
19	Lme10	54.1	.1
20	Nhi22	54.0	.1
21	**Lme_8**	**53.7**	.1
22	Lme_7	53.6	.1
23	Llo15	52.6	.1
24	Nhi24	52.1	.1
25	Nme29	51.7	.1
26	Llo13	51.5	.1
27	Llo14	51.4	.1
28	Lhi_6	50.8	.1
29	Lhi_2	49.3	.1
30	**Lhi_4**	**47.7**	.1
31	Lhi_1	46.0	.1

Figure 69. Lexical space extent displaying Rasch-Andrich thresholds in relation to person vocabulary size and strength on item's third thresholds. Ext A = Extent A; Ext B = Extent B; Ext C = Extent C; Ext D = Extent D; A = Person A (A216); B = Person B (F401); C = Person C (C167); D = Person D (D357); M = Mean; S = Standard Deviation; T = Two Standard Deviations.

polytomous scoring guides as used in this present study might be better suited to measuring where individual test-takers are located in relation to the cognitive linguistic threshold of imagining an alternative sentence context for an alternative keyword sense.

In this study, it is assumed that the test-takers can imagine an alternative sentence in which the keyword alternative sense meaning is used. There is a possibility that the polytomous rating scale model does not function properly because some participants were unable to imagine a keyword alternative sense in an alternative sentence. Imagining a sentence is complex and productive capacity is a requirement. In O'Connor's (1940) study, he was researching native speakers of English. However, in this present study non-native speakers of English were investigated and it is uncertain if all the participants had this generative capacity to imagine a keyword alternative sense in an alternative sentence.

Research Question 5: Person Ability Estimates and Analytical Approach

Research question 5 asked how the incremental vocabulary knowledge estimates changed depending on the type of analysis utilized. Three relational analyses were investigated: (a) baseline quaternary RSM 011223 (extent) and binary ordered triple rank (OTR) response DRM 000001 (size) (lossless data compression), (b) baseline quaternary RSM 011223 (extent) and ternary proxy single best answer (SBA) response RSM 001122 (extent) (lossy data compression), and (c) baseline quaternary RSM 011223 (extent) and binary proxy single best answer (SBA) response DRM 000011 (size) (lossy data compression).

Lossless (Extent to Size) Data Compression

(a) baseline quaternary RSM 011223 (extent) and binary ordered triple rank (OTR) response DRM 000001 (size) (lossless data compression)

In the first comparison, lossless data compression person ability measures from the baseline RSM 011223 extent scale and the DRM 000001 scale are investigated. The results indicate a Pearson's correlation coefficient of $r = .86$. Figure 64 in the results chapter displays a scatterplot of a flattened elliptical distribution. Figure 64's scatterplot indicates that persons with low person ability measures under RSM 011223 scoring have relatively lower person ability measures under DRM 000001 scoring. Conversely, persons with high person ability measures under RSM 011223 scoring have relatively higher person ability measures under DRM 000001 scoring. Therefore, the DRM 000001 sale is more conservative or stricter than the more liberal or permissive RSM 011223 baseline scale. This makes sense as under DRM 000001 scoring, only the very correct responses, most probably from persons with high person ability measures, are scored as correct. However, under RSM 011223 scaling, even partially correct responses, possibly from persons with low person ability measures, are able to score a partial credit.

Rescoring to the DRM 000001 scale from the RSM 011223 baseline scale is not exactly linearly diagonal, that is, following the identity line, on Figure 64's scatterplot. The spread of persons around the line indicates that the DRM 000001 scale does not result in exactly the same person ability measure hierarchy as the baseline RSM 011223 scale.

Therefore, the person ability measures of this sample depend on whether the RSM 011223 baseline scale or the collapsed and rescored DRM 000001 scale is chosen.

Lossless (Extent) to Lossy (Extent) Data Compression

(b) baseline quaternary RSM 011223 (extent) and ternary proxy single best answer (SBA) response RSM 001122 (extent) (lossy data compression).

In the second comparison, lossless data compression person ability measures from the baseline RSM 011223 extent scale and the lossy data compression RSM 001122 scale are investigated. The results indicate a Pearson's correlation coefficient of $r = .88$. Figure 65 in the results chapter displays a scatterplot of a diagonal elliptical distribution that was generally overlaying the diagonal identity line. Figure 65's scatterplot indicates that there are three anomalous persons in the distribution.

Rescoring to the RSM 001122 scale from the RSM 011223 baseline scale in general follows a diagonal line, that is, it overlays the identity line in Figure 65's scatterplot. The spread of persons around the line indicates that the RSM 0011122 scale does not result in exactly the same person ability measure hierarchy as the baseline RSM 011223 scale. Therefore, the person ability measures of this sample depend on whether the RSM 011223 baseline scale or the collapsed and rescored RSM 001122 scale is chosen.

Lossless (Extent) to Lossy (Size) Data Compression

(c) baseline quaternary RSM 011223 (extent) and binary proxy single best answer (SBA) response DRM 000011 (size) (lossy data compression).

In the third comparison, lossless data compression person ability measures from the baseline RSM 011223 extent scale and the lossy data compression from the DRM 000011 scale are investigated. The results indicated a Pearson's correlation coefficient of $r = .78$.

Figure 66 in the results chapter displayed a scatterplot of a flattened cigar shaped distribution. Figure 66's scatterplot indicated that persons with low person ability measures under RSM 011223 scoring had relatively lower person ability measures under DRM 000011 scoring. Conversely, persons with high person ability measures under RSM 011223 scoring had relatively higher person ability measures under DRM 000011 scoring. Therefore, the DRM 000011 sale is more conservative or stricter than the more liberal or permissive RSM 011223 baseline scale. This makes sense as under DRM 000011 scoring, only the correct responses, most probably from persons with high person ability measures, are scored as correct. However, under RSM 011223 scaling, even partially correct responses, possibly from persons with low person ability measures, are able to score a partial credit.

The relationship of rescoring the DRM 000011 scale from the RSM 011223 baseline scale is not exactly linearly diagonal, that is, it does not line up with the identity line exactly on Figure 66's scatterplot. The spread of persons around the line indicates that the DRM 000011 scale does not result in exactly the same person ability measure hierarchy as the baseline RSM 011223 scale. Therefore, the person ability measures of this sample depend on whether the baseline RSM 011223 scale or the collapsed and rescored DRM 000011 scale is chosen.

In summary, the highest correlation (r = .88) is between the baseline RSM 011223 scoring and the RSM 001122 scoring. Therefore, despite the change from lossless (RSM 011223) to lossy (RSM 001122) data compression, possibly the fact that both of these two scoring scales are measuring vocabulary extent is an overriding factor. Further, both of

these scoring formats are also polytomous which in this study are based on more of the available information about the test-takers ability.

The next highest correlation (r = .86) is between the baseline RSM 011223 scoring and the DRM 000001 scoring. Possibly, there was a high correlation as both of these scoring scales utilize lossless data compression. Finally, there is a correlation (r = .78) between the baseline RSM 011223 scoring and the DRM 000011 scoring. There was a high correlation even though there was a change from lossless RSM 011223 to lossy DRM 000011 data compression. The movement from polytomous scoring that utilizes all of the possible information available to either of the dichotomous scoring formats that do not utilize all of the possible information available might also account for the slightly lower correlation coefficients.

Comparing alternative dichotomous rescoring using DRM 001 and DRM 011 has been examined by Wright and Masters (1982). In a liking-for-science questionnaire, 75 children responded to 25 science activities indicating their attitudes on a 3-point Likert scale, where 0 = dislike (visualized as a frowning face), 1 = not sure/don't care (neutral face) and 2 = like (smiling face). Two dichotomous scoring approaches were compared, one with scoring 001 in which only a smile is scored correct and 011 in which only a frown is scored incorrect. In a series of scatterplot figures for item difficulty calibrations and person ability measures, the 001 and 011 ordinal-level raw scores were found to follow a curve, not the diagonal identity line as might be expected. However, once the raw scores were transformed into interval-level logit measures, they followed the linear identity line.

Wright and Masters (1982) also investigated a physics knowledge multiple choice test in which 321 students responded to thirty items with three categories each. If the student did not respond correctly on the first attempt, they were allowed a second attempt. Three alternative scoring formats were proposed: the first attempt was dichotomously scored 001, the second attempt was scored 011, and an overall polytomous partial credit was scored 012. An analysis of the two dichotomously scored formats compared 001 and 011 scoring schemes. The comparison visually revealed an elliptical shape overlaying the diagonal identity line in a scatterplot figure.

The receptive vocabulary test scoring comparisons in the present study therefore were similar to the dichotomous comparisons (001 and 011 scoring) for the science attitude survey and the physics knowledge test examined by Wright and Masters (1982).

Research Question 6: Guessing and Carelessness

Research question 6 asked to what degree guessing and carelessness were apparent in the types of analyses utilized. Four scoring analyses were investigated: (a) the baseline quaternary ordered triple rank (OTR) response RSM 011223 (extent), (b) binary ordered triple rank (OTR) response DRM 000001 (size) (lossless data compression), (c) ternary proxy single best answer (SBA) response RSM 001122 (extent) (lossy data compression), and (d) binary proxy single best answer (SBA) response DRM 000011 (size) (lossy data compression).

Examining the lower asymptote initial approximation through Winsteps as shown in the Results Table 61, an initial check of the analyses indicated that very few items were

eliciting item guessability. In the quaternary RSM 011223 extent scale and the binary DRM 000011 size scale, no items elicited item guessing. In the binary DRM 000001 size scale and the ternary RSM 001122 extent scale only 3 and 6 items respectively might have contributed to item guessability.

The number of items and the number of people susceptible to guessing are displayed in Table 70. The differences between the four different scales were influenced by the data cleaning procedure that was initially conducted to minimize misfitting data in the responses, that is, in the item and person data.

Table 70

Guessing Extent

Rasch Model Scale	Number of Items	Number of Persons
RSM 011223	0	0
DRM 000001	3	0, 0, 0
RSM 001122	6	2, 5, 4, 0, 2, 2
DRM 000011	0	0

Under the RSM 011223 scale, no items or persons were identified as being susceptible to guessing. In the next two proxy scales that subsequently followed, some items and/or persons were identified as being susceptible to guessing. Under the DRM 000001 scale, 3 items and 0 persons were identified as being susceptible to guessing, while under the RSM 001122 scale, 6 items and 2, 5, 4, 0, 2 and 2 persons were respectively identified as being susceptible to guessing. Under the DRM 000011 scale, no items or persons were identified as being susceptible to guessing.

Examining the upper asymptote initial approximation through Winsteps as shown in the Results Table 64, an initial check of the analyses indicated that some items were eliciting item carelessness. In the quaternary RSM 011223 extent scale, eleven items were flagged as contributing to item carelessness. In the ternary RSM 001122 extent scale, nine items were flagged as contributing to item carelessness. It is speculated here that these items might be due to the data cleaning procedure that took place to correct for quality control fit to the Rasch model. In both the binary DRM 000011 size scale and the binary DRM 000011 size scale, no items elicited item carelessness.

The number of items and the number of people susceptible to carelessness are displayed in Table 71. The differences between the four different scales were influenced by the data cleaning procedure that was initially conducted to minimize misfitting data in the responses, that is, in the item and person data.

Table 71

Carelessness Extent

Rasch Model Scale	Number of Items	Number of Persons
RSM 011223	11	0, 0, 0, 0, 1, 1, 0, 2, 1, 2, 1
DRM 000001	1	7
RSM 001122	9	0, 0, 0, 0, 0, 0, 0, 0, 0
DRM 000011	0	0

Under the RSM 011223 scale, 11 items and a few persons were identified as being susceptible to carelessness. In the next two proxy scales that subsequently followed, some items and/or persons were identified as being susceptible to guessing. Under the DRM

000001 scale, 1 item and 7 persons were identified as being susceptible to carelessness, while under the RSM 001122 scale, 9 items and 0 persons were respectively identified as being susceptible to carelessness. Under the DRM 000011 scale, no items or persons were identified as being susceptible to carelessness.

These findings confirm what had been stated by Wright (1988), "Research into who guesses usually shows that only a small proportion of persons do much guessing" (p. 9). This study's findings confirm this assertion. Gershon (1994) has stated that the cuthi and cutlo procedures can be utilized to eliminate the effects of poor targeting. This involves the transformation of suspect responses into missing data. In Gershon's (1992a) case, a large-scale dataset involved 1,294 items administered to 7,711 students and in Gershon's (1994) case, a large-scale dataset involved 1,304 items administered to 7,711 students. In the 1992 case 110 of the 1,294 items were eliminated after the cut high / cut low procedure. As this present study has 36 items the idea of eliminating items is not feasible. However, the cut hi and cutlo procedures give good insight into identifying which items elicit guessing and carelessness.

Implications for Vocabulary Assessment

This study contributes to second language testing in several ways. The unique use of traditional multiple-choice test sentence stems and options coupled with non-traditional instructions to sequence all three options according to a synonymy ordered triple rank rule presents a new way of conceptualizing this assessment approach. It is hoped that this might inspire other researchers to try new ways of testing methodology.

Additionally, the scoring guide based on mathematical permutations presents a new method of scoring the ordered triple rank responses elicited from the test-takers. Other researchers might try to find an alternative method of scoring their test if no other method is suitable.

The amendment of O'Connor's (1940) vocabulary acquisition stage theory, leading to three stages, and the subsequent bifurcation of those three stages into six stages was well-founded. The resulting collapsing and rescoring of the six categories down to four categories facilitated meaningful measurement. Researchers using polytomous Rasch model analyses should also collapse and rescore analyses if disordered thresholds appear in the initial analyses.

CHAPTER 7

CONCLUSION

Summary of the Findings

The two separate yet vital anchors securing this study were vocabulary and mathematics research domains. The focus was on vocabulary extent, with a particularly strong emphasis on vocabulary strength and mathematics in the form of various mathematical, modern measurement and statistical procedures. Despite these concepts being apart, like separated Rasch model measurement parameters, they worked together in conjunction in this study to construct a useful outcome. To quote Wright's (1997) profound statement once again, "After language, our greatest invention is numbers" (p. 33). Humanity's two greatest inventions have helped navigate this study through some rough seas to reach a relatively safe harbor. However, the next destination is in sight and a new voyage is being planned. As shown in this study, the Rasch model measurement analyses process is an iterative cycle, therefore, the ongoing process is much more important than a temporary port of call.

The findings indicated important outcomes for the two original contributions to knowledge for the written receptive vocabulary knowledge construct research domain proposed in this study. To reiterate, the two original contributions to knowledge were the proposal of an a priori metasynonymy awareness hypothesis and the construction of the Polysemous Academic Vocabulary Extent Test.

The first important finding was there is preliminary evidence that the a priori metasynonymy awareness hypothesis was empirically confirmed after a posteriori rescoring of the data. In other words, instead of the original a priori hypothesized six categories, only four categories were operational with this sample. As this hypothesis was operationalized at the nano-analysis level as presented in Table 10, the construction of item options was very sensitive as these ranked alternatives were at the fine-grained level of inter-item (i.e., threshold) analyses. This metasynonymy awareness hypothesis was based on an amendment to O'Connor's (1940) vocabulary acquisition theory and is focused on the vocabulary strength aspect.

The second important finding was the Polysemous Academic Vocabulary Extent Test is tentatively classified as being psychometrically sound. The two macro aspects of vocabulary extent, vocabulary size and strength, both contributed to operationalize the extent construct in the test. However, in this study the particular focus was on the a priori metasynonymy awareness hypothesis, therefore, vocabulary strength was the focal point of constructing this test. In the testing paradigm, there are four basic categories of tests for assessment: diagnostic, achievement, placement and proficiency. The Polysemous Academic Vocabulary Extent Test has qualities that make it an efficient diagnostic test, answering the call by Cronbach (1942, 1943) for improved and efficient diagnostic vocabulary strength tests. Unlike other vocabulary tests that attempt to assess size and strength by using a test battery, for example, a four test battery by Ishii and Schmitt (2009), the Polysemous Academic Vocabulary Test is just one test, which can administered in one test sitting of approximately 30 minutes.

Referring back to the introduction of this study, this study's findings address Cronbach's (1942, 1943) and Bruton's (2007) lament of the lack of incremental vocabulary assessments. The findings of this study indicated that respondents possess incremental academic vocabulary knowledge. Also, the findings indicated that loanwords are generally easier than non-loanwords. These incremental knowledge findings are congruent with Bruton (2007), "The acquisition of the meaning of any new lexical items from reading is almost always initially only partial, though some cognates or highly enhanced items may accelerate the assimilation in L2 acquisition" (p. 164).

Furthermore, Bruton (2007) asserted, "To reflect these increments, the inclusion of partial knowledge in a test might be paramount" (p. 165). I slightly disagree with Bruton on this point as the findings of this study indicate that partial knowledge *should be* paramount. On the topic of incremental word learning, Zimmerman (2009) stated, "If we acknowledge that learning a word is incremental and that it happens one step at a time, we will approach it realistically as a process rather than a single act of memorization" (p. 2). Further, aspects of word knowledge "are understood gradually, over time, as a result of seeing a word and using it" (p. 2). Therefore, the focus on incremental acquisition in this study was supported by research.

The three options used with the ordered triple rank (OTR) instruction, that is, null, decontextualized (i.e., partially correct) and synonym, used in this test worked well. This was again in congruence with Bruton (2007) who stated, "For most practical research purposes, it is probably sufficient to have one partial category for quantification" (p. 173). In other words, offering three options: incorrect, partially correct and correct, as used in this

study. Similarly, in a meta-analysis of 48 multiple-choice test studies spanning 80 years of research, Rodriguez (2005) concluded that three options are optimal, "MC items should consist of three options, one correct option and two plausible distractors" (11). Similarly, Downing (2006) noted that "even in very well-developed tests it is rare that more than three options are statistically functional" (p. 292). In the Japanese university context, Shizuka, Takeuchi, Yashima and Yoshizawa (2006) concluded that there were several advantages to using three-options instead of four-options in addition to greater item writing efficiency and increased test reliability.

In a critique of multiple choice vocabulary tests using single best answer responses, Brown (1958) asserted, "A vocabulary test undertakes to discover whether an individual knows the meanings of a set of words. No test exhausts the disposition attached to any word. Vocabulary tests always deal with single appropriate responses" (p. 105). Brown (1958) furthered his critique of vocabulary test designs with "We credit the subject with more than the particular demonstrated ability. This is clear from the fact that we should be confounded to discover that someone who knew that *idiosyncrasy* meant *peculiarity* knew absolutely nothing else about either word" (p. 105). In a softening of his strong stance, Brown (1958) concedes that, "For certain practical purposes (determination of vocabulary size) it is enough to know that individuals who possess one appropriate response will generally also possess many others" (p. 106). Therefore, according to Brown, vocabulary size tests with one correct form-meaning match and a single best answer response instruction do have an important role to play in vocabulary testing.

Limitations

The Polysemous Academic Vocabulary Extent test was not perfectly targeted to this sample. In fact, it was a little easy for the sample because two of the 36 items were characterized as very easy and the distribution of the person ability measure locations was higher on the Wright Map than the distribution of the items difficulty calibration locations. On the other hand, there was the a priori test design plan that the test should not be too difficult as this test uses a new instruction and response style. Therefore, this limitation is a minor limitation.

The findings were limited in the fact that the Rating Scale Model (Andrich, 1978) has fixed or constant thresholds. Therefore, the changes in thresholds between individual items were not analyzed. Instead, a rating scale model applied to the thresholds for all items was analyzed. However, this omnibus analysis using the Rasch-Andrich rating scale model (Andrich, 1978) was very important as it was needed to assess the a priori proposed metasynonymy awareness hypothesis. Thus, this limitation was a minor limitation.

Suggestions for Future Research

Future research is recommended to expand the limits of vocabulary testing beyond traditional formats. In his investigation of vocabulary strength, Read (2004) stated there "is a good argument to be made that the breadth vs. depth metaphor has served its rhetorical purpose of encouraging researchers and language teachers to look beyond conventional test items" (p. 223). Furthermore, Read (2004) concluded once broadening our view "is widely accepted, it opens up a whole range of possibilities for other types of test item and

alternative ways of measuring aspects of word knowledge" (p. 224). Therefore, researchers and language teachers are encouraged to invent and experiment with new ways of testing and assessing that might be different from conventional standards. This fresh experimentation helps keep the vocabulary test research domain in a state of progressive development.

One emphasis of this study has been on word form to multiple meanings for an individual word rather than a solitary word to form meaning. Miller (1999) in his study, *On knowing a word*, asks the rhetorical question "What does a person who knows a word know? A ready answer is that a person who knows a word must know its meaning(s)" (p. 3). Miller's focus on multiple meanings is deliberate and important. Therefore, future studies into word form to multiple meanings are recommended.

Contextualization is also important, notes Miller (1999). The Polysemous Academic Vocabulary Extent test constructed for this study uses both contextualized (i.e., written representation) and decontextualized (i.e., cognitive representation) modalities. Miller (1999) asserts "with every distinct meaning of a polysemous word there must be some cognitive representation of the contexts in which that word form can be used to express that meaning" (p. 17). Therefore, it is important to develop the "ability to exploit context in order to determine meaning and resolve potential ambiguities" (Miller, 1999, p. 1). Thus, future studies into the relationship between contextualization and decontextualization modalities are recommended.

Vocabulary size and strength can be conceptualized as having no differentiation (Vermeer, 2001). In Vermeer's view, "there is no conceptual distinction between breadth

and depth of word knowledge, and thus there should be a strong relationship between breadth and depth measures" (p. 218). This is a suggested area of future research related to this study.

Another suggestion for future research is more researchers trying Rasch model measurement analyses when testing. Dolch and Leeds (1953) critiqued vocabulary testing because test designers "were just trying roughly to compare a child with other children of his own age with regard to the general field of vocabulary" (p. 188). This focus on demography is one of the points that Rasch (1960) critiqued about test psychologists "not dealing with the testing of individuals, but…were studying how traits, such as intelligence, were distributed in populations. They were making demographic studies and not studies of individuals" (p. xii). By using Rasch model analyses findings from testing or surveying situations, researchers and classroom teachers can give focused attention to the particular needs of individual students.

A suggestion for future research is adding a battery of receptive and productive tests covering the four basic language macro skills of reading, writing, listening and speaking. This could be helpful as the four different vocabulary test forms analyzed in this study could then be correlated with multiple criterion variables. This might lead to further insights and pathways for future research.

Some research in testing recommends that 3 options is the optimum number of options for multiple-choice tests because of the difficulty in constructing a fourth option that acts as an effective distractor. A suggestion for future research is researchers should investigate 3-option versions of the Vocabulary Size Test. Furthermore, a suggestion to

future researchers is to base those three options on the results of this present study. One of the benefits of the 3-option multiple-choice test item is that more items can be administered because it takes less time for test-takers to complete each item.

A final suggestion for future researchers who utilize the imagined sentence generating task employed in this study is initially to test how well non-native test-takers at various proficiency levels can complete this task. This might lead to insights into how the capacity to generate the imagined sentence has an influence on the final person ability measures dependent variable.

Final Conclusions

In a critique of Birnbaum's 3-Parameter Logistic (3-PL) model, Wright (1992) asserted "I never met an item that guessed, so I look for guessing among people" (p. 196). To paraphrase Wright into the context of this study, I never met a traditional multiple-choice item with incremental knowledge, so I look for incremental knowledge among people. Therefore, if researchers continue attempting to assess incremental knowledge by requiring only multiple-choice item unary full knowledge single best answer responses, they can never meet their purpose of measuring partial or incremental knowledge. Such test formats are poorly designed for these purposes.

To overcome this deficiency, Fisher's (1935) advice of designing tests to maximize data information is essential. As shown in this study, multiple-choice items can be designed with ranked response outcome spaces and incremental scoring rubrics. Incremental knowledge responses can be analyzed only when the extracted information from the dataset

is facilitated by the correct item format design. The item format design can potentially activate dormant distractors amongst the item options. Therefore, it is essential to spend adequate time and preparation in the designing phase before the items are administered.

Wright (1992) further stated, "So I do a better job of identifying and controlling guessing when I look for it in persons' responses, label it and decide what I want to do with it" (pp. 196-197). Paraphrasing Wright once again, I do a better job of identifying and controlling incremental knowledge when I look for it in persons' responses, label it and decide what I want to do with it. One way to achieve this is to require the respondents to incrementally rank all multiple item options as their response to each individual multiple choice item stem and options. Measuring gradations of knowledge is important for educational research development. Indeed, incremental vocabulary knowledge degrees are latent and have the potential to be measured. "Essential features of what we measure, however, are that it can be expressed as 'more' and 'less' and that there are infinitely many gradations of it," (Bezruczko & Linacre, 2005, p. 24). Therefore, measuring with a Rasch model along a continuous variable scale is essential.

Using the measures obtained from Rasch measurement facilitates making inferences that might be useful for future research pathways. Therefore, we can look back on the past from today and then also look ahead into the future. Wright and Mok (2004) explain that we are not concerned with individual occurrences of success or failure from yesterday. On the contrary, we want to be able to use that information to make inferences about tomorrow. "The ability to expect and so to infer is the impetus of human development, the prime tool

of civilization" (Wright & Mok, 2004, p. 6). Rasch model measurement enables researchers to face the unknown and dynamic future.

Test design development and construction, using an ordered triple rank (OTR) response requirement and related scoring rubric, is one answer to the conundrum of measuring incremental written receptive vocabulary knowledge. Researchers interested in measuring incremental vocabulary knowledge are encouraged to apply the amended multiple-choice item response space and scoring rubric documented in this study. However, a much better alternative is to make one's own original test and learn hands-on about the adventure of constructing a test through the practical and iterative test item development process.

REFERENCES

Aizawa, K., Ishikawa, S., & Murata, M. (Eds.). (2008). *"Daigaku eigo kyoiku gakkai kihongo list" ni motozuku: JACET 8000 Eitango* [The Japan Association of College English Teachers (JACET) Basic List: JACET 8000 English word list]. Tokyo, Japan: Kirihara.

Allenby, R. B. J. T., & Slomson, A. (2011). *How to count: An introduction to combinatorics*. Boca Raton, FL: Chapman & Hall/CRC.

Anderson, R. C., & Freebody, P. (1981). Vocabulary knowledge. In J. T. Guthrie (Ed.), *Comprehension and teaching: Research reviews* (pp. 77-117). Newark, DE: International Reading Association.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*(4), 561-573. doi:10.1007/BF02293814

Andrich, D. A. (1996). Category ordering and their utility. *Rasch Measurement Transactions, 9*(4), 403-415. Retrieved from http://www.rasch.org/rmt/rmt.94f.htm

Andrich, D. (2004a). Controversy and the Rasch model: A characteristic of incompatible paradigms? In E. V. Smith Jr., & R. M. Smith (Eds.) *Introduction to Rasch measurement* (pp. 143-166). Maple Grove, MN: JAM.

Andrich, D. (2004b). Understanding resistance to the data-model relationship in Rasch's paradigm: A reflection for the next generation. In E. V. Smith Jr., & R. M. Smith (Eds.) *Introduction to Rasch measurement* (pp. 167-200). Maple Grove, MN: JAM.

Andrich, D. (2010). Understanding the response structure and process in the polytomous Rasch model. In M. L. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models* (pp. 123-152). New York, NY: Routledge.

Andrich, D. (2013a). An expanded derivation of the threshold structure of the polytomous Rasch model that dispels any "threshold disorder controversy." *Educational and Psychological Measurement, 73*(1), 78-124. doi:10.1177/0013164412450877

Andrich, D. (2013b). The legacies of R. A. Fisher and K. Pearson in the application of the Polytomous Rasch Model for assessing the empirical ordering of categories. *Educational and Psychological Measurement, 73*(4), 553-580. doi:10.1177/0013164413477107

Aristotle. (2004). *Rhetoric* (W.R. Roberts, Trans.). Mineola, NY: Dover. (Reprinted from *The works of Aristotle* (Vol. 9), by W. D. Ross, Ed., 1924, London, England: Oxford University Press)

Baguley, T. (2012). Serious stats: A guide to advanced statistics for the behavioral sciences. Hampshire, England: Palgrave Macmillan.

Beglar, D. J. (2000). *The validation of a breadth of vocabulary test using structural equation modeling*. (Unpublished doctoral dissertation). Temple University, Philadelphia, PA.

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing, 27*(1), 101-118. doi:10.1177/0265532209340194

Beglar, D., & Hunt, A. (1999). Revising and validating the 2000 word level and university level vocabulary tests. *Language Testing, 16*(2), 131-162. doi:10.1177/026553229901600202

Bernhardson, C. S. (1967). Comparison of the three-decision and conventional multiple-choice tests. *Psychological Reports, 20*(3), 695-698. doi:10.2466/pr0.1967.20.3.695

Bezruczko, N. (2005). *Introduction.* In N. Bezruczko (Ed.) *Rasch measurement in health sciences* (pp. 1-7). Maple Grove, MN: JAM.

Bezruczko, N., & Linacre, J. M. (2005). Measurement theory foundations. In N. Bezruczko (Ed.), *Rasch measurement in health sciences* (pp. 8-34). Maple Grove, MN: JAM.

Boone, W. J., Staver, J. R., & Yale M. S. (2014). *Rasch analysis in the human sciences*. New York, NY: Springer.

Borowski, E. J., & Borwein, J. M. (2006). *Collins web-linked dictionary of mathematics.* New York, NY: Harper Collins.

Breuer, J. (2006). *Introduction to the theory of sets*. Mineola, NY: Dover.

Brown, J. D. (2001). *Using surveys in language programs*. Cambridge, England: Cambridge University Press.

Brown, J. D., & Rodgers, T. S. (2002). *Doing second language research*. Oxford, England: Oxford University Press.

Brown, R. (1958). *Words and things: An introduction to language*. New York, NY: The Free Press.

Bruton, A. (2007). Partial lexical knowledge in tests of incidental vocabulary learning from L2 reading. *The Canadian Modern Language Review, 64*(1), 163-180. doi:10.3138/cmlr.64.1.163

Coombs, C. H., Milholland, J. E., & Womer, F. B. (1956). The assessment of partial knowledge. *Educational and Psychological Measurement, 16*(1), 13-37. doi:10.1177/001316445601600102

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213-238. doi:10.2307/3587951

Cronbach L. J. (1942). An analysis of techniques for diagnostic vocabulary testing. *The Journal of Educational Research, 36*(3), 206-217. doi:10.1080/00220671.1942.10881160

Cronbach L. J. (1943). Measuring knowledge of precise word meaning. *The Journal of Educational Research, 36*(7), 528-534. doi:10.1080/00220671.1943.10881192

Daller, H., Milton, J., & Treffers-Daller, J. (2007). Editors' introduction: Conventions, terminology and an overview of the book. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.) *Modelling and assessing vocabulary knowledge* (pp. 1-32). Cambridge, England: Cambridge University Press.

Daulton, F. E. (2008). *Japan's built-in lexicon of English-based loanwords*. Clevedon, England: Multilingual Matters.

Davidson, F., & Lynch, B. K. (2002). *Testcraft: A teacher's guide to writing and using language test specifications*. New Haven, CT: Yale University Press.

Dolch, E. W., & Leeds, D. (1953). Vocabulary tests and depth of meaning. *The Journal of Educational Research, 47*(3), 181-189. doi:10.1080/00220671.1953.10882095

Downing, S. M. (2006). Selected-response item formats in test development. In S. M. Downing & T. M. Haladyna (2006). *Handbook of test development* (pp. 287-301). Mahwah, NJ: Erlbaum.

Eccles, P. J. (2007). *An introduction to mathematical reasoning: Numbers, sets and functions*. Cambridge, England: Cambridge University Press.

Elgort, I. (2012). Effects of L1 definitions and cognate status of test items on the Vocabulary Size Test. *Language Testing, 30*(2), 253-272. doi:10.1177/0265532212459028

Ellis, N. C. (2002). Frequency effects in language processing. *Studies in Second Language Acquisition, 24*(2), 143-188. doi:10.1017/S0272263102002024

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). London, England: Sage.

Fisher, R. A. (1935/2003). *The design of experiments.* Oxford, England: Oxford University Press.

Gakken. (2007). *Yourei de wakaru katakana shingo jiten: kaitei dai 2 han* [A dictionary of katakana words: Revised 2nd edition] (2nd ed.). Tokyo, Japan: Author.

Gass, S. M., & Mackey, A. (2002). Frequency effects and second language acquisition: A complex picture? *Studies in Second Language Acquisition, 24*(2), 249-260. doi:101017/S0272263102002097

Gatfield, T., & Larmar, S.A. (2006). Multiple choice testing methods: Is it a biased method of testing for Asian international students? *International Journal of Learning, 13*(1), 103-111. Retrieved from http://www98.griffith.edu.au/dspace/bitstream/handle/10072/13918/40461.pdf?sequence=1

Gershon, R. (1992a). Guessing and measurement. *Rasch Measurement Transactions, 6*(2), 209-210. Retrieved from http://www.rasch.org/rmt/rmt62a.htm

Gershon, R. (1992b, April). The effects of person-item mismatches on the integrity of the item characteristic curve. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA. Abstract retrieved from http://files.eric.ed.gov/fulltext/ED400284.pdf

Gershon, R. (1994, April). Analyzing multiple choice tests with the Rasch model: Improving item calibrations by deleting person-item mismatches. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA. Abstract retrieved from http://files.eric.ed.gov/fulltext/ED400282.pdf

Gries, S. T. (2008). Corpus-based methods in analyses of second language acquisition data. In P. Robinson & N. C. Ellis (2008). *Handbook of cognitive linguistics and second language acquisition* (pp. 406-431). New York, NY: Routledge.

Gyllstad, H. (2013). Looking at L2 vocabulary knowledge dimensions from an assessment perspective–Challenges and potential solutions. In C. Bardel, B. Laufer, & C. Linqvist (Eds.), *L2 vocabulary acquisition, knowledge and use: New perspectives on assessment and corpus analysis*. (pp. 11-22). Eurosla Monographs Series, 2. EUROSLA. Retrieved from http://www.eurosla.org/monographs/EM02/Gyllstad.pdf

Hudson, R. (1994). About 37% of word-tokens are nouns. Language, 70(2), 331-339. doi:10.2307/415831

IBM Corp. (2012). IBM SPSS Statistics for Windows (Version 21.0) [Computer Software]. Armonk, NY: IBM Corp.

Iramaneerat, C., Smith, E. V., & Smith, R. M. (2008). An introduction to Rasch measurement. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 50-70). Thousand Oaks, CA: Sage.

Ishii, T., & Schmitt, N. (2009). Developing an integrated diagnostic test of vocabulary size and depth. *RELC Journal, 40*(1), 5-22. doi:10.1177/0033688208101452

Jiyu. (2011). *Gendaiyogo no kiso chishiki katakana gairaigo/ryakugo jiten dai 4 han* [Contemporary glossary of basic knowledge of katakana loanwords/abbreviations dictionary, 4th edition] (4th ed.). Tokyo, Japan: Horiuchi, K. (Ed.) & Omori, R.

Johansson, S., & Hofland, K. (1989). *Frequency analysis of English vocabulary and grammar based on the LOB corpus, Vol. 1: Tag frequencies and word frequencies*. Oxford, England: Oxford University Press.

Jones, L. V., & Thissen, D. (2007). A history and overview of psychometrics. In C. R. Rao, & S. Sinharay (Eds.), *Handbook of statistics, Vol. 26: Psychometrics* (pp. 1-27). Amsterdam, The Netherlands: Elsevier.

Kuhn, T. S. (1996). *The structure of scientific revolutions* (3rd ed.). Chicago: The University of Chicago Press.

Laufer, B. (1992). How much lexis is necessary for reading comprehension? In P.J. Arnaud & H. Bejoint (Eds.). Vocabulary and applied linguistics (pp. 126-132). London, England: Macmillan. Retrieved from http://english.haifa.ac.il/staff/BL-Downloadable-Publications.pdf

Leech, G., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English: Based on the British National Corpus*. Harlow, England: Longman.

Linacre, J. M. (*n.d.*). Misfit diagnosis: infit outfit mean-square standardized [Online help for Winsteps Rasch measurement software]. Retrieved from http://www.winsteps.com/winman/diagnosingmisfit.htm

Linacre, J. M. (1989/1994). *Many-facet Rasch measurement* (2nd ed.). Chicago, IL: Mesa.

Linacre, J. M. (2000). Redundant items, overfit and measure bias. *Rasch Measurement Transactions*, 14, 755. Retrieved from http://www.rasch.org/rmt/rmt143a.htm

Linacre, J.M. (2002). What do outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, 16, 878. Retrieved from http://www.rasch.org/rmt/rmt162f.htm

Linacre, J. M. (2004). Optimizing rating scale category effectiveness. In E. V. Smith Jr. & R. M. Smith (Eds.) *Introduction to Rasch Measurement* (pp. 258-278). Maple Grove, MN: JAM.

Linacre, J. M. (2010). Transitional categories and usefully disordered thresholds. *Online Educational Research Journal.* Retrieved from http://www.oerj.org/View?action=viewPDF&paper=2

Linacre, J. M. (2012). Winsteps Rasch Tutorial 1 [Online help webpage]. Retrieved from http://www.winsteps.com/a/winsteps-tutorial-1.pdf

Linacre, J. M. (2013). Re: 46. Asymptotes [Online forum comment]. Retrieved from https://mmm1406.sanjose14-verio.com/rascho/forum2013b.htm

Linacre, J. M. (2015a). Winsteps® Rasch measurement computer program User's Guide. Beaverton, OR: Winsteps.com

Linacre, J.M. (2015b). Winsteps® (Version 3.90.0) [Computer software]. Beaverton, OR: www.winsteps.com

Lord, F. M., & Novick, M. R. (1968/2008). *Statistical theories of mental test scores*. Charlotte, NC: Information Age.

Lovell, G. D. (1941). Interrelations of vocabulary skills: Commonest versus multiple meanings. *The Journal of Educational Psychology*, *32*(1), 67-72. doi:10.1037/h0062751

Makinson, D. (2012). Sets, logic and maths for computing (2nd ed.). London, England: Springer.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2),149-174. doi:10.1007/BF02296272

McNamara, T. F. (1996). *Measuring second language performance*. Essex, England: Pearson.

Meara, P., & Buxton, B. (1987). An alternative to multiple choice vocabulary tests. *Language Testing, 4*(2), 142-154. doi:10.1177/026553228700400202

Messick, S. A. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13-103). New York, NY: American Council on Education/Macmillan.

Ministry of Education, Culture, Sports, Science and Technology (MEXT). (2009). *Chugakko gakushu shidou yoryou kaisetsu gaigokugo hen* [Junior high school study guide points explanation of foreign language compilation] (2nd ed.). Tokyo, Japan: Kairyudo.

Ministry of Education, Culture, Sports, Science and Technology (MEXT). (2010). *Kotogakko gakushu shidou yoryou kaisetsu gaigokugo hen: Eigo hen* [Senior high school study guide points explanation of foreign language compilation: English language compilation]. Tokyo, Japan: Kairyudo.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review*, *63*(2), 81-97. doi:10.1037/h0043158

Miller, G. A. (1999). On knowing a word. *Annual Review of Psychology, 50*(1), 1-19. doi:10.1146/annurev.psch.50.1.1

Miller, G. A. & Fellbaum, C. (1999). Semantic networks of English. *Cognition, 41*(1-3), 197-229. doi:10.1016/0010-0277(91)90036-4

Nation, I.S. P. (1983). Testing and teaching vocabulary. *Guidelines, 5*(1), 12-25. Retrieved from http://www.victoria.ac.nz/lals/about/staff/publications/paul-nation/1983-Testing-and-teaching.pdf

Nation, I.S.P. (1990). *Teaching and learning vocabulary*. Boston, MA: Heinle and Heinle.

Nation, I.S.P. (2012, October 23). The vocabulary size test [Online instructions, description, information and specifications]. Retrieved from http://www.victoria.ac.nz/lals/about/staff/publications/paul-nation/Vocabulary-Size-Test-information-and-specifications.pdf

Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher, 31*(7), 9-13. Retrieved from http://www.victoria.ac.nz/lals/about/staff/publications/paul-nation/2007-Beglar-TLT.pdf

Nedelsky, L. (1954). Ability to avoid gross error as a measure of achievement. *Educational and Psychological Measurement, 14*(3), 459-472. doi:10.1177/001316445401400303

Nelson, D. (Ed.). (2008). *The Penguin dictionary of mathematics* (4th ed.). London, England: Penguin.

O'Connor, J. (1940). *Unsolved business problems.* Boston, MA: Johnson O'Connor Research Foundation.

Palmer, H. E. (1917). *The scientific study and teaching of languages.* London, England: Harrap.

Poizner, S. B., Nicewander, W. A., & Gettys, C. F. (1978). Alternative response and scoring methods for multiple-choice items: An empirical study of probabilistic and ordinal response modes. *Applied Psychological Measurement, 2*(1), 83-96. doi:10.1177/014662167800200109

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research. (Expanded edition, 1980. Chicago, IL: University of Chicago Press.)

Rasch, G. (2006) On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. In D. J. Bartholomew (Ed.). *Measurement: Sage benchmarks in social research methods* (*Vol. II)* (pp. 3-41). London, England: Sage. (Reprinted from *Danish yearbook of philosophy* (Vol. 14) pp. 58-94, by G. Rasch, 1977, Copenhagen, Denmark: University of Copenhagen)

Read, J. (2000). *Assessing vocabulary*. Cambridge, England: Cambridge University Press.

Read, J. (2004). Plumbing the depths: How should the construct of vocabulary knowledge be defined? In P. Bogaards, & B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition, and testing* (pp. 209-227). Amsterdam, The Netherlands: Benjamins.

Read, J. (2013). Research timeline: Second language vocabulary assessment. *Language Teaching, 46*(1), 41-52. doi:10.1017/S0261444812000377

Remark Office OMR® (Optical Mark Recognition) (2010). (Version 8) [Computer software]. Malvern, PA: www.gravic.com

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice, 24*(2), 3-13. doi:10.1111/j.1745-3992.2005.00006.x

Ruhl, C. (1989). On monosemy: A study in linguistic semantics. Albany, NY: State University of New York Press.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. (*Psychometric Monograph No. 17*). Richmond, VA: Psychometric Society. Retrieved from https://www.psychometricsociety.org/sites/default/files/pdf/MN17.pdf

Sanseido. (2008). *Konsaisu katakanago jiten dai 3 pan* [Sanseido's concise dictionary of katakana words, 3rd edition] (3rd ed.). Tokyo, Japan: Author.

Sayood, K. (2012). *Introduction to data compression* (4th ed.). Waltham, MA: Elsevier.

Schmidt, C.K. (2009). Loanwords in Japanese. In M. Haspelmath & U. Tadmor (Eds.), Loanwords in the world's languages: A comparative handbook (pp. 545-574). Berlin, Germany: De Gruyter Mouton.

Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge, England: Cambridge University Press.

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behavior of two new versions of the Vocabulary Levels Test. *Language Testing, 18*(1), 55-88. doi:10.1177/026553220101800103

Seashore, R. H., & Eckerson, L. D. (1940). The measurement of individual differences in general English vocabularies. *The Journal of Educational Psychology, 31*(1), 14-38. doi:10.1037/h0053494

Sheskin, D. J. (2007). *Handbook of parametric and nonparametric statistical procedures* (4th ed.). Boca Raton, FL: Chapman & Hall/CRC.

Shizuka, T., Takeuchi, O., Yashima, T., & Yoshizawa, K. (2006). A comparison of three- and four-option English tests for university entrance selection purposes in Japan. *Language Testing, 23*(1), 35-57. doi:10.1191/0265532206lt319oa

Smith, R. M. (1987). Assessing partial knowledge in vocabulary. *Journal of Educational Measurement, 24*(3), 217-231. doi:10.1111/j.1745-3984.1987.tb00276.x

Stevens, S. S. (2006). On the theory of scales of measurement. In D. J. Bartholomew (Ed.) *Measurement: Sage benchmarks in social research methods* (Vol. 1.) (pp. 119-126). London, England: Sage. (Reprinted from *Science, 103*(2684), pp. 677-680, by S. S. Stevens, 1946)

Stone, M. H. (1998). Rating scale categories: Dichotomy, double dichotomy, and the number two. *Popular Measurement: Journal of the Institute for Objective Measurement, 1*(1), 61-65.

Stone, M. H., & Stenner, A. J. (2014). Comparison is key. *Journal of Applied Measurement, 15*(1), 26-39.

Tomoda, T. (2009). *The foreign word tsunami*. Saarbrucken, Germany: VDM Verlag.

Vermeer, A. (2001). Breadth and depth of vocabulary in relation to L1/L2 acquisition and frequency of input. *Applied Psycholinguistics, 22*(2), 217-234. doi:10.1017/S0142716401002041

Waller, N.G., & Reise, S.P. (2010). Measuring psychopathology with nonstandard Item Response Theory models: Fitting the four-parameter model to the Minnesota Multiphasic Personality Inventory. In S. Embretson (Ed.), *New directions in psychological measurement with model-based approaches* (pp. 147-173). Washington, DC: American Psychological Association.

Weisstein, E. W. (2003). *CRC concise encyclopedia of mathematics* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.

Willey, C. F. (1960). The three-decision multiple-choice test: A method of increasing the sensitivity of the multiple choice item. *Psychological Reports, 7*(3), 475-477. doi:10.2446/pr0.1960.7.3.475

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Erlbaum.

Wolfe, E. W., & Smith, E.V., Jr. (2007). Instrument development tools and activities for measure validation using Rasch models: Part 1–Instrument development tools. In E. V. Smith, Jr., & R. M. Smith (Eds.), *Rasch measurement: Advanced and specialized applications* (pp. 202-242). Maple Grove, MN: JAM.

Wright, B. D. (1992). IRT in the 1990s: Which models work best? *Rasch Measurement Transactions, 6,* 196-200.

Wright, B. D. (1997). A history of social science measurement. *Educational Measurement: Issues and Practice, 16,* 33-45. Retrieved from http://brainimaging.waisman.wisc.edu/~perlman/Kristin/wright-educ-meas-1997.pdf

Wright, B. D. (1988). Some comments about guessing. *Rasch Measurement Transactions, 1,* 9.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean square fit values. *Rasch Measurement Transactions,* 8, 370. Retrieved from http://www.rasch.org/rmt/rmt83b.htm

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago, IL: Mesa.

Wright, B. D., & Mok, M. M. C. (2004). An overview of the family of Rasch measurement models. In E. V. Smith Jr., & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 1-24). Maple Grove, MN: JAM.

Xue, G., & Nation, I. S. P. (1984). A university word list. *Language Learning and Communication, 3*(2), 215-229. Retrieved from http://www.victoria.ac.nz/lals/staff/publications/paul-nation/1984-Xue-UWL.pdf

Zimmerman, C. B. (2009). *Word knowledge: A vocabulary teacher's handbook*. Oxford, UK: Oxford University Press.

APPENDIX A

KEYWORDS BY JACET 8000 WORD FREQUENCY BAND AND KATAKANA DICTIONARY PAGE REPRESENTATION

	Item keyword	JACET 8000	Frequency band	Gakken (2007)	Sanseido (2008)	Jiyu (2011)
1.	link	1227	High	722	1188	760
2.	export(ed)	1760	High	83	148	94
3.	positive	1435	High	604	1016	644
4.	volume	1936	High	620	1039	656
5.	relax(ed)	1477	High	721	1187	759
6.	authority	1415	High	102	185	113
7.	license	4033	Medium	692	1144	728
8.	involve(d)	3258	Medium	68	117	75
9.	mature	3936	Medium	639	1065	672
10.	portion	3943	Medium	604	1017	---
11.	pose(d)	3382	Medium	605	1017	---
12.	depression	3515	Medium	391	677	413
13.	distortion	5189	Low	376	652	397
14.	monitor(ing)	5442	Low	678	1121	713
15.	sustainable	6210	Low	231	404	246
16.	deduction	6381	Low	377	669	398
17.	refine(d)	6365	Low	718	1182	756
18.	simulation	6268	Low	258	452	274
19.	consequence	2156	High	---	---	---
20.	affect(ed)	874	High	---	---	---
21.	significant	2003	High	---	---	---
22.	attitude	886	High	---	---	---
23.	survey(ed)	1349	High	---	---	---
24.	equipment	1568	High	---	---	---
25.	exploitation	4283	Medium	---	---	---
26.	commence	4060	Medium	---	---	---
27.	consistent	3320	Medium	---	---	---
28.	thesis	3769	Medium	---	---	---
29.	derive(d)	3135	Medium	---	---	---
30.	evaluation	3381	Medium	---	---	---
31.	inhibition	5178	Low	---	---	---
32.	invoke(d)	5051	Low	---	---	---
33.	incompatible	6378	Low	---	---	---
34.	deviation	5625	Low	---	---	---
35.	unifi(ed)	7165	Low	---	---	---
36.	contradict(ion)	6476	Low	---	---	---

APPENDIX B

ANSWERS TO ITEM OPTIONS

	Loan Keyword	0 Null	1 Decontextualized	2 Synonym
1.	link	b) fuzz (lint)	c) chain ring	a) connection
2.	exported	a) inspected (exposed)	b) introduced	c) sold overseas
3.	positive	c) environmental (positional)	a) sure	b) good
4.	volume	b) vent (hole)	a) loudness	c) amount
5.	relaxed	c) carried (relayed)	b) softened	a) rested
6.	authority	a) scribbler (author)	c) public corporation	b) expert
7.	license	b) file (line)	a) freedom	c) permit
8.	involved	a) clothed (enveloped)	c) complicated	b) included
9.	mature	c) inherited (maternal)	b) due	a) ripened
10.	portion	b) mixture (potion)	c) slice	a) part
11.	posed	b) owned (possessed)	a) modeled	c) presented
12.	depression	a) decry (depreciation)	b) dent	c) sadness
13.	distortion	c) amusement (distraction)	a) signal change	b) corruption
14.	monitoring	b) warning (motioning)	c) supervising	a) checking
15.	sustainable	a) questionable (suspicious)	c) confirmable	b) supportable
16.	deduction	c) action (deed)	b) conclusion	a) subtraction
17.	refined	c) supplied (refilled)	a) cultivated	b) polished
18.	simulation	a) announcement (simulcast)	b) pretense	c) representation

	Non-Loan Keyword	0 Null	1 Decontextualized	2 Synonym
19.	consequence	b) agreement (consensus)	a) importance	c) effect
20.	affected	c) associated (affiliated)	b) impressed	a) altered
21.	significant	b) autographic (signature)	c) meaningful	a) substantial
22.	attitude	c) dress (attire)	a) posture	b) outlook
23.	surveyed	a) endured (survived)	b) measured	c) assessed
24.	equipment	c) twins (equivalent)	a) apparatus	b) appliances
25.	exploitation	b) outburst (explosion)	c) victimization	a) development
26.	commence	c) praise (commend)	b) take a degree	a) embark on
27.	consistent	a) thoughtful (considerate)	c) persistent	b) congruent
28.	thesis	a) synonym book (thesaurus)	b) dissertation	c) premise
29.	derived	a) denied (declined)	c) inferred	b) obtained
30.	evaluation	b) elimination (exclusion)	c) ranking	a) appraisal
31.	inhibition	b) acquisition (inheritance)	a) ban	c) suppression
32.	invoked	c) concerned (involved)	b) conjured	a) solicited
33.	incompatible	a) inadequate (incomplete)	b) unsuitable	c) discordant
34.	deviation	c) adoration (devotion)	a) perversion	b) deflection
35.	unified	b) clothed (uniformed)	a) wedded	c) merged
36.	contradiction	a) compression (constriction)	c) difference	b) denial

APPENDIX C

POLYSEMOUS ACADEMIC VOCABULARY EXTENT TEST (PAVET)

Name: Student #:

MEANING MATCH INSTRUCTIONS

○ Please read these instructions carefully. Ask your teacher any questions if you are not sure.

○ In this question booklet there are written questions. You will also be given an answer mark sheet.

○ Usually, in multiple-choice tests only one answer is needed for each question.

○ But, this test uses a new way of answering. Three answers are needed for each question.

○ Each answer space has 3 rows (a, b, c) and 3 columns (0, 1, 2).

○ Use a pencil to fill in a bubble circle around a letter (a, b, c) to make a black dot ●.

○ For each question, make 3 dots ●●●. In other words:

- in each row, make one dot ● for a, one dot ● for b, and one dot ● for c.

- in each column, make one dot ● for 0, one dot ● for 1, and one dot ● for 2.

○ Find the meaning match (between the bolded **key word** and the option (a, b, c) in columns (0, 1, 2) where:

2 = this written sentence.

1 = another imagined sentence.

0 = no sentences.

○ When you have finished answering double-check your answers. For each answer, you must have:

● = ONE dot in each row (a, b, c), and

● = ONE dot in each column (0, 1, 2).

○ Try to clearly fill in the bubble circles. Do not write outside of the bubble circles.

○ Do not bend or fold the answer mark sheet.

○ It is alright if you want to write notes in the question booklet.

○ Please answer every question.

1. There is a strong **link** between reading and knowledge.
link: a) connection b) fuzz c) chain ring

2. Japan **exported** many cars last year.
exported: a) inspected b) introduced c) sold overseas

3. All the reviews of the new movie were **positive**.
positive: a) sure b) good c) environmental

4. The **volume** of food almost covered the entire plate.
volume: a) loudness b) vent c) amount

5. I **relaxed** by reading the newspaper after work.
relaxed: a) rested b) softened c) carried

6. She is an **authority** on the subject.
authority: a) scribbler b) expert c) public corporation

7. He has a **license** to drive a car.
license: a) freedom b) file c) permit

8. The manager always **involved** me in the final decisions.
involved: a) clothed b) included c) complicated

9. The bananas will be delicious when they are **mature**.
mature: a) ripened b) due c) inherited

10. The middle **portion** of the bridge fell down.
portion: a) part b) mixture c) slice

11. Their attitude **posed** a threat to our hope for peace.
posed: a) modeled b) owned c) presented

12. She was in a **depression** for 3 months.
depression: a) decry b) dent c) sadness

13. **Distortion** of the facts is typical of politicians.
distortion: a) signal change b) corruption c) amusement

14. The police are **monitoring** their phone calls.
monitoring: a) checking b) warning c) supervising

15. Solar powered energy is a **sustainable** resource.
sustainable: a) questionable b) supportable c) confirmable

16. They make a **deduction** from his monthly income.
deduction: a) subtraction b) conclusion c) action

17. When brown rice is **refined** it becomes white.
refined: a) cultivated b) polished c) supplied

18. A **simulation** is used in the classroom to develop communication skills.
simulation: a) announcement b) pretense c) representation

19. His decision had an unintended **consequence**.
consequence: a) importance b) agreement c) effect

20. The injury from the car accident **affected** his running ability.
affected: a) altered b) impressed c) associated

21. Our section has had a **significant** decrease in complaints.
significant: a) substantial b) autographic c) meaningful

22. They had the **attitude** that life was meant to be enjoyed.
attitude: a) posture b) outlook c) dress

23. She **surveyed** her new colleagues carefully from across the room.
surveyed: a) endured b) measured c) assessed

24. They bought some new kitchen **equipment**.
equipment: a) apparatus b) appliances c) twins

25. The **exploitation** of iron ore deposits has been very profitable.
exploitation: a) development b) outburst c) victimization

26. She will **commence** her next adventure in June.
commence: a) embark on b) take a degree c) praise

27. The statement from the witness was **consistent** with the other evidence.
consistent: a) thoughtful b) congruent c) persistent

28. His **thesis** is that global warming has been exaggerated by the media.
thesis: a) synonym book b) dissertation c) premise

29. We **derived** a new medicine from the jungle plants.
derived: a) denied b) obtained c) inferred

30. The **evaluation** indicated the antique was worth 50 million yen.
evaluation: a) appraisal b) elimination c) ranking

31. Her **inhibition** of sneezing in public was a sign of her strict upbringing.
inhibition: a) ban b) acquisition c) suppression

32. They **invoked** support for their rights.
invoked: a) solicited b) conjured c) concerned

33. John's worldview is **incompatible** with Mary's.
incompatible: a) inadequate b) unsuitable c) discordant

34. The magnet caused **deviation** of the compass needle.
deviation: a) perversion b) deflection c) adoration

35. Germany **unified** in the late twentieth century.
unified: a) wedded b) clothed c) merged

36. The defendant's **contradiction** of the accusations was not surprising.
contradiction: a) compression b) denial c) difference

APPENDIX D

OPTICAL MARK RECOGNITION (OMR) ANSWER SHEET

Name: TOEIC-IP Score: i.) Listening:

Student #: ii.) Reading:

Gender: Woman / Man Age: …… Years iii.) Total:

	❶	❷	❸		❶	❷	❸		❶	❷	❸
1.	a	a	a	6.	a	a	a	11.	a	a	a
	b	b	b		b	b	b		b	b	b
	c	c	c		c	c	c		c	c	c
2.	a	a	a	7.	a	a	a	12.	a	a	a
	b	b	b		b	b	b		b	b	b
	c	c	c		c	c	c		c	c	c
3.	a	a	a	8.	a	a	a	13.	a	a	a
	b	b	b		b	b	b		b	b	b
	c	c	c		c	c	c		c	c	c
4.	a	a	a	9.	a	a	a	14.	a	a	a
	b	b	b		b	b	b		b	b	b
	c	c	c		c	c	c		c	c	c
5.	a	a	a	10.	a	a	a	15.	a	a	a
	b	b	b		b	b	b		b	b	b
	c	c	c		c	c	c		c	c	c

Office use only

0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9

	❶	❷		❶	❷		❶	❷
16.	ⓐ ⓑ ⓒ	ⓐ ⓑ ⓒ	ⓐ ⓑ ⓒ	23.	ⓐ ⓑ ⓒ	ⓐ ⓑ ⓒ	ⓐ ⓑ ⓒ	30.	ⓐ ⓑ ⓒ	ⓐ ⓑ ⓒ	ⓐ ⓑ ⓒ
17.	ⓐ ⓑ ⓒ	ⓐ ⓑ ⓒ	ⓐ ⓑ ⓒ	24.	ⓐ ⓑ ⓒ	ⓐ ⓑ ⓒ	ⓐ ⓑ ⓒ	31.	ⓐ ⓑ ⓒ	ⓐ ⓑ ⓒ	ⓐ ⓑ ⓒ
18.	ⓐ ⓑ ⓒ	ⓐ ⓑ ⓒ	ⓐ ⓑ ⓒ	25.	ⓐ ⓑ ⓒ	ⓐ ⓑ ⓒ	ⓐ ⓑ ⓒ	32.	ⓐ ⓑ ⓒ	ⓐ ⓑ ⓒ	ⓐ ⓑ ⓒ
19.	ⓐ ⓑ ⓒ	ⓐ ⓑ ⓒ	ⓐ ⓑ ⓒ	26.	ⓐ ⓑ ⓒ	ⓐ ⓑ ⓒ	ⓐ ⓑ ⓒ	33.	ⓐ ⓑ ⓒ	ⓐ ⓑ ⓒ	ⓐ ⓑ ⓒ
20.	ⓐ ⓑ ⓒ	ⓐ ⓑ ⓒ	ⓐ ⓑ ⓒ	27.	ⓐ ⓑ ⓒ	ⓐ ⓑ ⓒ	ⓐ ⓑ ⓒ	34.	ⓐ ⓑ ⓒ	ⓐ ⓑ ⓒ	ⓐ ⓑ ⓒ
21.	ⓐ ⓑ ⓒ	ⓐ ⓑ ⓒ	ⓐ ⓑ ⓒ	28.	ⓐ ⓑ ⓒ	ⓐ ⓑ ⓒ	ⓐ ⓑ ⓒ	35.	ⓐ ⓑ ⓒ	ⓐ ⓑ ⓒ	ⓐ ⓑ ⓒ
22.	ⓐ ⓑ ⓒ	ⓐ ⓑ ⓒ	ⓐ ⓑ ⓒ	29.	ⓐ ⓑ ⓒ	ⓐ ⓑ ⓒ	ⓐ ⓑ ⓒ	36.	ⓐ ⓑ ⓒ	ⓐ ⓑ ⓒ	ⓐ ⓑ ⓒ

APPENDIX E

LOANWORD RECOGNITION YES/NO CHECKLIST SURVEY

AND ANSWER SHEET

Name: Student #:

1. link
2. consequence
3. affect
4. export
5. significant
6. positive
7. volume
8. attitude
9. relax
10. authority
11. survey
12. equipment
13. license
14. exploitation
15. commence
16. involve
17. mature
18. portion
19. consistent
20. thesis
21. pose
22. depression
23. derive
24. evaluation
25. distortion
26. inhibition
27. monitor
28. invoke
29. sustainable
30. incompatible
31. deduction
32. deviation
33. refine
34. unify
35. simulation
36. contradiction

Name: Student #:

INSTRUCTIONS:

Please mark Ⓨ or Ⓝ for each word.

Ⓨ = **Yes**, this word is a katakana loanword, or, Ⓝ = **No**, this word is not a katakana loanword.

1.	Ⓨ	Ⓝ	13.	Ⓨ	Ⓝ	25.	Ⓨ	Ⓝ
2.	Ⓨ	Ⓝ	14.	Ⓨ	Ⓝ	26.	Ⓨ	Ⓝ
3.	Ⓨ	Ⓝ	15.	Ⓨ	Ⓝ	27.	Ⓨ	Ⓝ
4.	Ⓨ	Ⓝ	16.	Ⓨ	Ⓝ	28.	Ⓨ	Ⓝ
5.	Ⓨ	Ⓝ	17.	Ⓨ	Ⓝ	29.	Ⓨ	Ⓝ
6.	Ⓨ	Ⓝ	18.	Ⓨ	Ⓝ	30.	Ⓨ	Ⓝ
7.	Ⓨ	Ⓝ	19.	Ⓨ	Ⓝ	31.	Ⓨ	Ⓝ
8.	Ⓨ	Ⓝ	20.	Ⓨ	Ⓝ	32.	Ⓨ	Ⓝ
9.	Ⓨ	Ⓝ	21.	Ⓨ	Ⓝ	33.	Ⓨ	Ⓝ
10.	Ⓨ	Ⓝ	22.	Ⓨ	Ⓝ	34.	Ⓨ	Ⓝ
11.	Ⓨ	Ⓝ	23.	Ⓨ	Ⓝ	35.	Ⓨ	Ⓝ
12.	Ⓨ	Ⓝ	24.	Ⓨ	Ⓝ	36.	Ⓨ	Ⓝ

Office use only

⓪	①	②	③	④	⑤	⑥	⑦	⑧	⑨
⓪	①	②	③	④	⑤	⑥	⑦	⑧	⑨
⓪	①	②	③	④	⑤	⑥	⑦	⑧	⑨

APPENDIX F

WRITTEN PERMISSION REQUESTED VIA EMAIL FOR ADAPTION OF A COPYRIGHTED TABLE INTO TABLE 2 (PAGE 40)

Sent**:** Sun, 18 Oct 2015
Subject**:** Re: A request for adapting a table permission

Dear Richard,

I would like to request written copyright permission (for reuse in print and electronic form) to adapt a table from a book published by you (JAM Press). I would like to adapt a small part of the table and add some of my own points for making a simpler table for use in my doctoral dissertation.

The table is from the book, Bezruczko, N. (2005). "Rasch measurement in health sciences". Part One - Introduction, Table 1: Prominent Differences Among Rasch Models, Item Response Theory, and Classical Theory" on page 3, written by Nikolaus Bezruczko.

With your permission, I would be able to provide the source material in a table note. This would be in line with APA publication manual (6th edition, 2010) guidelines.

Looking forward to hearing from you.

Best regards,

Phil.

APPENDIX G

WRITTEN PERMISSION GRANTED VIA EMAIL FOR ADAPTION OF A COPYRIGHTED TABLE INTO TABLE 2 (PAGE 40)

Sent: Thursday, 22 October 2015
Subject**:** Re: A request for adapting a table permission

Dear Phil,

You have my permission to adapt the table that appears on page 3 of the Rasch Measurement in Health Sciences book published in 2005 for your dissertation, with the appropriate attribution.

Good luck with your work. If you decide to publish any work from your dissertation, I hope that you will remember JAM as one of your publishing outlets.

Best Regards,
Richard

Richard M. Smith, Editor
Journal of Applied Measurement
P.O. Box 1283
Maple Grove, MN 55311, USA