

Faculty of Cognitive Sciences and Human Development

Design, Validation, and Analysis of Perceived Sign Usability Scale Vis-à-vis Cognitive Sign Characteristics

Donald Anak Stephen

Doctor of Philosophy
2018

Design, Validation, and Analysis of Perceived Sign Usability Scale Vis-à-vis Cognitive Sign Characteristics

Donald Anak Stephen

A thesis submitted

In fulfilment of the requirements for the degree of Doctor of Philosophy

(Cognitive Sciences)

Faculty of Cognitive Sciences and Human Development
UNIVERSITI MALAYSIA SARAWAK
2018

DECLARATION

I, Donald Anak Stephen (12010017), Faculty of Cognitive Sciences and Human Development hereby declare that the work entitled "Design, Validation, and Analysis of Perceived Sign Usability Scale Vis-à-vis Cognitive Sign Characteristics" is my original work. I have not copied from any other students' work or from any other sources where due references or acknowledgement is made explicitly in the text, nor has any part been written for me by another person. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature of Student:

Name:

Date:

ACKNOWLEDGEMENT

The only thing that we can be certain about research is the fact that it will never be certain. This is something that we eventually learn to realize but at times refuse to believe. I have grown and learned a lot in this journey. It was at times brutal, but it was a steep learning curve that I will always treasure. Although I have learned a lot, I realized there are a lot more that I do not know. This thesis was conceived with the help of an incredible support system, good coffees, and Elvis Presley's records. Therefore, writing this acknowledgement is an honour because it is an opportunity to formally express my gratitude. Of course, after working on this for many years, it would not be possible to list everyone who had helped. The Ministry of Higher Education has been incredibly generous by providing me with MyBrain15 scholarship. I am grateful to UNIMAS, especially the staffs (academic and non-academic), colleagues and fellow students in the Faculty of Cognitive Sciences and Human Development (FCSHD) as well as the Institute of Borneo Studies who have provided invaluable resources and support throughout my study.

I owe my development as a researcher and the completion of this thesis to my dear supervisor, Assoc Prof. Dr. Shahren Ahmad Zaidi Adruce who has been instrumental in this journey. Not only he has given me invaluable opportunities, inputs, and guidance, he also believed in me when I was struggling with self-doubt. For that, I am touched and profoundly grateful. I couldn't have climbed all the way up here with my puny knowledge and strength alone. Growing up with serious stuttering, my schoolteachers tirelessly trained my speech every morning before the first period. It was from that experience I aspire to be in the world of academia and to inspire others through education. This is one of the reasons why I wrote

this thesis (be sure to read problem statements for more reasons I wrote this thesis. Those are legit reasons as well).

I would also like to dedicate special acknowledgement to Dr. Philip Nuli Anding who was the Deputy Dean for Postgraduate and Research, who have worked tirelessly to help graduate students fulfill their potential. Not forgetting my co-supervisor, Dr. Ng Giap Weng, my proposal defense panels, item reviewers, friends as well as examiners who have all played an important role in my development as a researcher. I would like to dedicate this thesis to my family who have been unconditionally supportive throughout the entire duration of my study. I am sure that skills, experiences, and information I gained from this wonderful support system will be invaluable in future endeavors.

ABSTRACT

This study was galvanized by the lack of instruments to assess users' perception on sign usability. This study introduces Perceived Sign Usability Scale (PSUS), which is an instrument to assess sign usability pertaining to its cognitive sign characteristics (CSC). This study delineates the process of designing and developing the PSUS. The design of the instrument involves the development of items as well as a formula to calculate the composite score of PSUS. PSUS utilizes an open-ended sign comprehension test and Likert-type items to assess sign usability. The combination of both methods results in a more intuitive and valid instrument. This study adapts statistical methods such as the c-coefficient, Multiple Responses Analysis, Cochran's Q with Pairwise McNemar as post-hoc tests, as well as Rasch Analysis. This study has the potential to spearhead the emergence of sign usability studies both locally and internationally because of the identification of CSC, the introduction of PSUS, design framework/technique, as well as the derivation of formula to calculate the composite score of Likert-type items. Although PSUS is proved to be valid and reliable, it is still subjected to inevitable limitations. PSUS may not cover the complete extent of sign usability because it is an indubitable fact that human cognition and sign usability are complex. There could be more CSC emerged in the future to be integrated into the new version of PSUS. However, this limitation is not undesirable as it also serves as a catalyst that allows for continuous improvement of the instrument.

Keywords: Cognitive sign characteristics, Perceived Sign Usability Scale, instrument development.

Reka Bentuk, Pensahihan, dan Penganalisaan Instrumen Kebolehgunaan Papan Tanda berlandaskan Ciri-ciri Kognitif

ABSTRAK

Kajian ini didorong oleh masalah ketandusan instrumen untuk mengkaji persepsi pengguna jalan raya terhadap tahap kebolehgunaan papan tanda. Kajian ini menampilkan instrumen Skala Kebolehgunaan Papan Tanda untuk mengkaji kebolehgunaan papan tanda berdasarkan ciri-ciri kognitif. Kajian ini menghurai proses pengolahan instrumen yang merangkumi perolehan rumus untuk mengira skor gabungan. Instrumen ini menggunakan soalan kefahaman berbentuk terbuka dan juga item-item berskala Likert untuk mengukur tahap kebolehgunaan. Implikasi daripada pembauran kedua-dua kaedah ini menyuguhkan keputusan analisa yang sahih dan mudah difahami. Proses pengolahan instrumen Skala Persepsi Kebolehgunaan Papan Tanda ini menggunakan pelbagai kaedah statistik termasuklah pekali c, analisis respon majmuk, ujian Q Cochran, ujian McNemar dan juga model pengukuran Rasch. Kajian ini memangkin kajian-kajian lanjutan hasil daripada cetusan idea dengan adanya skala kebolehgunaan, kerangka kerja reka bentuk, dan juga formula skor gabungan. Meskipun kesahihan telah dapat dipastikan, instrumen ini tidak lekang daripada kekangan ekoran kerumitan proses kognitif dan juga kebolehgunaan papan tanda yang tidak dapat disangkal. Kajian ini tidak menafikan kemungkinan wujudnya ciri-ciri kognitif lain yang seharusnya diasimilasikan ke dalam instrumen. Namun demikian, perihal ini bukanlah sesuatu yang tidak diingini kerana kelansungan proses penambahbaikan instrumen dapat dilaksanakan.

Kata kunci: *Ciri-ciri kognitif, Skala Persepsi Kebolehgunaan Papan Tanda, penghasilan instrumen*

TABLE OF CONTENTS

Page

DECLARATION i

ACKNOWLEDGEMENT ii

ABSTRACT iv

ABSTRAK v

TABLE OF CONTENTS vi

LIST OF TABLES xiii

LIST OF FIGURES xvii

LIST OF ABBREVIATIONS xx

CHAPTER 1: INTRODUCTION 1

1.1 Overview 1

1.2 Introduction 1

1.2.1 Background of study 2

1.2.2 Road safety climate and initiatives in Malaysia 4

1.3 Problem Statement 7

1.3.1 Poor traffic sign comprehension 7

1.3.2 Lack of sign usability studies and usability instrument 9

1.3.3 Difference on usability of sign and other products 10

1.3.4 Issues related to existing sign usability methods 11

1.3.5 Controversial composite score calculation methods 13

1.3.6 Summary of problems/issues 16

1.4 Objectives of the Study 17

1.4.1 General objective 17

1.4.2 Specific objectives 17

1.5 Significance of Study 20

1.6 Assumptions 22

1.7 Definition of Terms 22

1.7.1 Traffic signs 22

1.7.2 Usability 23

1.7.3 Perception 24

1.7.4 Cognitive Sign Characteristics (CSC) 24

1.8 Chapter Organization 25

CHAPTER 2: LITERATURE REVIEW 26

2.1 Overview 26

2.2 Part One: Traffic Sign 27

2.2.1 About traffic signs 27

2.2.2 Human perception in relation to traffic signs 31

2.2.3 Traffic signs usability component 35

2.3 Part Two: Existing Sign Usability Researches 37

2.3.1 Existing methods for assessing sign usability 38
2.3.2 Locality 39
2.3.3 Characteristics of participants in sign usability researches 40
2.4 Part Three: Cognitive Sign Characteristics Constituting Sign Usability 45
2.4.1 Definition of cognitive sign characteristics (CSC) 46
2.4.2 Identifying CSC constituting sign usability 58
2.4.3 CSC code occurrence 60
2.4.4 CSC code co-occurrence 62
2.5 Part Four: Usability Instruments 64
2.5.1 Analysis of generic usability instruments 64
2.5.2 System Usability Scale (SUS) 68
2.5.3 Composite score of SUS 70
2.6 Chapter Summary 71
CHAPTER 3: METHODOLOGY 72
3.1 Overview 72
3.2 Research Approach 74
3.3 Instrument Development Flow 74
3.4 Literature Analysis Methods 76
3.4.1 Collection of sources 78
3.4.2 Coding 82

3.5 Item Development 83
3.5.1 Redundancy analysis 84
3.5.2 Relevancy analysis 85
3.5.3 Readability Tests 86
3.5.4 Multiple Responses Analysis (MRA) 87
3.6 Justifications for the Nature Participants and Sample Size 90
3.6.1 Determining the nature of participants 90
3.6.2 Sample size requirement for Cochran's Q test 94
3.6.3 Sample size requirement for Rasch 95
3.7 Procedure and Instrumentation 96
3.7.1 Item feedback and translation of instrument 96
3.7.2 Instrument validation 98
3.8 The Rasch Model 102
3.8.1 Rasch fit statistics 106
3.8.2 Assessment of reliability and separation index 108
3.8.3 Unidimensionality and PCAR 111
3.8.4 Differential Item Functioning 113
3.8.5 Rasch item-person mapping 114
3.9 Development of Composite Score Formula for PSUS 114
3.9.1 Concept and mechanism 114

3.9.2 Development 116

3.9.3 Formula testing and analysis 120

3.10 Suggestion for Interpretation of Results 123

3.10.1 Interpretation for open-ended sign comprehension test 123

3.10.2 Interpretation for Likert-type items 125

3.10.3 Classification of sign 128

3.11 Chapter Summary 131

CHAPTER 4: RESULTS 133

4.1 Overview 133

4.2 Demographic Information 134

4.3 Item Development 136

4.3.1 Item feedback 137

4.3.2 Item rationales and justification 142

4.3.3 Readability tests 146

4.4 Multiple Responses Analysis (MRA) 150

4.4.1 Item categorization based on CSC 150

4.4.2 Analysis of item – Cochran's Q and pairwise McNemar 156

4.4.3 Summary for MRA, Cochran's Q and McNemar tests 162

4.5 Rasch Analysis 163

4.5.1 Test level statistics for all signs 164

4.5.2 Rasch Analysis of sign A (RP.06 p.78) 165

4.5.3 Rasch Analysis of sign B (RP.10 p.79) 172

4.5.4 Rasch Analysis of sign C (RP.05 p.78) 178

4.5.5 PCAR analysis and paired t-test of item subsets 184

4.5.6 Differential Item Functioning 185

4.5.7 Summary and decision on item removal 193

4.6 Assessment of Test Signs 201

4.6.1 Cognitive mapping 201

4.6.2 Classification of sign 205

4.7 Discussion 208

4.7.1 Major findings 208

4.7.2 Sign usability vs. usability of other systems 209

4.7.3 Cognitive sign characteristics (CSC) 211

4.7.4 Interpretation of scale 213

4.7.5 Individual characteristics in sign comprehension 214

4.8 Chapter Summary 216

CHAPTER 5: SUMMARY AND CONTRIBUTION 218

5.1 Overview 218

5.2 Summary of the Findings 218

5.3 Contribution and Implication of Study 222

5.3.1 A new dimension of sign usability research 222

5.3.2 Methods/framework to develop instrument 223

5.3.3 PSUS formula to calculate composite score of Liker-type items 224

5.3.4 Methods/technique to assess sign usability 226

5.3.5 Practice: Instrument to assess sign usability 227

5.4 Limitations and Delimitations 228

5.5 Suggestions for Future Research 230

5.5.1 Revisiting and expanding theories/concept 231

5.5.2 Adaptation of instrument and interpretation methods 232

5.5.3 Intervention 233

5.6 Concluding Remarks 234

REFERENCES 237

APPENDICES 270

LIST OF TABLES

Page

Table 2.1 Overview of Chapter Two 26

Table 2.2 Colour Combinations for Signs of Different Purpose 30

Table 2.3 Aspects of Sign Usability 36

Table 2.4 Participants in Sign Usability Research 41

Table 2.5 Individual Characteristics Affecting Sign Usability 42

Table 2.6 Individual Characteristics Not Affecting Sign Usability 43

Table 2.7 CSC and Article Sources 59

Table 2.8 Distribution of CSC based on Year of Publication in 2 Decades 60

Table 2.9 Relative Code Occurrence Frequency of CSC 61

Table 2.10 C-coefficient Table of CSC 63

Table 2.11 Redundant Concepts and Number of Item 65

Table 2.12 Co-occurrence of Concepts 65

Table 2.13 Item Pool Redundancy Analysis Summary 66

Table 2.14 Most used Qualifying, Subject and Object Words 66

Table 3.1 Overview of Chapter Three 72

Table 3.2 Source of Data and Methods for Objective 1 73

Table 3.3 Source of Data and Methods for Objective 2-4 74

Table 3.4 Calculation on Data Table 89

Table 3.5 Value for 100/(sh-sl)i based on No. of Items and Scale Size 117

Table 3.6 Operation on Item Responses 118

Table 3.7 Independent T-test Results Comparison for All Composite Scores 121

Table 3.8 Mann-Whitney U Test Comparison for All Composite Scores 121

Table 3.9 Spearman and Pearson Comparison for all Composite Scores 122
Table 3.10 PSUS Classification of Sign 129
Table 3.11 Sign Usability Class and Description (Ascending) 130
Table 4.1 Overview of Chapter Four 133
Table 4.2 Objectives and Results 134
Table 4.3 Demographics of Participants 135
Table 4.4 Modification of SUS 138
Table 4.5 Researcher's Notes and Action on Item Feedback (version 1) 139
Table 4.6 Researcher's Notes and Action on Item Feedback (version 2) 141
Table 4.7 The Perceived Sign Usability Scale (Version 3) 147
Table 4.8 Test Statistics for 20 Items (Version 3) 148
Table 4.9 Readability Grade Levels and Readability Scores (Version 3) 149
Table 4.10 Correlation of Readability Tests 150
Table 4.11 List of Items and CSC 151
Table 4.12 MRA of Item 1 to 4 152
Table 4.13 MRA of Item 5 to 8 152
Table 4.14 MRA of Item 9 to 12 153
Table 4.15 MRA of Item 13 to 16 153
Table 4.16 MRA of Item 17 to 20 154
Table 4.17 Responses on Each Item for MRA 155
Table 4.18 Complexity, Priming, and Learnability for Item 2 157
Table 4.19 Complexity and Memorability for Item 4 157
Table 4.20 McNemar Test for Item 4 158
Table 4.21 Complexity and Standardization for Item 6 158

Table 4.22 McNemar Test for Item 6 159
Table 4.23 Meaningfulness, Comprehension, and Learnability for Item 9 159
Table 4.24 McNemar Test for Item 9 160
Table 4.25 Legibility, Surface Properties and Meaningfulness of Item 19 161
Table 4.26 McNemar Test for Item 19 161
Table 4.27 Legibility, Surface Properties, and Complexity of Item 20 162
Table 4.28 CSC assigned to each item 163
Table 4.29 Test Level Statistics for All Signs 165
Table 4.30 Item Analysis Summary on Text + Symbol Sign A (RP.06) 165
Table 4.31 Reliability Analysis Text + Symbol Sign (RP.06) 167
Table 4.32 Rasch Reliability Statistics Text + Symbol Sign (RP.06) 167
Table 4.33 Score, Ability Estimate, and Standard Error (RP.06) 169
Table 4.34 WMS and UMS Fit Statistics for Text + Symbol Sign (RP.06) 171
Table 4.35 Item Difficulty and Std. Fit Statistics for Text + Symbol Sign (RP.06) 172
Table 4.36 Item Analysis Summary on Symbol-only Sign (RP.10) 173
Table 4.37 Reliability Analysis on Symbol-only Sign (RP.10) 173
Table 4.38 Rasch Reliability Statistics for Symbol-only Sign (RP.10) 174
Table 4.39 Scores, Ability Estimate, and Standard Error (RP.10) 175
Table 4.40 WMS and UMS fit Statistics for Symbol-only Sign (RP.10) 177
Table 4.41 Item Difficulty and Std. Fit Statistics for Symbol-only Sign (RP.10) 177
Table 4.42 Item Analysis Summary on Text-only Sign (RP.05) 178
Table 4.43 Reliability analysis on Text-only Sign (RP.05) 179
Table 4.44 Rasch Reliability Statistics for Text-only Sign (RP.05) 179
Table 4.45 Score, Ability Estimate, and Standard Error (RP.05) 181

Table 4.46 WMS and UMS fit Statistics for Text-only Sign (RP.05) 183
Table 4.47 Item Difficulty and Std. Fit Statistics for Text-only Sign (RP.05) 183
Table 4.48 PCAR Analysis and Paired T-test of Item Subsets 184
Table 4.49 DIF Analysis on Gender for All Signs 186
Table 4.50 DIF Analysis on License Status for All Signs 187
Table 4.51 DIF Classification Based on Gender and License Status 191
Table 4.52 Independent T-Test for PSUS Score (Gender and License Status) 193
Table 4.53 Screening of Problematic Items 200
Table 4.54 Comprehension Test Result for Sign A RP.06 201
Table 4.55 Comprehension Test Result for Sign B RP.10 202
Table 4.56 Comprehension Test Result for Sign C RP.05 202
Table 4.57 Cross-tabulation and Classification of All Signs 207
Table 4.58 PSUS Classification of Test Sign 207
Table 4.59 Final version of PSUS 208
Table 5.1 Summary of Contributions 227
Table 5.2 Summary- Differences in Usability Aspects 234
Table 5.3 Summary - Poor Sign Comprehension & Lack of Sign Usability Studies ... 235
Table 5.4 Summary – Problems in Composite Score Calculations 236

LIST OF FIGURES

Page

Figure 1.1 Road accidents in Sarawak 2003-2017 5

Figure 1.2 Problems leading to research objectives 19

Figure 1.3 PSUS design and development framework 20

Figure 2.1 Shapes of sign and their use 29

Figure 2.2 Different colours of prohibitive signs 31

Figure 2.3 Human cognition and understanding of sign (Clarkson et al., 2013) 32

Figure 2.4 Semiotic approach of understanding sign 34

Figure 2.5 Gestalt principles of sign perception 35

Figure 2.6 Methods to assess sign usability 39

Figure 2.7 The locality of sign usability research 40

Figure 2.8 Example of sign that have low and high guessability score. 48

Figure 2.9 Number of code occurrence for CSC 62

Figure 2.10 Frequency chart of qualifying words and subject words 67

Figure 3.1 Development of instrument (Brancato et al., 2006; Radhakrishna, 2007) 75

Figure 3.2 Documentary search flow, adapted from Ridley (2008) and Page (2008) 79

Figure 3.3 Participants in past sign usability studies 90

Figure 3.4 Number of road accident deaths by age group 92

Figure 3.5 Distribution of sample size in sign usability studies 93

Figure 3.6 Test signs (from left: RP.06, RP.10, and RP.05) 101

Figure 3.7 Mean SUS score for adjective ratings (Bangor et al., 2009, 2008) 126

Figure 3.8 Grade scale and acceptability range (Bangor et al., 2009, 2008) 127

Figure 3.9 Dichotomous range of PSUS 128

Figure 4.1 Item development from version 0 to the final version 136
Figure 4.2 Number of categorization for all items 155
Figure 4.3 Weight of characteristics 156
Figure 4.4 Theoretical relationship of Theta vs. Likert scale score (RP.06) 168
Figure 4.5 Conditional standard error and person ability (RP.06) 170
Figure 4.6 Theoretical relationship of Theta vs. Likert scale score (RP.10) 174
Figure 4.7 Conditional standard error and person ability (RP.10) 176
Figure 4.8 Theoretical relationship of Theta vs. Likert scale score (RP.05) 180
Figure 4.9 Conditional standard error and person ability (RP.05) 182
Figure 4.10 TCC of DIF for RP.06 based on gender 188
Figure 4.11 TCC of DIF for RP.10 based on gender 188
Figure 4.12 TCC of DIF for RP.05 based on gender 189
Figure 4.13 TCC of DIF for RP.06 based on license status 189
Figure 4.14 TCC of DIF for RP.10 based on license status 190
Figure 4.15 TCC of DIF for RP.05 based on license status 190
Figure 4.16 Descriptive plots for PSUS score based on gender and license status 192
Figure 4.17 Option Characteristic Curves for All Options 195
Figure 4.18 Item Characteristics Curve 195
Figure 4.19 Item map for text + symbol sign (RP.06) 196
Figure 4.20 Item map for symbol-only sign (RP.06) 197
Figure 4.21 Item map for text-only sign (RP.06) 197
Figure 4.22 Kernel Density Function of person estimates for RP.06 198
Figure 4.23 Kernel Density Function of person estimates for RP.10 199
Figure 4.24 Kernel Density Function of person estimates for RP.05 199

Figure 4.25 Cognitive mapping for sign A RP.06 203
Figure 4.26 Cognitive mapping for sign B RP.10 204
Figure 4.27 Cognitive mapping for sign C RP.05 205
Figure 4.30 PSUS of sign C RP.05 206
Figure 4.29 PSUS of sign B RP.10 206
Figure 4.28 PSUS of sign A RP.06 206
Figure 4.31 Structure of SUS and PSUS 214
Figure 5.1 Aspects of sign usability 223

LIST OF ABBREVIATIONS

ARI	Automated Readability Index
CI/Cont.Int	Confident Interval
CLI	Coleman-Liau Index
CSC	Cognitive Sign Characteristics
DIF	Differential Item Functioning
Diff.	Difficulty
Discri.	Discriminant
FKGL	Flesch-Kincaid Grade Level
FOG	Frequency of Gobbledygook
GFS	Gunning-Fog Score
JMLE	Joint Maximum Likelihood Estimation
JMLE Diff.	Joint Maximum Likelihood Estimation Difficulty
MRA	Multiple Responses Analysis
MSE	Mean Square Error
PCAR	Principal Components Analysis of Rasch Residuals
PSUS	Perceived Sign Usability Scale
RP	Regulatory Prohibitive
RPP	Relative Participant Percentage
RRP	Relative Response Percentage
SD/Std. Dev.	Standard Deviation
SEM	Standard Error of Measurement
Sig.	Significant

SMOG	Simple Measure of Gobbledygook
Std. UMS	Standardized Unweighted Mean Squares
Std. WMS	Standardized Weighted Mean Squares
SUS	System Usability Scale
TCC	Test Characteristics Curve
UMS	Unweighted Mean Squares
Var.	Variance
WD	Warning Danger
WMS	Weighted Mean Squares
α	Alpha
$\sum$	Summation of
$\approx$	Approximately Equals to
$\geq$	Greater than or Equal to
$>$	Greater than
$<$	Smaller than
π	Proportion
χ	Chi
$\in$	Element of

CHAPTER 1

INTRODUCTION

1.1 Overview

This study attempts to design, develop, and analyse Perceived Sign Usability Scale (PSUS) Instrument to assess sign usability in relation to cognitive sign characteristics. Sign characteristics are attributes of a sign that constitutes to its overall usability. The identification of cognitive sign characteristics will serve as the groundwork for designing and constructing the instrument. Subsequently, calculation method and interpretation are introduced to facilitate the assessment of sign usability. The instrument is then analysed using various quality measures. Lack of comprehensive instrument to assess sign usability based on cognitive sign characteristics shows there is a pressing need to develop an instrument.

This chapter narrates the background of study and regional road safety climate. Problems and issues that prompted the development of the instruments are explained. This chapter also includes objectives and assumptions of the study. Lastly, common terms used are also defined to avoid misunderstanding.

1.2 Introduction

This section provides introductory information about usability and usability in the context of traffic signs. The background of study introduces the pivoting ideas and important concepts that are used throughout this study. This section also details road safety climate in Malaysia. The information will lead into problem statements.

1.2.1 Background of study

The concept of usability is diverse, always evolving and does not exist in an absolute sense (Brooke, 1996; Dumas & Redish, 1999). A generally accepted definition of usability as defined in ISO 9241-11 is "extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use"(ISO 9241-210:2010(E), 2010). In the perspective of web usability, usability is defined as an attribute of quality that assesses the extent to which a user interface is easy to use, which comprises of five components: learnability, efficiency, memorability, errors prevention and recovery as well as user satisfaction (Nielson, 2012).

Although the principles of usability originated from stand-alone software systems development to improve computer systems, usability concepts and definitions have experienced constant evolution along with the increased interest and awareness about the usability of consumer products (Ryu & Smith-jackson, 2005). In the past, there were efforts to broaden and enrich Human-Computer Interaction's (HCI) concept of usability (Hassenzahl, 2006). This effort is necessary because the concepts and quality components of usability in HCI maybe not be fully applicable for other products. Because of this, usability principles have later been adapted to the development and assessment of communication systems, such as browsers and websites. (Ketola & Röykkee, 2001). From then, usability engineering was extended to cover a wider range of products. Resulting from the variability of application, measures of usability are diverse and covering multiple and distinct aspects. Over time, new usability aspects emerged, practitioners started to measure task completion and response time, error rates, user's satisfaction and enjoyment, ease-of-use, etcetera. There are many usability assessment methods, but past researches offer limited

help in how to select usability measures (Hornbæk & Law, 2007). Therefore, researchers must carefully examine the usability aspects for every purpose.

The usability of traffic signs will be the focal point in this study. The concept of sign usability is still vague as researchers were divided as to what sign usability represents. This ambiguity arises owed to the difficulty in specifying attributes of sign. Simply adopting a generic usability instrument can be hazardous because of the variability of context as well as the characteristics of the product (Abran, Khelifi, Suryn, & Seffah, 2003). Researchers refer usability instruments as a tool to assess usability of products, often as questionnaires and rating scales (Ryu & Smith-jackson, 2005). There are many generic usability instruments available, namely the After Scenario Questionnaire (ASQ) (Lewis, 1995), Post Study System Usability Questionnaire (PSSUQ) (Lewis, 1992, 1995), Software usability measurement inventory (SUMI) (Kirakowski & Corbett, 1993), System Usability Scale (SUS) (Brooke, 1996, 2013) and Standardized Universal Percentile Rank Questionnaire (SUPR-Q) (Sauro, 2004). Refer to Appendix A for some of the generic usability instruments.

There was an attempt to modify generic usability instrument to measure sign usability. Ng, Lo, & Chan (2011b) adapted and modified System Usability Scale (SUS) to assess the usability of safety signs. The modifications include substituting the subject word "system" to "safety signs". Although the authors claimed that the modified SUS can be used to assess the usability of safety sign, it is an established fact that every type of product/interface has different aspects of usability. For example, usability component of software involves the evaluation of interactive graphical user interface. This may not be relevant for other products that do not have a similar type of interface (refer to problem statement in sub-section 1.3.3).

There were attempts in the past to identify cognitive aspects of sign usability because researchers aware that road users "use" signs by perceiving them (McDougall, Curry, & de Bruijn, 1999). This knowledge prompted investigations on how road users perceived sign (Brézillon et al., 2004). Researchers attempted to explain sign usability in terms of human perception using different approaches, such as semiotic approach (Bazire, Tijus, Brézillon, & Cambon, 2005; Brézillon et al., 2004; Tijus, Barcenilla, Lavalette, & Meunier, 2007) and semiotic approach of sign comprehension (Clarkson, Coleman, Hosking, & Waller, 2013) (details in Sub-section 2.2.2). Researchers started to identify cognitive sign characteristics (CSC) and used them to assess sign usability. Cognitive sign characteristics (CSC) are design features of signs that relate to human cognitive processes, such as perception on sign meaning and its memorability (McDougall et al., 1999; Ng & Chan, 2007a). Among the earlier concepts of cognitive sign characteristics include several characteristics such as concreteness, complexity, meaningfulness, familiarity, and semantic distance (McDougall et al., 1999). Ng & Chan (2007a) then refer these sign characteristics as cognitive design features.

1.2.2 Road safety climate and initiatives in Malaysia

Road accident is the leading cause of fatal injuries internationally (Makinde & Oluwasegunfunmi, 2014) with estimated fatal injuries at over 1.2 million while non-fatal injuries recorded 20 to 50 million annually (Rohayu, Sharifah, Jamilah, & Wong, 2012). In 2004, World Health Organization (WHO) stated that road traffic injuries ranked ninth in the list of frequent causes of death. It was predicted that they would become the fifth most frequent cause of death by 2030 (World Health Organization, 2009). Locally, the accident rate in Sarawak is alarming, with the observed nonlinear increment of reported road accidents over 14 years from 2003 to 2017, as shown in Figure 1.1. Figure 1.1 is based on

data taken reported by Road Safety Department of Malaysia (2018). Over this period, the number of reported traffic accidents recorded an overwhelming increase of 58.7%.

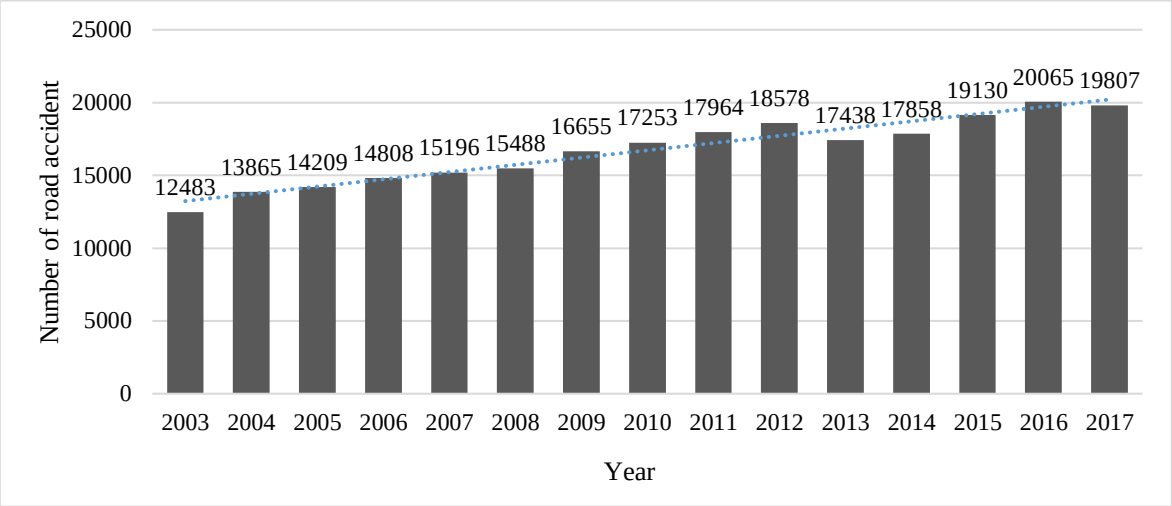


Figure 1.1: Road accidents in Sarawak 2003-2017

Up until 31st December 2017, Road Transport Department in Malaysian Ministry of Transport reported 28,738,194 motor vehicles in Malaysia (Road Safety Department of Malaysia, 2018). The exponential increase of vehicle on the road indicates a crucial need for a more comprehensive road safety measures. Resulting from the increasing number of motor vehicles, organizations dealing with road safety issues emerged within government departments, private institutions and non-profit organizations (Mustafa, 2005). Road Safety Department of Malaysia was established on September 15, 2004, to enhance awareness of the importance of road safety in Malaysia, subsequently helps to reduce fatality and accident rate (Road Safety Department of Malaysia, 2013). Furthermore, Malaysian Institute of Road Safety Research (MIROS) was established on the 3rd of January 2007 to conduct researches related to regional road safety. MIROS also generate and disseminate road safety information through the print media and concerted training programs (Malaysian Institute of Road Safety Research, 2016).

Malaysia is committed in enhancing road safety (Rohayu et al., 2012). The Malaysian government has initiated several measures to enhance road safety through various campaigns and programs. In the aspect of traffic sign comprehension, Malaysian Department of Road Transport has included sign training as part of driver's license training modules. It is compulsory for drivers to pass sign comprehension test (Malaysian Department of Road Transport, 2016). The Malaysian government established a 5-year national road safety target with the aim to reduce road accidents death by 30% by the year 2000. In the year 2001, another national road safety target has been specified. The target was to contain accident rate at around 4 road accidents per 10,000 vehicles by the year 2010 (Law, Radin Umar, & Wong, 2005). A study analysing Malaysian Road Safety Plan 2006-2010 explained that the plan includes many strategies to reduce accident rate and fatality towards achieving 10.0 road deaths per 100,000 population by the year 2010 (Law et al., 2005). This ambitious target prompted a case study to assess its efficacy. From the case study, it was concluded that the Malaysian Road Safety Plan 2006-2010 did not achieve its target (Eusofe & Evdorides, 2017).

Continuity of efforts in road safety initiatives must be preserved with enhanced effectiveness and intensity (Road Safety Department of Malaysia, 2014). The conception of Road Safety Plan of Malaysia 2014-2020 (RSPM 2014-2020) by the Ministry of Transport through Road Safety Department of Malaysia ensured the continuity of road safety initiatives. RSPM 2014-2020 was prepared with the collaboration from MIROS to support Global Decade of Action for Road Safety by the United Nations Commission for Global Road Safety by placing emphasis on outcomes and interventions. RSPM 2014-2020 is implemented based on five strategic pillars, which are road safety management, safer road and mobility, safer vehicles, safer road users, and post-crash management (Road Safety

Department of Malaysia, 2014). Improvement of signs through standardization and regulation was also included as part of the initiative under the second strategic pillar (Safer Mobility and Roads) of the road safety plan. However, its method of execution was not detailed in the action plan.

1.3 Problem Statement

This section outlines existing issues and area of concern leading up to the research objectives. This section details problems that need to be improved upon, difficulties to be addressed, and troubling questions that were found in existing literature. Issues discussed in this section includes poor traffic sign comprehension, lack of sign usability studies and usability instrument, distinction on usability of sign and other products, controversial composite score calculation methods, as well as issues related to existing sign usability methods.

1.3.1 Poor traffic sign comprehension

Adequate knowledge of traffic signs is important for road users regardless of driving experience. Road signs are considered the only way of informing drivers about potential dangers on the road, such as slippery road, rock falling, etcetera (Costa et al., 2014). Despite its role as a tool to relay vital information to users, correct sign comprehension is not guaranteed (McDougald & Wogalter, 2011). Researchers are unanimous in this claim, noting that road users have a below-par understanding of traffic sign (Duarte, Rebelo, Teles, & Wogalter, 2014; Makinde & Oluwasegunfunmi, 2014; Talab, Meshkani, & Mofidi, 2013; Zamanian, Afshin, Davoudiantalab, & Hashemi, 2013). It has been demonstrated that many warning symbols currently in use are poorly understood and that the difficulty may be even

greater for older than for younger adults (Al-Madani & Al-Janahi, 2002; Bazire, Tijus, & Lavalette, 2006; Razzak & Hasan, 2010). A study reported only 56% out of 28 signs tested in the research were well understood by the participants (Al-Madani & Al-Janahi, 2002). Another study noted that signs can be confusing to users (Lay, 2009). This unanimous claim accentuates the realization that it is crucial to have a deeper understanding of road users' perception on the usability of traffic signs.

Road Safety Department of Malaysia in its road statistics book in 2018 has noted that 80.6% of road accidents are resulting from human errors and complacency. Only 19.4% of road accidents are caused by external factors such as road condition (13.2%) and car condition (6.2%) (Road Safety Department of Malaysia, 2018). Poor traffic sign comprehension can cause human errors on the road. It has been frequently reported that poor knowledge on traffic signs leads to accidents on the road, especially when road users failed to execute expected behaviour (Al-Madani & Al-Janahi, 2002; Bazire et al., 2005; Razzak & Hasan, 2010). Since signs are not always elucidative to road users, there is a great variability and misunderstanding in sign comprehension. This misunderstanding threatens the safety of road users (Ben-Bassat & Shinar, 2006). Moreover, poor sign recognition and sign design can lead to impaired communication on the road (Zakowska, 2001).

There is a need to evaluate the design of the existing traffic signs as some signs may not be the most effective design for their respective purpose (Podany, Wogalter, & Mayhorn, 2004). This situation creates another need to assess users' perception of traffic signs to avoid misinterpretation. In addition to poor design and familiarity, traffic signs are not uniform across countries (Ben-Bassat & Shinar, 2006). It was noted that the variation in signing system across countries can be a problem when using symbolic signing for communication (Zakowska, 2001).

1.3.2 Lack of sign usability studies and usability instrument

Sign usability was not given much attention by researchers before 1960 (Horberry et al., 2004). Road safety researches emerged more prominently after 1972, following demands from developing countries. Initially, researches were focusing on determining the magnitude and the nature of road accident. At later stage, the emphasis has been placed on aspects such as trends of road accidents, advancement of data analysis, as well as development and augmentation of appropriate low-cost countermeasures to meet the need to find practical solutions (Downing, Baguley, & Hills, 1991).

Although the usability of traffic control devices particularly traffic signs is one of the most significant and life-saving measures in risk reduction on the road (Ben-Bassat & Shinar, 2006; Ng & Chan, 2007b, 2011), it was not given priority until recent decade (Horberry et al., 2004). In the past, little attention has been devoted in improving sign usability as one of the road safety initiatives. Due to the increased awareness of the importance of signs in the past decade, studies on the usability of signs have been a topic of considerable interest, covering diverse aspects such as engineering, traffic safety, comprehension, design and human differences (Al-Madani & Al-Janahi, 2002; Makinde & Oluwasegunfunmi, 2014). Most sign usability studies are engineering-based. Significant amount of studies were conducted on automated road signs detection (Shneier, 2006; Zakir, Leonce, & Edirisinghe, 2010), visibility of sign (Obeidat, Rys, Rys, & Du, 2016; Sagawa, 2002; Siegmann, 2005), sign luminance (Ho, Scialfa, Caird, & Graw, 2001; Reisinger, 2015; Schnell, Aktan, & Li, 2004), sign conspicuity (Charlton, 2006; Inman, 2012; Miska, Leur, & Sayed, 2002) and cultural issues affecting road signs (Ng & Chan, 2007b; Ward, Wogalter,

& Mercer, 2004). It was noted that studies on cognitive aspects of sign usability are still scarce (Ng & Chan, 2011, 2013; Ng et al., 2011b).

Furthermore, lack of proper usability instruments based on perception on sign characteristics indicates a crucial need for a new method of assessing signs usability. Existing methods to assess sign usability includes visual tracking/eye movement, recall, verbal protocol, naming, driver responses, tachistoscopic presentation, situational awareness and analysis of accident rates (Lansdown, 2004). In Malaysia, no study on sign usability was found, this provides a crucial need to investigate sign usability in local context.

1.3.3 Difference on usability of sign and other products

Different products have different facets of usability. Researchers must examine usability aspects of a product to conduct a valid usability research (Ketola & Röykkee, 2001). Evidently, the lack of user-sign interaction was regarded as the main reason of why sign usability is different than the usability of other products or systems (Lay, 2009). Because of this, adapting generic instruments to assess sign usability is an ill-founded decision. In the case of traffic signs, finding a suitable approach to measure its usability is problematic because currently there is no comprehensive sign usability framework was found because past researches studied usability aspects in isolation. This problem prompted the identification of sign usability aspects as well as the identification of Cognitive Sign Characteristics (CSC) that constitute sign usability. A study noted the importance of researches to identify other signs characteristics and suggested that future researches should examine the possibility of adopting a more inclusive approach (Zhang & Chan, 2013). This recommendation from Zhang & Chan (2013) fuelled the need to study sign usability

involving more sign characteristics. Sign characteristics that affect sign usability will be discussed in Sub-section 2.2.3.

1.3.4 Issues related to existing sign usability methods

A study noted a problem with existing methods of assessing sign usability (Houldin, Stuart, & Murphy, 2012). Problems arise when road users read a sign and are certain that they understand it, but in truth their understanding of the sign is incorrect. Therefore, even if a participant is confident of the meaning, it may not be necessarily correct. When designing the instrument, great care should be taken to ensure respondents' understanding of the signs is genuine and not just relying on respondent's claim saying they understood it (Houldin et al., 2012).

One of the solutions for this is to request the participant to select the meaning of the sign from a list of possible answers (multiple-choice question). This method is a popular method to assess sign usability, as observed in several studies (Dissanayake & Lu, 2001; Jian, Ruili, Zhonghua, & Yongyan, 2011; Makinde & Opeyemi, 2012). However, multiple-choice question has a serious limitation. Researchers do not know if the correct response is due to understanding or a lucky guess. A study noted that the major argument against the use of multiple choice question is that low quality of distracters (wrong options) can increase signs comprehension and thus make the correct meaning of a sign easier to recognize (Ben-Bassat & Shinar, 2006).

To avoid the effect of distractors in multiple choice question, several studies opted for open ended question whereby participants are requested to state the meaning of the sign in the form of free text (Ben-Bassat & Shinar, 2006; Shinar & Vogelzang, 2013; Wogalter,

Murray, Glover, Shaver, & Carolina, 2002). The responses from participants were recorded as written text or verbal response. The responses were then coded according to the degree of correctness (Shinar, Dewar, Summala, & Zakowska, 2003; Shinar & Vogelzang, 2013; Zakowska, 2001). While this approach augments the validity of sign comprehension test, there is no way for researchers to examine other aspects of sign usability.

While single-item instrument (multiple-choice comprehension test and open-ended comprehension test) is quick to administer, single-item instruments lack the inclusiveness and reliability of multiple-item instruments. Psychometric theory suggests that multiple-item instruments are superior and more favourable than a single-item instrument as multiple-item instruments as multiple-item instruments tend to yield more reliable results (Bangor, Kortum, & Miller, 2009). Single-item instruments could not uniquely address specific usability aspects and tend to dilute usability score, thus producing a higher score. Therefore, to incorporate more aspects of usability, using more items is necessary.

To incorporate more aspects of sign usability, researchers experimented with the use of self-administered items. The participants were requested rate whether the presented signs comply to ergonomics principles. Generally, there are five ergonomic principles that were regarded as classical principles of icon design, which are spatial compatibility (arrangement of space), conceptual compatibility (association between symbol and user's expectation), physical representation (symbol being representative of the reality it presents), familiarity, as well as standardization (Ben-Bassat & Shinar, 2006; Jian et al., 2011). The major drawback of this method is that it requires the participants to fully understand each principle. Therefore, this method is unsuitable for road users who are not the expert in ergonomic principles.

A better way to assess sign usability is by adapting a more generalized self-administered instrument by asking about user's perception on sign characteristics instead of asking whether the sign conforms to ergonomic principles without giving contexts. A notable attempt of this is the modification of System Usability Scale (SUS) to the usability of safety signs (Ng et al., 2011b; Ng, Lo, & Chan, 2011a). System Usability Scale is a 10-item Likert Scale intended to assess the usability of computer system; therefore, modifications were needed to cater for safety signs. The modifications include word simplification, paraphrasing, and substitution of subject words (Ng et al., 2011a). Section 2.4 details in-depth analysis and modifications of System Usability Scale. While this is a good attempt to capture user's perception on sign characteristics, the lack of inclusiveness of cognitive sign characteristics cannot be overlooked. Hence, this establishes the need to develop an instrument that is more inclusive, in a way that it can capture sign comprehension with precision as well as allows researchers to examine other aspects of sign usability.

1.3.5 Controversial composite score calculation methods

Sub-section 1.3.4 revealed that self-administered instrument can be the solution to problems of using multiple-choice question, and open-ended question to assess sign usability. In usability research, Likert scale is the most popular scaling method. Although Likert scale is widely adapted, one of the concerns on the usage of Likert scale is its abuse. Rensis Likert (pronounced */ˈlɪkərt/)* first introduced the idea of Likert scale in 1932 as a new technique for attitudinal research (Likert, 1932; Wuensch, 2005). Initially used for psychological and social researches, the use of Likert scale has extended to various fields, such as computational and medical sciences. Commonly, respondents are required to rate their level of agreement or disagreement with a given statement (Sullivan & Artino, 2013). Researchers can also adapt different response anchors such as level of acceptability,

importance, desirability, and frequency (Vagias, 2006). Likert scales yield ordinal data as it is ranked from one extreme to another extreme (Robertson, 2012). Likert scale data are obtained by combining four or more Likert-type items measuring the same construct. This is executed by computing the composite score (typically mean or sum) of those Likert-type items (Boone & Boone, 2012; Warmbrod, 2014). There are many aspects of Likert Scales that are often ignored by researchers. A notable example of abuse is the calculation of composite score and interpretation of Likert Scale instrument. The computation has been subjected to unending controversies.

Likert Scales gained significant popularity since its conception and gave rise to various instruments. It has been adopted extensively over the years, thus became one of the most popular scaling methods because of its simplicity and robustness. Researchers were divided on how to treat Likert scales correctly. There are controversies in interpreting Likert scales, on whether they can be treated as interval data and subsequently used in parametric tests (Sullivan & Artino, 2013). One of the major arguments on the treatment of Likert scales is on the computation of mean and standard deviation to draw conclusions about the data. Likert scales have a rank order, and therefore should be treated as ordinal data (Jamieson, 2004). Some researchers believe that the confusion on the difference of Likert scales and Likert-type items is the root of the problem (Carifio & Perla, 2007). In an attempt to differentiate between the two, another study defined a Likert-type item as a single statement that uses Likert response anchors (Boone & Boone, 2012) as listed by a study (Vagias, 2006). Meanwhile, a Likert scale is a set of four or more Likert-type items that are combined into a single composite score or variable during data analysis (H. N. Boone & Boone, 2012).

Norman (2010) argues that parametric tests can be adapted for Likert scales However, Jamieson (2004) disagreed with this notion, saying researchers tend to analyse Likert scale data incorrectly by adopting inappropriate statistical tests and treat the data as interval level of measurement. In many cases, test assumptions are often violated. The report recognized the problems in using the arithmetic mean for ordinal data. Although it is a common practice to use mean as the composite score for a set of Likert-type items, computation of the mean (and standard deviation) is inappropriate for ordinal data. By adding a response of "Agree" (4) and "Neutral" (3) would yield a mean of 3.5, but it is difficult to interpret what 3.5 means in absolute terms. It is dubious to assume 3.5 refers to "Half Agree" (Jamieson, 2004; Robertson, 2012). Furthermore, when respondents chose the lowest or highest available alternative on the scale, it produces distributions showing a floor or ceiling effect, making the computation of mean in these situations problematic and may not be measurable (Clason & Dormody, 1994). Despite the controversial usage of mean and parametric statistical tests on Likert items and scales, most researchers continue to do so (Brown, 2011).

Some researchers use sum, median, or mode instead of mean to calculate the composite score (Mogey, 1999). They argued that computing sum can provide a way out of aforementioned controversies of using mean (Clason & Dormody, 1994). However, the problem of using sum as the composite score is that the highest value is different across questionnaires and it is difficult to interpret the score. For example, an instrument that has 7 scale size (1-7 scale) and 15 number of items, the highest sum of scores is 105. However, for an instrument that also has 7 scale size and 15 number of items but uses 0-6 scale would have 90 as the highest score. It is cumbersome to interpret a scale from 0-105 or 0-90. Worse, this range will be different for a different instrument. Moreover, conventional methods for calculating composite score encourage researchers to re-code their negative statements into

different variables so that they won't lose the original data when there is an error in the re-coding process (Pallant, 2010). This method will create many extra variables in the data sheet. By using the formula above, there is no need to re-code negative statements into different variables

In designing an instrument, it is crucial to design not only the items and administration method but also the method to calculate the composite score. Researchers were often discombobulated by intricate problems and arguments surrounding the use of conventional composite score calculation method. Because of the discrepancy on the use of traditional composite score methods, this study attempts to find an alternative method to calculate a composite score that devoid of the intricacy of using sum as well as the controversy of using mean. The desired output is a composite score that is more intuitive to comprehend. A formula is introduced to convert Likert scales into a 0-100 scale. This scaling method has already been practised by researchers to present their findings, as it is a better way to handle Likert scales data. However, there is no known universal formula to convert the responses for all variants of Likert-type instruments.

1.3.6 Summary of problems/issues

Poor comprehension of traffic sign was well documented worldwide albeit regarded as one of the significant risk reduction measures on the road. This problem is made worse by the lack of sign usability studies both locally and internationally. The lack of sign usability study is a stumbling block to authorities and policy makers to devise intervention plans to increase sign comprehension or increase sign usability. Sign usability studies were conducted in various ways, each with unique limitations. As this study adapts Likert scale as the scaling method for a new instrument to assess sign usability, controversies on the

computation of composite score are discussed. This study attempts to explain and provide solutions for the issues detailed.

1.4 Objectives of the Study

1.4.1 General objective

This study aims to develop, validate, and analyse the Perceived Sign Usability Scale (PSUS), which is an instrument to assess the sign usability. The development process starts with the analysis of past researches and existing methods to assess sign usability. The analysis serves as the justification of need and gaps in current instruments. The construction of items was performed using several methods. The design of the instrument includes the decision on instrument administration method as well as its interpretation. Subsequently, the instrument is validated using several test signs. Based on the main objective, specific objectives are listed in Sub-section 1.4.2.

1.4.2 Specific objectives

Section 1.3 discussed issues and problems that lead to the research objectives. To achieve the general objective of the study, which is to develop, validate, and analyse the Perceived Sign Usability Scale (PSUS) instrument, this research had several specific objectives:

i. To develop an instrument to assess sign usability based on its cognitive sign characteristics. This includes the following processes:
    a. Developing a framework for the design and development of Perceived Sign Usability Scale (PSUS).
    b. Reviewing existing methods to assess sign usability.
    c. Identifying cognitive sign's characteristics that constitute road signs usability.

d. Developing items for a new measure to assess road users' perception on traffic sign usability.

e. Examining psychometric properties of PSUS.

ii. To develop a formula to calculate the composite score of the new instrument

iii. To determine the best methods to interpret the data obtained with PSUS.

iv. To assess the usability of test signs using PSUS.

To develop an instrument to assess sign usability based on its cognitive sign characteristics, there are four processes involved. After the process of conceptualization and item development is completed, the psychometric properties of the instrument are assessed to ensure the instrument is reliable and valid. The design of the instrument involves the framework for PSUS development (Figure 1.3), analysis of past researches (Section 2.3), identification of CSC (Section 2.4), item development (Section 3.5), and establishing the validity and reliability of the instrument (Section 4.5). After the completion of the instrument, its composite score calculation (Section 3.9) and interpretation methods are suggested (Section 3.10).

Figure 1.2 illustrates how problem statements lead to the specific objectives in this study. Issues about sign usability prompted the development of PSUS. Subsequently, the instrument is tested using test signs. Problems regarding composite score calculation provide the need of developing a composite score formula and the most suitable method to interpret PSUS.

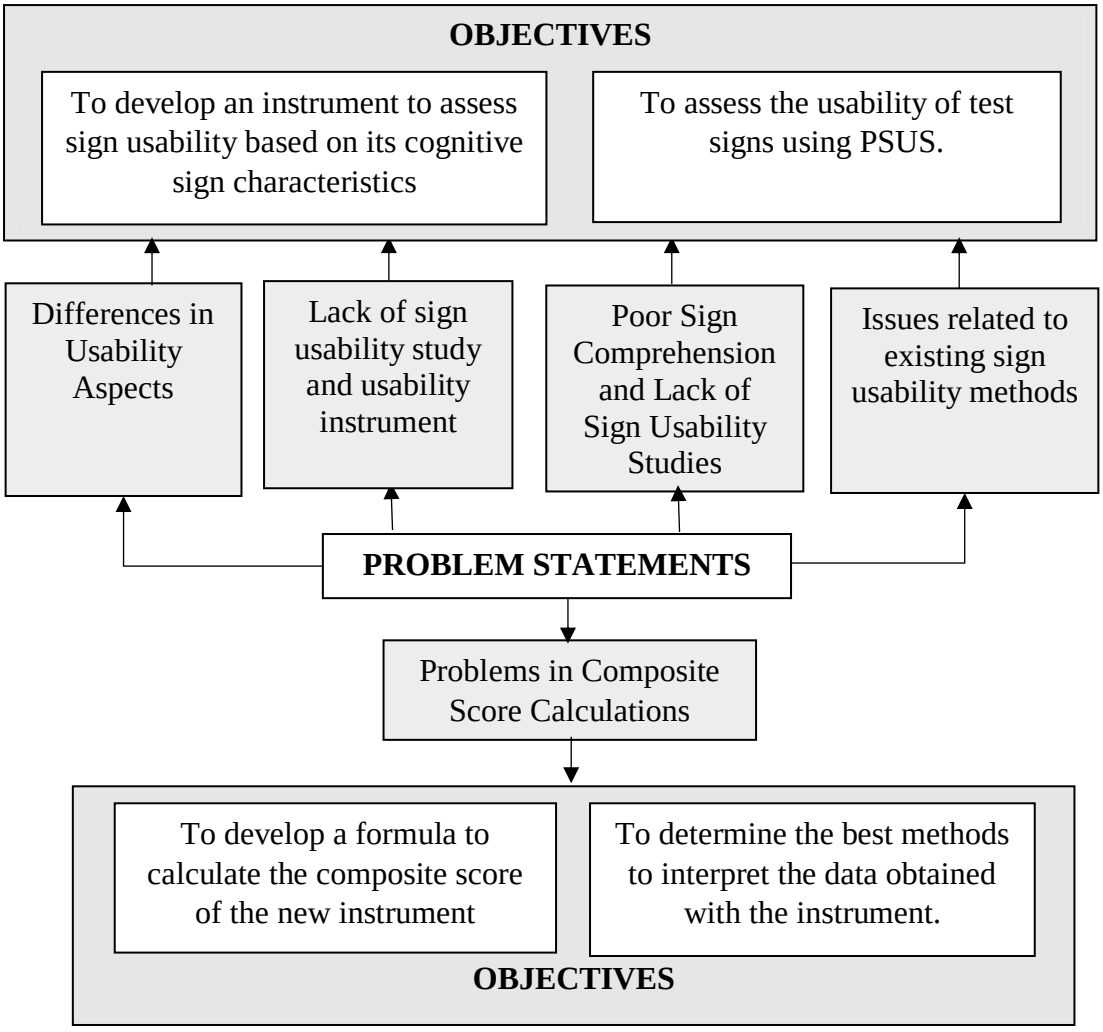


Figure 1.2: Problems leading to research objectives

Based on the design flow as depicted in specific objectives of the study are established to achieve the main objective of the study. The interactions between these stages/actions are also illustrated in Figure 1.3. This serves as the framework to develop PSUS. Instrument development process begins with the analysis of related literature. This process is important to assess the need of developing a new instrument as well as identifying gaps in past researches. Underlying theories from previous researches are identified to serve as the basis of the new instrument. From this information, researcher can decide the best method to calculate composite score, interpret the result, as well as the best method to

administer the instrument. As in Figure 1.3, instrument development process is iterative. Researcher may need to repeat certain process if needed.

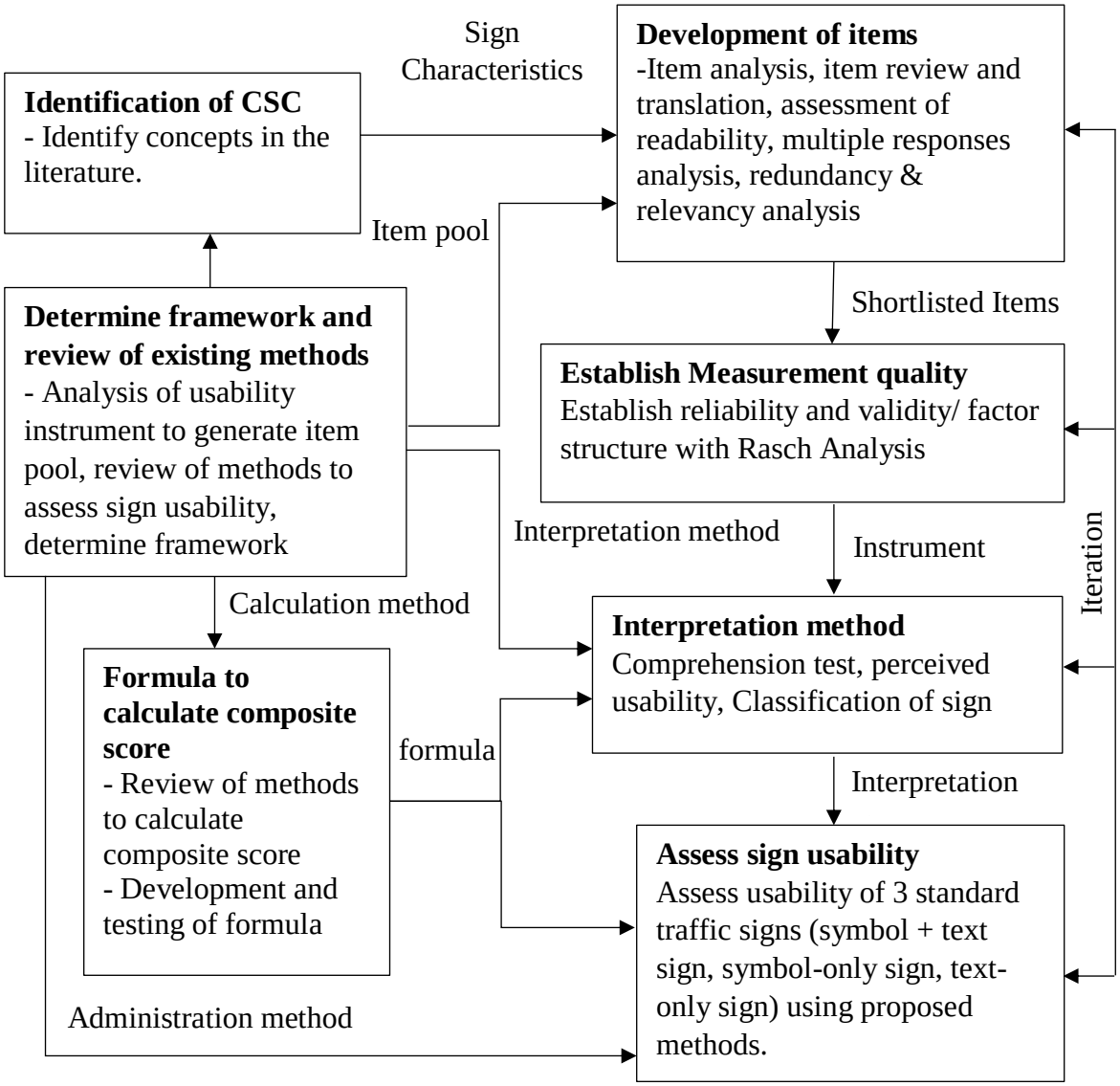


Figure 1.3: PSUS design and development framework

1.5 Significance of Study

Generally, this study can spearhead the emergence of sign usability studies resulting from the identification of cognitive sign characteristics as well as a new instrument to assess sign usability. The knowledge obtained through this study helps to increase awareness and

enhance road safety. Specifically, this study will be of great benefit particularly to the following:

I. Researchers conducting sign usability studies

Organizations and independent researchers on sign usability studies can benefit directly from this study as it serves as a manual on how to use PSUS to assess sign usability. This study provides an instrument as well as interpretation methods for the instrument. Because of this, this study can benefit researchers by changing the way sign usability research will be conducted in the future. The identification of sign usability aspects opens for countless of possible sign usability studies.

II. Researchers developing a new instrument

This study benefits researchers developing new measure for a specific context of use. This study generally acts as a guide to develop an instrument. The processes and methods to develop an instrument are detailed in this study. This study provides a framework for instrument development. Researchers can also adapt methods used in this study to develop an instrument. This includes various techniques, statistical approaches, and software adapted in this study.

III. Researchers and organizations who use Likert Scales

Researchers adapting the PSUS formula to calculate the composite score of Likert-types items not only can avoid the controversial use of conventional composite score computation, they will also benefit from the intuitiveness of the scale produced by PSUS formula. Suggestions to interpret PSUS can be used as a guideline to interpret their instrument.

IV. Authorities and policy makers

In Malaysia, there are several entities responsible for road safety issues, as well as issues related to traffic sign. These entities include Road Safety Department, Road Transport Department, Public Works Department, and Malaysian Institute of Road Safety Research. This study contributes to the works of these entities in relation to traffic sign, such as sign training, redesign, construction and assessment.

V. Road users

PSUS allows road users to detect problematic signs. Interventions by authorities and policy makers in relation to traffic sign improvement can ultimately help road users to enhance sign comprehension of which is beneficial to reduce risk factors associated to poor comprehension of traffic signs.

1.6 Assumptions

Assumptions are uncontrollable occurrences in the study. Assumptions are claims that were regarded as "probably" true before progressing with the study as their absence will render the study irrelevant (Simon, 2011). Literature identified various factors that could contribute to sign usability. It is assumed that in the test, sign usability is influenced by cognitive factors and not physical and environmental factors such as placement, viewing distant, weather, time of the day, and deterioration. Factors contributing to sign usability are further explained in Chapter 2. Instrument administration is conducted in a more controlled laboratory setting to reduce the effect of external factors and nuisance variables.

1.7 Definition of Terms

1.7.1 Traffic signs

Conceptual Definition

Traffic signs are placed along or above a highway, roadway, or pathway. Traffic signs control the flow of traffic, provide directional guidance, reassure drivers about their location, indicate sites of local services, indicate the rules that the user must abide, and warn of upcoming changes or hazards (Mitchell, 2010).

Operational Definition

Traffic signs in this study are standard traffic signs that were taken from the list of standard regulatory and warning sign listed in *Arahan Teknik Jalan* manuals (guide books for traffic control devices) prepared by the Ministry of Works, Malaysia. This study uses three types of sign (text & symbol sign, symbol-only sign and text-only sign). The selection of traffic signs in this study is further detailed in Sub-section 3.7.2.3.

1.7.2 Usability

Conceptual Definition

Generally, usability can be defined as user's satisfaction and quality of experience. International Organization for Standardization stated that products that have good usability allow users to achieve intended goals effectively, efficiently, and with satisfaction (ISO 9241-210:2010(E), 2010).

Operational Definition

Aspects of usability can be different for various types of product. This study focuses on sign usability based on cognitive sign characteristics (CSC). Sign usability can be divided into 3 aspects: Drivers/individual differences, physical/environmental as well as cognitive characteristics. This study focuses on cognitive aspect of sign usability.

1.7.3 Perception

Conceptual Definition

Human perceive a perceptual object by grouping, organizing, and segmenting perceptual elements such as colours, sizes, and shapes that form the object (Firestone & Scholl, 2016; Pinna, 2010). Perceptual elements are captured by visual experiences of which give meaning to the perceptual object (Siegel, 2016).

Operational Definition

The instrument developed in this study is based on user's perception on traffic sign. The perception of sign includes user's opinion on certain characteristics of the sign. The characteristics is listed in Sub-section 1.7.4 and further explained in Sub-section 2.4.1. Participants were requested to provide meaning for the presented signs and rate their agreement with items in the instrument. The responses were captured in the format of open-ended question and Likert Scales.

1.7.4 Cognitive Sign Characteristics (CSC)

Conceptual Definition

Mitchell (2010) has indicated that driver's attention to traffic signs depends on several factors such as driver's characteristics, environmental variables as well as cognitive sign characteristics. Driver's characteristics include demographic, driving experience, driving goal, physical differences (Horberry et al., 2004; Lansdown, 2004; Mitchell, 2010). Environmental factors include weather, placement, time of the day, the number of signs, as well as pollution (Horberry et al., 2004; Lansdown, 2004; Mitchell, 2010; Public Works Department Malaysia, 1985b). Cognitive sign characteristics involve road users' perception on aspects of a sign.

Operational Definition

CSC comprises of 12 characteristics that were identified using documentary analysis: Authoritativeness, Complexity, Comprehension, Familiarity, Learnability, Legibility, Meaningfulness, Memorability, Priming, Saliency, Standardization, and Surface Properties. These characteristics are then reflected into rating scale instrument and subsequently used to assess sign usability. Each CSC is further elaborated in Sub-section 2.4.1.

1.8 Chapter Organization

This study consists of five chapters. Chapter 1 introduces the research project. Research motivation and the statement of problems are discussed, as well as the design framework for the development of PSUS. Chapter 2 is the conceptual basis for the instrument. Chapter 2 answers two specific objectives of this study: to review existing methods to assess sign usability and to identify cognitive sign's characteristics that constitute road signs usability. Review of existing methods includes the understanding of usability instruments, gaps in existing measures and identification of items and approaches that can be adapted in PSUS. Chapter 3 discusses the methodology being implemented in this study and provides answers for three specific objectives: to develop items for new measure to assess road users' perception on traffic sign usability to develop a formula to calculate composite score of the new instrument as well as to suggest methods to interpret the data obtained with the instrument. In Chapter 4, readability, reliability, and validity of the instrument are assessed and subsequently used to test the usability of test signs. Chapter 5 summarizes the study. Limitations and contributions of this research are also addressed. Lastly, recommendations for further studies on sign usability are provided.

CHAPTER 2

LITERATURE REVIEW

2.1 Overview

This chapter is divided into four parts, which are traffic signs, the analysis of sign usability researches, cognitive sign characteristics, (CSC), as well as the analysis of usability instruments. This chapter provides analysis of documents to provide more insightful analysis on the subject and conceptual background. Description of traffic signs includes importance, sign comprehension climate as well as sign usability components. Subsequently, CSC are identified and defined. The understanding of generic usability instrument is important to provide astute analysis on common usability criteria in usability studies and to serve as initial item pool. This chapter also reviews System Usability Scale of which forms the foundation of the scaling method and interpretation of PSUS. This chapter is divided into four different, but related parts as shown in Table 2.1.

Table 2.1: Overview of Chapter Two

Chapter 2: Literature Review			
Part one: Traffic Sign (Section 2.2)	Part two: Sign Usability Researches (Section 2.3)	Part three: Cognitive Sign Characteristics (Section 2.4)	Part four: Usability Instruments (Section 2.5)
Contains general information about traffic signs as well as perception theories related to sign usability. This section provides underlying background for the identification of CSC in Part two.	CSC in part three originated from selected researches on sign usability. This section analyses these researches, in terms of methods used, locality and sample characteristics.	Describes and analyses CSC used in the study. This section also analyses code-occurrence of CSC.	Analyses generic usability instruments that will subsequently used in item pool for the instrument.

2.2 Part One: Traffic Sign

This section provides information on traffic sign in Malaysian context. Because of cultural differences such as language and the way of life, signing system is different across countries (Shinar et al., 2003). Sub-section 2.2 explains the traffic signing system being implemented in Malaysia. This includes standard shape and colour of traffic signs in Malaysia. Sub-section 2.2.2 explains human perception in relation to sign comprehension by using three models of sign perception: schematic model, semiotic model and gestalt principles. The theories explained in this section further justify the appropriateness of assessing sign usability based on human perception of sign. Finally, since it was established in the problem statement (Sub-section 1.3.3) that sign usability is different than the usability of other products, Sub-section 2.2.3 further elucidates different components of sign usability.

2.2.1 About traffic signs

Traffic signs are the most common graphical symbol along the roads and highways and the use of signs is one of the significant measures in road safety (Mukhopadhyay et al., 2013; Ng & Chan, 2008; Ng et al., 2011a, 2011b). Traffic signs are graphical cues that utilize shapes, colours, words, and symbols to communicate with road users. Traffic signs are placed along, beside, or above a highway, roadway, pathway, or other with the purpose of regulating the flow of traffic, provide location, and directional and guidance, show reminders, as well as warn of potential changes or hazards with the aim of reducing road accidents (Mitchell, 2010). There are four categories of traffic signs in Malaysia: regulatory signs (prohibitive and mandatory), warning signs, temporary signs, as well as guide signs (Destination, directional, distance, information, route/section markers, kilometre post) (Public Works Department Malaysia, 1985b, 1985a).

A sign should be placed in a way that the drivers can see and comprehend it within a short time period to make a decision and execute appropriate behaviour (Mitchell, 2010). This is important because signs regulate the flow of traffic, provide location and directional and guidance, indicate sites of local services, rules that the user must abide, as well as warn of potential changes or hazards with the aim of reducing road accidents (Makinde & Oluwasegunfunmi, 2014; Mitchell, 2010).

The importance of traffic signs has been recorded in early researches dated back in 1971, saying drivers need information when they are uncertain of their own perception and judgment on a particular situation (Allen, Lunenfeld, & Alexander, 1971). Traffic signs are crucial in directing, informing and controlling road users' behaviour to enhance road safety for all users. It is important for all drivers including experienced professional drivers to have adequate knowledge of traffic signs (Public Works Department Malaysia, 1985b, 1985c). Moreover, traffic signs help to maintain discipline on the road by providing information on potential dangers (Mitchell, 2010).

The standardization of standard road signs brings benefits to road users, authorities as well as the traffic sign manufacturer. The Ministry of Work has published several manuals on traffic control devices focusing on standardized traffic signs (Public Works Department Malaysia, 1985b, 1985a). Its function is to establish and maintain uniformity in design and application of all traffic signs and control devices in Malaysia and acts as one of the main guides to build traffic signs. It consists of four parts:

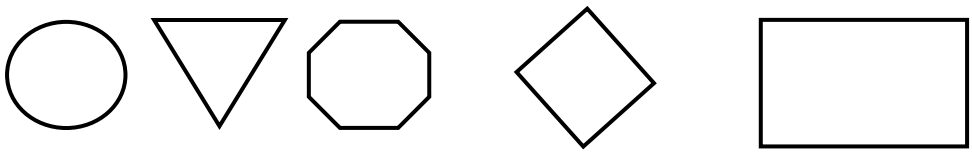
I. Arahan Teknik (Jalan) 2A/85 - Standard Traffic Signs: Standard traffic signs drawings for regulatory, warning, and guide signs. It includes acceptable colours for each sign.

II. Arahan Teknik (Jalan) 2B/85 - Traffic Sign Application: It consists of design criteria and location requirements for all traffic signs. However, this requirement is not applicable for temporary signs.

III. Arahan Teknik (Jalan) 2C/85 - Temporary Road Signs and Work Zones Control: It consists of standard temporary signs drawings and their application in the work zones.

IV. Arahan Teknik (Jalan) 2D/85 – Road Marking and Delineation: This section does not involve traffic signs. It is a guideline for standard road line paints, road marking, and delineator.

These 1985 standards for traffic sign have been revised and reprinted numerous times in different forms throughout the years in print and electronic versions. However, there is no significant change to the document and the traffic signs listed in the standard. This is an alarming situation because some of the signs may be obsolete over the years. According to Public Works Department Malaysia (1985b, 1985c), there are three basic types of traffic sign based on its function: give orders, warn and give information. Each type has a different shape. Figure 2.1 shows different shapes of sign and their context of use.



Regulatory Warning Guide

Figure 2.1: Shapes of sign and their use

As in Figure 2.1, Guide/Information signs are rectangles. Round signs are commonly used for regulatory prohibitive and mandatory signs. The main purpose is to inform drivers

about traffic regulation. The common shape for warning signs is diamond square. The purpose of warning signs is to warn drivers about potential dangers. This includes temporary signs placed near a construction area. Temporary signs are either in diamond shape or rectangle. They are often placed in construction area to avoid accidents in the area due to ongoing construction works (Malaysian Department of Road Transport, 2016; Public Works Department Malaysia, 1985b, 1985c).

Aside from shape, all signs are colour coded based on their meaning and purposes. Table 2.2 summarizes the colour of signs for different purpose works. These colour combinations are adapted in Malaysia, as explained in the manual issued by the Public Works Department of Malaysia (Malaysian Department of Road Transport, 2016; Public Works Department Malaysia, 1985b, 1985c). As it is a standardized colour code, it should be applied consistently to avoid confusion.

Table 2.2: Colour Combinations for Signs of Different Purpose

Purpose	Colour Combination
- River names - Historical and cultural interest	- White on Green background
- Town names	- Black icon/text on white background
- Other information signs	- Blue and orange background
- Warning signs	- Yellow with a black border
- Prohibitive	- Red on a blue background - Red border and white background - Black icon/text on white background
- Mandatory signs	- White icon on blue background
- Temporary signs	- Orange with a black border

Figure 2.2 shows different colour code for standard prohibitive signs as delineated by the Public Works Department of Malaysia in Table 2.2. The example for red

on a blue background is No Stopping sign (RP.11). Restriction on Use of Horns sign (RP.12) is the example for red border and white background, whereas End of Speed Limit (RP.09) is the example for black icon/text on white background.



RP.11 RP.12 RP.09

Figure 2.2: Different colours of prohibitive signs

2.2.2 Human perception in relation to traffic signs

The idea of sign usability in this study centres around the contents of perception, which is defined as the subject of perceptual experience (Siegel, 2016). As described in Sub-section 2.2.1, it was noted that the combination of different colours and shapes is used for different purpose. This sub-section describes how users perceive and give meaning to signs based on the surface properties (shape, colour, icon). In this case, user's visual experience conveys that a sign exhibits certain properties such as colour, shape, and texture. With the understanding of how users perceive signs, the usability aspects of sign can be identified. There are various perception theories can be used to explain the mechanism of how humans perceive signs. This study attempts to explain three models of sign perception: Schematic model, Semiotic model and Gestalt principles. The knowledge of various models explaining human perception in relation to traffic signs facilitates and verifies the identification of CSC (as explained in Sub-section 2.4.1).

2.2.2.1 Schematic Model of Sign Perception

The main predicament in processing iconic information represented in pictograms relates to its meaningfulness (Tijus et al., 2007). This is because a symbol can have many interpretations depending on its context of use. For example, a left arrow at the corner of a street has a different meaning as a left arrow on the edge of a bottle cap. When the arrow is displayed on the street, it represents the message "turn left", while on the bottle cap, the intended message would be "rotate the bottle cap to the left". The perception of the combination of surface properties gives meaning to a sign. This combination of surface properties (icon, shape, and colour) gives information about sign category and subsequently the meaning of the sign (Brézillon et al., 2004). Based on sign perception, road users ultimately decide the appropriate action. A study has called for researches to investigate cognitive mechanism on how road users perceived sign (Brézillon et al., 2004). Figure 2.3 depicts how sign properties and different cognitive processes are involved in sign comprehension (Clarkson et al., 2013).

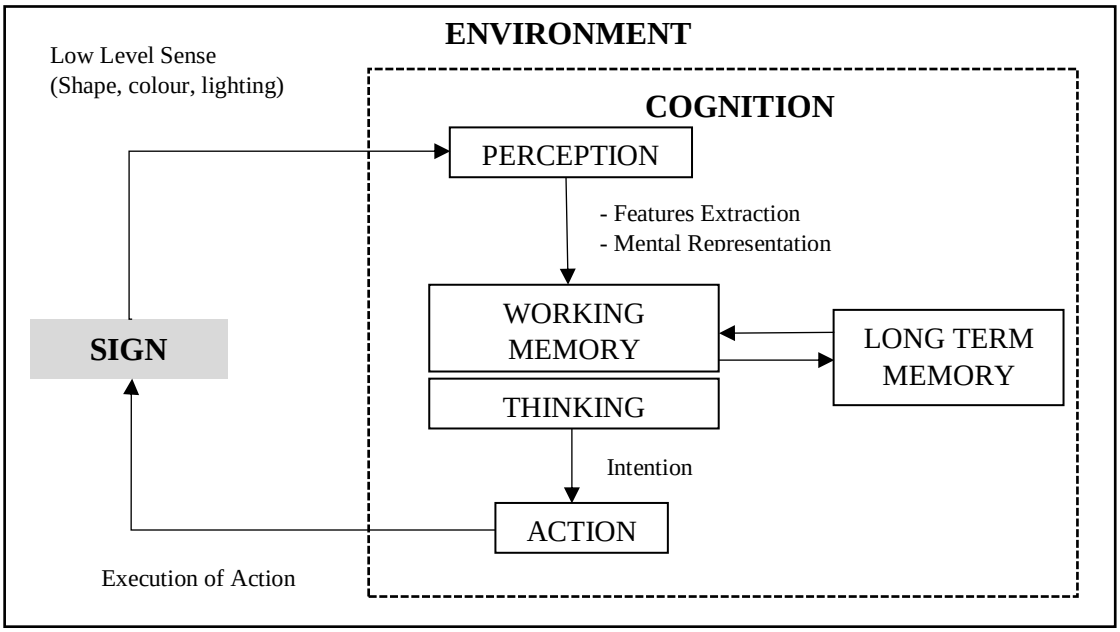


Figure 2.3: Human cognition and understanding of sign (Clarkson et al., 2013)

This schematic model (Figure 2.3) illustrates processes involved in the perception of sign and graphical product in general. Based on this model, low-level sense refers to the breakdown of an object (high-level sense) into smaller pieces of information. In the context of traffic signs, it refers to surface properties such as text, shape, icon, and colour. Features are extracted and assigned a meaning based on mental representation derived from experience. Information is processed in working memory of which prompted actions.

2.2.2.2 Semiotic Model of Sign Perception

Another model explaining the mechanism of sign comprehension is the semiotic model of sign comprehension. This model is an extension to the previous model as this approach further explains features extraction on sign comprehension schematic. Icons and graphic signs, such as road signs have a unique coded meaning. Different colour and shape of a sign provide different coded meaning that must be well understood by road users. The understanding of coded meaning facilitates sign comprehension. Figure 2.4 shows the semiotic approach of understanding sign adapted from multiple studies (Bazire et al., 2005; Brézillon et al., 2004; Tijus et al., 2007) based on sign meaning from Public Works Department Malaysia (1985b). The semiotic approach of sign understanding explains how people comprehend signs by assembling the isolated meaning of surface properties (colour, shape, icon, and text).

Figure 2.4 shows an example of semiotic approach of understanding sign. The example as in Figure 2.4 indicates how different shape (square and circle) gives a different meaning to a sign. First, the human silhouette represents “pedestrian” and a yellow square with black border means “warning”. Therefore, the pedestrian icon on a yellow square prompts road user to be more aware of their surrounding for any pedestrian crossing.

Meanwhile, a red circle means, “Prohibited”. A red line across the red circle also indicates prohibition. Therefore, a pedestrian icon on a red circle would represent “pedestrian prohibited. Knowing the meaning of the red triangle and the red circle makes it easier for road users to interpret the coded meaning of signs.

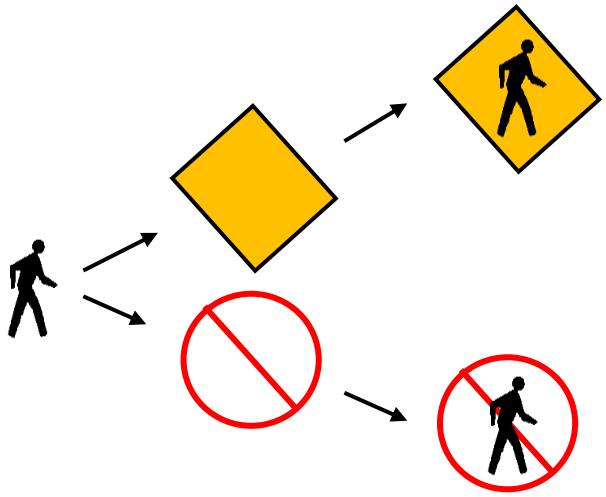


Figure 2.4: Semiotic approach of understanding sign

2.2.2.3 Gestalt Principles of Sign Perception

Gestalt principles facilitate visual perception processing (Peterson & Berryhill, 2013). Gestalt is a German word for 'shape' or 'forms'. Gestalt principles, or gestalt laws, are a collection of rules about the organization of perceptual scenes (Todorovic, 2008). Gestalt law is regarded as a classical law of perception as it has been widely discussed for almost a century. Because of this, this set of rules is not constant, as researchers attempted to reduce and add into the existing laws. Generally, Gestalt principles covers at least 11 laws: balance/symmetry (asymmetrical object appears as incomplete), continuation (instinctively following a direction), closure (instinctively complete unfinished forms), figure-ground (foreground and background), focal point (point of interest), isomorphic correspondence (meaning is shaped by experience), prägnanz (good form), proximity (groping based on placement), similarity (similar objects are perceived collectively), simplicity (unconscious

simplification of object), and unity/harmony (congruity of elements) (Chang, Dooley, & Tuovinen, 2002; Todorovic, 2008).

It is noted that Gestalt principles centres around visual elements of a sign that are similar in proximity, direction, shape, size, and colour. These elements are perceived as part of a group (Graham, 2008). An example of how gestalt principles are used in the design of traffic sign is illustrated in Figure 2.5. This figure explains some of the Gestalt principles used to ease comprehension of Rock Falling Sign.

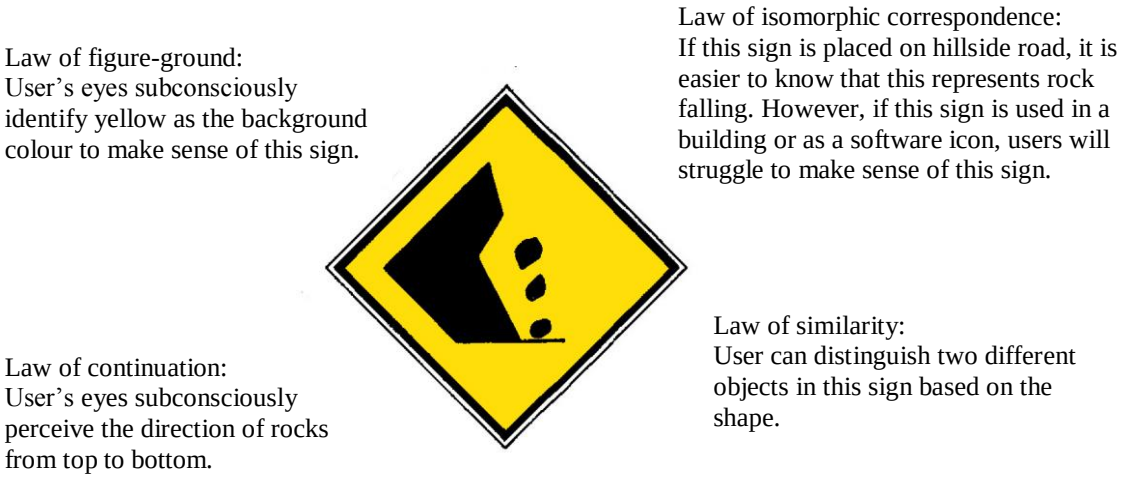


Figure 2.5: Gestalt principles of sign perception

2.2.3 Traffic signs usability component

Signs that conform to ergonomic principles of good design (spatial compatibility, conceptual compatibility, physical representation, familiarity, and standardization) are more comprehensible (Ben-Bassat & Shinar, 2006). In most cases, usability testing of traffic sign only involves sign comprehension (Jian et al., 2011). With the identification of sign usability components, this approach is considered too narrow to represent sign usability. A sign is considered effective if it can be noticed and comprehended well by road users.

The understanding of factors affecting road signs usability is crucial in item conceptualization/development. Mitchell (2010) indicated that driver's attention to traffic signs depends on several factors such as human differences, environmental variables as well as sign characteristics. This review expands this concept by compiling aspects of sign usability and presents the finding in Table 2.3.

Table 2.3: Aspects of Sign Usability

Individual Characteristics	Physical/Environmental Characteristics	Cognitive Sign Characteristics
- Cultural differences/ nationality - Demographic - Driving goal - Experience - Physical differences - Route familiarity - Mental load - Cognitive impairment - Fatigue - Language	- Day/night - Sign placement - Obstruction - Weather condition - Traffic density - Pollution - Sign condition/ deterioration - Number of signs - Materials - Vehicle used	- Authoritativeness - Complexity - Comprehension - Familiarity - Learnability - Legibility - Meaningfulness - Memorability - Priming - Saliency - Standardization - Surface properties (shape, color, font, icon, contrast)

Road users are different in their level of sign comprehension. Individual differences, such as cultural difference can influence comprehension of road signs (Shinar et al., 2003). Shinar et al. (2003) noted that locally used signs are much easier to comprehension as compared to non-local signs. If culture specific signs are used, the use of supplementary text may be needed to avoid cultural bias (Ng & Chan, 2007b). In Malaysia, examples for culture specific signs include signs related to trains and trishaw because these modes of transportation are uncommon in certain states, especially in East Malaysia. Individual differences have always been regarded as nuisance variable and should be circumvented

where possible in usability testing and ergonomic product development (Karwowski, Weaver, Gielo-perczak, Washburn, & Hancock, 2003). However, addressing individual differences in sign usability studies is relevant because it is unlikely for a design to be favourable by every road user. Individual characteristics that affect sign usability are further discussed in Sub-section 2.3.3.

Physical and environmental characteristics include time of the day (day/night), sign placement, obstruction (Lay, 2009), weather condition (Fleyeh, 2008; Peschke, Olk, & Hilgetag, 2013), traffic density (Mitchell, 2010), pollution (Peschke et al., 2013), sign condition (Public Works Department Malaysia, 1985b), natural deterioration (Lay, 2009), number of signs (Public Works Department Malaysia, 1985b), sign sheeting material (Schnell et al., 2004), and vehicle used (Ismail, 2012).

Aside from physical aspect of the sign, researchers have started to study sign usability based on cognitive design features of signs. This shifted the paradigm of studies on surface properties by incorporating user's perception and cognitive processes. Instead of physical attributes of a sign such as the material used, typeface, and placement, CSC deals with sign familiarity, meaningfulness, and complexity among others. A full aspect of CSC is explained in Section 2.4.

2.3 Part Two: Existing Sign Usability Researches

The main purpose of this section is to analyse previous sign usability researches to scope out the key patterns in terms of methods, and participants' characteristics. This provides useful justification for data collection requirements for this study so that the approach adopted was in line with good practices in current researches.

2.3.1 Existing methods for assessing sign usability

Studies on traffic sign cover a wide range of methods (Sun, Lu, Yao, & Rong, 2011; Yuan, Ma, Lei, & Xu, 2014). Most of the traffic sign studies are engineering-based. For example, significant amount of studies have been conducted on automated road signs detection (Shneier, 2006; Zakir et al., 2010), sign visibility (Obeidat et al., 2016; Sagawa, 2002; Siegmann, 2005), sign luminance (Ho et al., 2001; Reisinger, 2015; Schnell et al., 2004), sign conspicuity (Charlton, 2006; Inman, 2012). Cognitive aspect of road signs is still a relatively new research, such as the study on sign comprehensibility (Ben-Bassat & Shinar, 2006; Makinde & Oluwasegunfunmi, 2014; Shinar et al., 2003; Yuan et al., 2014) and cultural issues affecting traffic sign (Ng & Chan, 2007b).

Although the understanding of road signs is crucial, studies on the usability of traffic signs are extremely scarce (Ng et al., 2011b; Yuan et al., 2014). There are several methods of assessing sign usability, namely visual tracking/eye movement, recall, verbal protocol, naming, driver responses, tachistoscopic presentation, situational awareness, and analysis of accident rates (Horberry et al., 2004; Lansdown, 2004). Lack of psychometric usability instruments based on sign characteristics indicates a crucial need for a new method of assessing signs usability. There are three approaches in implementing questionnaire is research. The first is to adapt ready-made instruments developed by other researchers. The second approach is through modification of existing questionnaire to match researcher's interests and context of use. Lastly is to develop new questionnaire if there is no known suitable questionnaire to assess the constructs (Estabrooks & Wallin, 2004).

The self-report methodology is regarded as the most appropriate tool in traffic psychology research to measure cognitive processes, perception, opinion, beliefs, and

attitude. However, unlike reporting perception or opinion, it was noted that self-report methodology could be misused and biased when reporting behaviour, accidents, violation history, frequency, experience etcetera (Lajunen & Özkan, 2011).

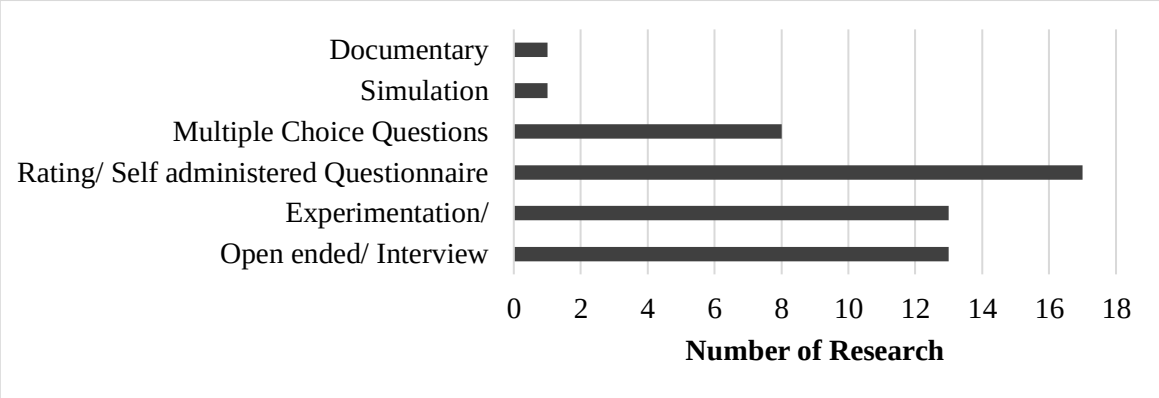


Figure 2.6: Methods to assess sign usability

Figure 2.6 summarizes the methods of assessing sign usability. In conclusion, Self-administrated rating scale and open-ended comprehension test are widely accepted as an appropriate and effective format to assess sign usability. This acts as the basis for developing rating scales instrument and adapting open-ended comprehension test to measure sign usability.

2.3.2 Locality

This sub-section examines the location of existing sign usability researches. Sign usability studies are mostly conducted locally due to differences in signage system. Although international standard exists for traffic signs exist, the variation of icons in signage system is still prominent in many countries. Local authorities tend to modify traffic signs based on the local context of use. Modifications include translation and design tweaks. This modification can be attributed to efforts of the local governing body to bridge cultural and language barriers (Ng & Chan, 2007b). Localized setting maintains the relevancy of

research, but precautions should be taken to avoid over-generalizing the findings. Results from studies conducted in a country may not be applicable in another country. Figure 2.7 shows the places from where sign usability researches was conducted. The data was based on documentary analysis on research papers from 1994 to 2014.

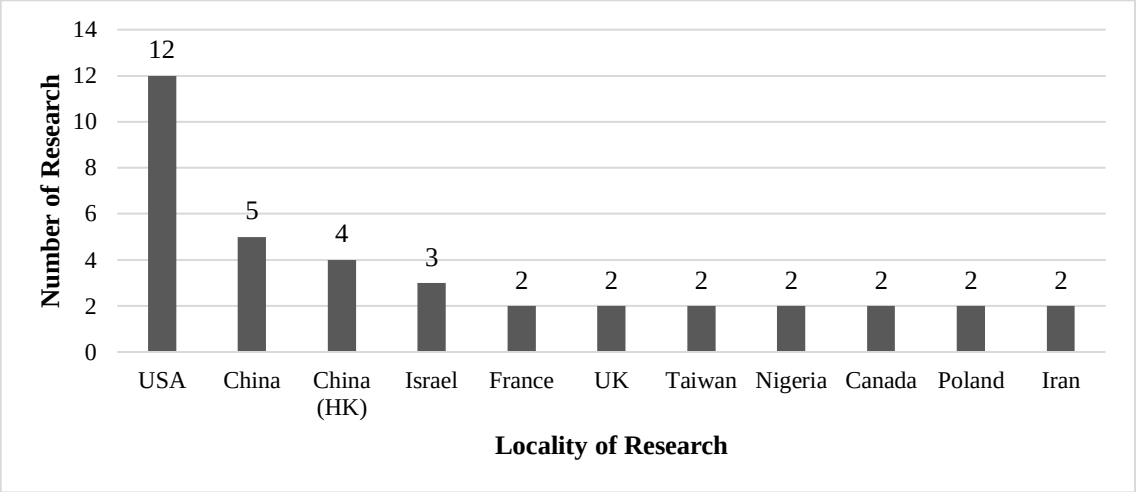


Figure 2.7: The locality of sign usability research

As shown in Figure 2.7, a large proportion of sign usability researches focused on road users in the USA. Researches in China and Hong Kong were mainly conducted by the same group of researchers, using international as well as local signs. In the case of South East Asia, researches on sign usability are extremely limited.

2.3.3 Characteristics of participants in sign usability researches

Describing the nature of the participants is important to understand how sign usability researches were conducted. Participants are generally obtained with non-random sampling technique because of the ambiguous population of road users and difficulties to locate the sample if they are randomly selected for the study. This problem is more prominent for experimental researches. Table 2.4 shows groups of road users who participated in sign usability studies (data from 1994 to 2014).

Table 2.4: Participants in Sign Usability Research

Sample	Authors	Total
College/ university students	(Al-Gadhi, Naqvi, & Abdul-Jabbar, 1994; Ben-Bassat & Shinar, 2006; Campbell & Cowley, 2007; Castro, Moreno-Ríos, & Tornay, 2012; Crundall & Underwood, 2001; Duarte et al., 2014; Koyuncu & Amado, 2008; Laughery et al., 2002; Liu, 2005; McDougald & Wogalter, 2011; McDougall et al., 1999; Mukhopadhyay et al., 2013; Ng & Chan, 2011; Ng et al., 2011b; Podany et al., 2004; Purduski & Rys, 1999; Shinar & Vogelzang, 2013; Ward et al., 2004; Wogalter et al., 2002)	20
Nonspecific drivers	(Akple & Biscoff, 2012; Al-Gadhi et al., 1994; Brémond & Deugnier, 2006; Dissanayake & Lu, 2001; Jamson, Tate, & Jamson, 2005; Jian et al., 2011; Makinde & Oluwasegunfunmi, 2014; Makinde & Opeyemi, 2012; Ng et al., 2011b; Ou & Liu, 2012; Shinar et al., 2003; Yuan et al., 2014; Zakowska, 2001)	13
Public/ general	(Bazire et al., 2006; Borowsky, Shinar, & Parmet, 2008; Ismail, 2012; Moroni & Aymone, 2012; Ng & Chan, 2011)	5

Students in higher educated were often selected as participants because they are easily gathered for experimental research and the hypothesized ideal that they tend to score higher in comprehension test than other drivers. If the purpose of research is mainly about the usability of sign based on the sign characteristics, individual characteristics are regarded as nuisance variable that should be controlled. Researches based on experimental laboratory setting tend to use student sample.

Individual differences in terms of sign comprehension sparked researchers to investigate individual characteristics that affect sign usability. Table 2.5 shows individual characteristics that were found to affect sign usability. Individual differences cited are age, education background, driving experience, gender, culture/nationality, physical differences, driver's familiarity with route, fatigue, mental load and cognitive impairment. Based on the table, educational background, and age were regarded as two of the most influential individual characteristics.

Table 2.5: Individual Characteristics Affecting Sign Usability

Individual differences	Research	No. of Article
Age	(Hancock, Rogers, Schroeder, & Fisk, 2004; Lesch, 2003, 2004; Lesch, Horrey, Wogalter, & Powell, 2011; Lesch, Powell, Horrey, & Wogalter, 2013; Makinde & Opeyemi, 2012; Scialfa et al., 2008; Talab et al., 2013)	8
Education background	(Al-Gadhi et al., 1994; Duarte et al., 2014; Makinde & Oluwasegunfunmi, 2014; Makinde & Opeyemi, 2012; Ng et al., 2011b; Yuan et al., 2014)	6
Driving experience	(Brémond & Deugnier, 2006; Costa et al., 2014; Crundall & Underwood, 2001; Mitchell, 2010; Shaver et al., 2006)	5
Gender	(Al-Madani & Al-Janahi, 2002; Borowsky et al., 2008; Ismail, 2012; Shinar et al., 2003)	4
Culture/ Nationality	(Ng & Chan, 2007c; Ou & Liu, 2012; Shinar et al., 2003; Ward et al., 2004)	3
Language	(Borowsky et al., 2008)	1
Physical differences	(Shinar et al., 2003)	1
Familiarity with route	(Shinar et al., 2003)	1
Fatigue	(Sun et al., 2011)	1
Mental Load	(Scialfa, McPhee, & Ho, 2000)	1
Cognitive Impairment	(Scialfa et al., 2008)	1

Researchers heavily investigated the influence of driver's age. However, there is ambiguity in defining young and old drivers because there is no fixed guideline in categorizing driver's age groups. Generally, young drivers are those who are younger than 35 years of age. Meanwhile, older drivers are those who are older than 50 (Lesch, 2003, 2004, Lesch et al., 2011, 2013). It has been cited frequently that younger drivers record a higher score in sign comprehension test compared to older drivers. Various researchers show many signs currently in use are poorly understood and that the difficulty may be even greater for older than for younger drivers (Al-Madani, 2004; Al-Madani & Al-Janahi, 2002; Bazire et al., 2006; Ben-Bassat & Shinar, 2006; Hancock et al., 2004). Although several studies found no correlation between age and comprehension, there are more overwhelming

evidences suggest the effect of age is significant. A study compares sign comprehension between younger and older drivers before and after sign training. It was concluded that training improves comprehension in both groups, but the effect of age was retained (Lesch, 2004). This observation is expected because of the decrease in information processing capability, attention, and vision (Al-Gadhi et al., 1994).

In Table 2.5, it was also observed that sign comprehension increases with driving experience. However, novice and experienced driver performed differently in sign comprehension test whereby experienced drivers have difficulty to provide the meaning of signs when presented in isolation but not in context. This is different from new drivers who recently completed education on sign comprehension as part of their driver's license requirement (Bazire et al., 2005).

Table 2.6 shows individual characteristics that were observed to be insignificant in sign usability. Gender, age, education background, occupation, marital status, driving distance, and frequency, income, as well as driving experience, were regarded as unrelated to sign usability based on findings of past researches. Among the notabilia is the abundance of researches noting that gender has no relationship on sign usability.

Table 2.6: Individual Characteristics Not Affecting Sign Usability

Individual differences	Research	No. of Article
Gender	(Arphorn, Augsornpeug, Srisorrachatr, & Pruktharathikul, 2003; Brémond & Deugnier, 2006; Costa et al., 2014; Duarte et al., 2014; Makinde & Oluwasegunfunmi, 2014; Makinde & Opeyemi, 2012; Ng et al., 2011a, 2011b; Talab et al., 2013; Yuan et al., 2014)	10
Age	(Brémond & Deugnier, 2006; Chan & Chan, 2011; Costa et al., 2014; Duarte et al., 2014; Jian et al., 2011; Lesch, 2008; Makinde & Oluwasegunfunmi, 2014; Ng et al., 2011a; Yuan et al., 2014)	10

Table 2.6 continued

Occupation	(Chan & Chan, 2011; Jian et al., 2011; Ng et al., 2011a, 2011b)	4
Education background	(Chan & Chan, 2011; Jian et al., 2011)	2
Marital status	(Al-Madani & Al-Janahi, 2002; Ismail, 2012; Makinde & Opeyemi, 2012)	2
Driving distant/frequency	(Ben-Bassat & Shinar, 2006; Jian et al., 2011)	2
Income	(Jian et al., 2011)	1
Driving Experience	(Shinar et al., 2003)	1

Past researches noted that young drivers with good education perform better in sign comprehension tests. Using this group of participants in assessing sign usability would not give an underestimation on underlying usability issues on a sign. It was generally known that external factors need to be controlled in assessing sign usability in laboratory settings (Katz, Dagnall, & Bertola, 2012). Therefore, uncontrolled individual variables may interfere in the study of sign usability based on CSC.

As shown in Table 2.5 and Table 2.6, it was conspicuous that individual characteristics such as education background, driving experience, gender, and age have been studied extensively in the past. However, the influence of these individual characteristics has not been unanimous. These widely discrepant results are apparent even when studies were conducted by the same researcher(s). Blindly conducting more researches on individual characteristics would not be effective in solving this contradiction of result. To solve this problem, researchers should address issues that hinder valid and reliable comparison of individual characteristics. There are three main issues in researches investigating the effect of individual characteristics on sign usability. The most notable issue is the failure to control nuisance variables when conducting the assessment. Sign usability is affected by various

factors. Researchers need to make sure those other factors would not interfere with the validity of the result. To control unwanted variables, controlled laboratory setting was recommended (Brémond & Deugnier, 2006).

Furthermore, there is a lack of uniformity in conducting sign usability testing. Researchers are divided in determining a suitable method to assess sign usability. While some prefer computer-assisted simulation studies (Borowsky et al., 2008; Campbell & Cowley, 2007; Costa et al., 2014; Crundall & Underwood, 2001; Jamson et al., 2005; Liu, 2005) and field test (Al-Gadhi et al., 1994; Purduski & Rys, 1999), some researchers used tachistoscopic still-image representation (Ben-Bassat & Shinar, 2006; Shinar et al., 2003) and rating scale(Mukhopadhyay et al., 2013; Ng et al., 2011b; Podany et al., 2004). This diversification renders it difficult to compare the findings. When there are significant variations in methodology and sample used, the results would not be consistent (Zhang & Chan, 2013).

Lastly, lack of conclusiveness was noted in a study as one of the reasons for contradicting results (Lesch, 2004). For example, researchers should compare the difference in comprehension with more test signs as the difference in sign comprehension may be attributed to a specific sign.

2.4 Part Three: Cognitive Sign Characteristics Constituting Sign Usability

Sub-section 2.2.3 explained that traffic sign usability is affected by physical/ environmental characteristics, individual characteristics, and cognitive sign characteristics (CSC). This section defines and analyses CSC that constitutes sign usability of which acts as the basis for item development. The process of identifying CSC in the literature is explained in 2.4.2. CSC occurrence will be further investigated in Sub-section 2.4.3. Sub-

section 2.4.4 investigates CSC co-occurrence, which is the frequency of CSC occurring together.

2.4.1 Definition of cognitive sign characteristics (CSC)

2.4.1.1 Authoritativeness

Drivers may know the meaning of a sign and what action is needed, but they still ignore the sign. Authoritativeness is an intangible attribute that shows the extent to which a sign prompt the user (driver) to comply or take note of the message. Low authoritativeness results to the violation, and it can lead to accidents (Akple & Biscoff, 2012). Analysis of road accident climate in Tel Aviv as reported by Borowsky, Shinar, & Parmet (2008) noted that violation of traffic signs was cited as the third major cause of road accident. The tendency to comply with traffic sign is one of the measures of sign effectiveness, along with saliency and comprehension (Podany et al., 2004). Authoritativeness of a sign corresponds to hazard, time stress, and mental workload. If the risk perceptive is not deemed to be critical, it increases the likelihood of sign violation, especially if the cost of compliance overweighs the benefit. However, proper placement and interactivity may enhance sign compliance (Wogalter, 2004).

2.4.1.2 Complexity

A sign should have a simple and comprehensible design (Mitchell, 2010). Complex signs are intricate and have many details of which can reduce reaction time. Meanwhile, simple signs contain only few elements or little detail (Ng & Chan, 2011; Zhang & Chan, 2013). The complexity of a sign must be appropriate to optimize saliency, legibility, and comprehension (Tijus et al., 2007). Traffic signs with complex semantic affect the

readability of a visual sign (Bazire et al., 2005). The detrimental effects of complex signs are more apparent when road users are presented with complex information (e.g. traffic and road condition and visual clutter). When road users need to process complex information while driving, they have to react swiftly, thus influencing saliency in visual search performance (Brémond & Deugnier, 2006; Castro et al., 2012; McDougall, Tyrer, & Folkard, 2006). The complexity of sign often associated with surface properties (Bazire et al., 2005; Campbell & Cowley, 2007), the information available to the driver (Jamson et al., 2005) as well as sign familiarity (Ng & Chan, 2011; Zhang & Chan, 2013).

Simplification of sign can increase comprehension (Chan & Chan, 2011; McDougall et al., 1999; Wogalter et al., 2002; Zhang & Chan, 2013), thus minimizing the effect of barrier created by language difference as well as enhancing the effectiveness of sign training (Makinde & Oluwasegunfunmi, 2014; Ng & Chan, 2011). Based on an experiment conducted on comprehension of warning signs, complex design can contribute to misunderstanding and confusion on the meaning of a sign (Yuan et al., 2014). However, in an another comprehension test that was conducted on Taiwanese drivers shows no significant relationship between complexity and comprehension (Ou & Liu, 2012). If icon complexity is inevitable, colour highlighting could be used to increase comprehension (McDougald & Wogalter, 2011).

2.4.1.3 Comprehension

Comprehension is one of the most crucial aspects of sign usability (Ward et al., 2004). The ability to comprehend signs is influenced by individual differences (Shinar et al., 2003). As a sign conveys a message to drivers, it should be comprehensible so that the drivers can execute appropriate actions (Jamson et al., 2005). While the meaningfulness of a sign

reflects whether a sign provides a meaning, comprehension of a sign, however, reflects the correctness of perception and how easy it is to understand the sign. In many cases, level of comprehension is often attributed to the percentage of correct answers in sign comprehension test and correct identification of sign meaning (Dissanayake & Lu, 2001; Ismail, 2012; Jian et al., 2011; Makinde & Oluwasegunfunmi, 2014; Makinde & Opeyemi, 2012; Yuan et al., 2014).

Ng & Chan (2007b) referred to the percentage of correct sign comprehension as the "Guessability" of a sign. Ng & Chan (2007b) compared several similar traffic signs in Mainland China and Hong Kong to examine the guessability of those signs. Figure 2.8 shows the example of signs that have low and high guessability score in Ng & Chan (2007b) study. As in Figure 2.8, M5 (Keep Right) sign recorded the lowest guessability score, with the mean percentage of 2.44%. Meanwhile P20 (No Left Turn) sign recorded the highest guessability score with the mean percentage of 100%. M5 sign was often misinterpreted as turn right, steep descend, and U-turn sign (Ng & Chan, 2007b). Signs that have low level of comprehensibility will not only fail to convey the intended meaning but may also cause confusion that can distract driver's attention and increase the risk of traffic accident (Ward et al., 2004).



M5 – Keep Right (Low Guessability)



P20 – No Left Turn (High Guessability)

Figure 2.8: Example of sign that have low and high guessability score.

The American National Standards Institute (2007) in ANSI Z535.3 report stated that the minimum percentage of correct answers for an open-ended sign comprehension test is at 85%. However, it is generally believed that most signs (traffic signs and safety signs) fail to fulfill this requirement (Duarte et al., 2014; McDougald & Wogalter, 2011; Ward et al., 2004; Wogalter et al., 2002).

Even if a sign is regarded as meaningful, the perceived meaning is not necessarily correct. A study identified two measures of sign comprehension, namely dynamic and static comprehension. Dynamic comprehension refers to correct identification of a sign in a presence of other distractors such as objects and motion. Meanwhile, static comprehension refers to correct identification when there are no distractors, such as comprehension test with still-image (Charlton, 2006). Comprehension was found to be associated with other CSC. For example, when the sign comprehension is poor, its memorability and its priming effect will suffer (Charlton, 2006). Comprehension is also linked to familiarity, as a study noted that comprehension was higher for more familiar signs (Shinar et al., 2003).

2.4.1.4 Familiarity

Familiarity if defined as the extent to which the user is familiar with the sign from previous encounters with the sign (Sanders & McCormick, 1993). Familiarity is an important factor that determines sign usability (McDougall et al., 1999). Sign familiarity reflects the frequency of prior exposure with the sign (Hancock et al., 2004; McDougall et al., 1999). In familiarity test, participants are often asked to rate their familiarity of signs in self-administered rating scales. Familiarity is sometimes investigated independently. However, this should be done with circumspection because it was noted that administering familiarity test anterior to the comprehension test could have detrimental effects on validity as

familiarity increases comprehension (Hancock et al., 2004). This is explained by a study that noted that the participants could assign a meaning a meaning of a sign by drawing from their experiences, although the meaning might not be correct (McDougall et al., 1999).

Familiarity appears to influence conspicuity of a sign. Participants who are familiar with the tested signs felt less visual pressure than participants who are not familiar with the tested signs. A cross-analysis revealed participants that were familiar with the tested signs generally have better comprehension (Liu, 2005). Familiarity also showed association with sign priming, as a study noted that exposure to traffic signs influence driver's behaviour on the road (Crundall & Underwood, 2001).

2.4.1.5 Learnability

Learnability of a sign is regarded as the extent to which the meaning of a sign can be perceived without elucidation and help (Chan & Chan, 2011; Ng et al., 2011b). It involves the cost required to achieve an acceptable level of comprehension (Ng et al., 2011a). Learnability of a sign depends on the context to which a sign is presented. Learnability of a sign is greater if the user is familiar with the sign (Shinar et al., 2003; Shinar & Vogelzang, 2013).

Young adults have much greater ability to learn signs (Lesch, 2008). While young drivers tend to memorize the meaning, experienced drivers tend to understand traffic sign more when given a context (Bazire et al., 2005). Paired-associate learning, recall training, and recognition training are the three methods of learning traffic signs. Paired-associate learning is performed by pairing a sign with its meaning while recall method is conducted by having learners to recall the meaning of signs they have learned. On the other hand,

recognition method involves the identification of sign meaning from a list of option (Ng & Chan, 2011).

2.4.1.6 Legibility

Sign legibility has been moderately investigated in the past especially in the field of traffic engineering. Although physical or external factors: namely, luminance (material), placement, distance, and ambience can affect the legibility (Reisinger, 2015), sign legibility can also be a cognitive characteristic as it also involves users' perception and how typeface, layout, and surface properties are recognizable (DuBay, 2004; Funkhouser, Chrysler, Nelson, & Park, 2008).

Legibility is not to be confused as saliency/conspicuity. A study noted the relationship between conspicuity and legibility. It was noted that excessive luminance could worsen legibility although it increases sign conspicuity (Reisinger, 2015). In a laboratory setting, environmental factors are eliminated (Katz et al., 2012). This allows participants to judge whether a sign is easy to read based on the design in relation to icon design, text legibility and surface properties such as colour, contrast ratio, and lettering (Funkhouser et al., 2008). Symbols and texts do not have the same level of legibility. It is expected that symbols have higher chances of being recognized than text at a greater (Xie et al., 2007).

2.4.1.7 Meaningfulness

Meaningful signs convey messages effectively and increase situational awareness. A sign with a high degree of meaningfulness is likely to be understood by users. Effective signs convey simple and clear meaning (Charlton & Baas, 2006). This is regarded as the most important and popular aspect of sign usability. This claim is backed by the abundance of

researches investigating road users' understanding of traffic signs. Attributes such as semantic distance and concreteness were investigated pertaining to the meaningfulness of sign. The semantic distance of a symbol refers to the relatedness or closeness to what it represents (McDougall et al., 1999; Mcdougall & Curry, 2001; Ou & Liu, 2012; Zhang & Chan, 2013). Concreteness refers to the extent to which an icon is abstract or arbitrary. It was argued that semantic closeness should be the prime determinant for comprehensibility (Mcdougall & Curry, 2001). However, a vast amount of traffic signs do not contain any symbols. This limitation acts a hindrance factor to incorporate semantic closeness and concreteness to evaluate meaningfulness of sign. A universal solution is to assess if a sign makes sense to the user and provides clear meaning.

2.4.1.8 Memorability

Memorability is the extent to which users can easily recall and recognize sign that they have learned. Memorability refers to the tendency to recall sign meaning as well as the ability to recall sign encounter. Although road users are aware of the traffic signs during driving, poor memorability will delay the execution of appropriate behaviour (Akple & Biscoff, 2012). In a study conducted by Costa et al. (2014) using eye-tracking technique to investigate sign recall rate, most drivers failed to recall test signs they encountered. This situation is concerning, as traffic sign recall rate is alarmingly low, at only 6.66% (Costa et al., 2014). Moreover, there is a possibility whereby drivers can recall road signs they have seen but failed to comprehend them (Jamson et al., 2005). Memorability can be associated with familiarity as signs can be perceived more efficiently if the drivers have been exposed to them in the past, regardless of whether they remember having seen them (Koyuncu & Amado, 2008).

The concept of sign memorability is consistent with the psychological theory of memory (Lesch, 2008). In order for a sign to meet its purpose, user's conscious remembrance and compliance of a sign are crucial (Akple & Biscoff, 2012). A study conducted by drivers Al-Gadhi et al. (1994) found that memorability is related to age. From the result of the study, it was noted that older drivers have poorer detection and recall rate as compared to younger drivers. In another study, Hancock, Rogers, Schroeder, & Fisk, (2004) disagreed, saying that older adults tend to perform better. Studies on memorability are mostly utilizing eye-recognition systems to detect eye fixation, in addition to verbal and behavioural responses (Jamson et al., 2005).

It is likely for a sign to have low comprehension rate when the level of memorability is low (Akple & Biscoff, 2012). Therefore, recall and recognition training can be conducted to enhance comprehension (Ng & Chan, 2011; Ou & Liu, 2012). Liu (2005) concluded that pictorial signs have a higher degree of memorability as the use of icons positively affects the rate of correct responses. Although the result is true for younger drivers, using pictorial signs can have an adverse effect on recall rate by older drivers (Scialfa et al., 2008).

2.4.1.9 Priming

After successful comprehension of a sign, its priming effect prompted drivers to execute an appropriate action(s). Knowing the meaning of a sign is not an indication that road users know what action(s) is expected of them. Road users should know the action expected from them when encountering a sign in a road scenario. It can be hazardous when executed behaviour does not match intended meaning of a sign (Bazire et al., 2005). For example, prohibition signs are intended to warn about forbidden actions while obligations signs are intended to remind about compulsory actions. The priming effect often differs

across signs. Text-only signs and a text + symbol signs elicit a greater magnitude of priming effect compared to symbol only signs (Koyuncu & Amado, 2008). In emergency cases, the more confident the drivers with the course of action, the quicker they can respond to avoid hazards (Lesch, 2008). Comprehension, priming, and authoritativeness of a sign are interrelated. Even if a person understands the warning and knows about the course of actions to be taken, there is still a possibility that they do not obey the warning's message, nor do they motivated to take action (Wogalter, 2004).

2.4.1.10 Saliency

Sign saliency is the ability to attract driver's attention who are not specifically searching for it, especially in a complex environment (Brémond & Deugnier, 2006). Saliency is measured by the likelihood of a sign to be noticed by drivers (Campbell & Cowley, 2007; Laughery et al., 2002; Podany et al., 2004). Some signs can hardly attract driver's attention because of the existent of less relevant information embedded in the sign that can lessen the conspicuity of a sign (Akple & Biscoff, 2012). Saliency is an important aspect because traffic signs need to be conspicuous in order to be effective as the inadequacy of sign conspicuity can lead to sign violation (Campbell & Cowley, 2007; Charlton & Baas, 2006). Although saliency generally involves cognitive processes, it is heavily dependable on the placement, situation, gender, language, surface properties and car speed (Akple & Biscoff, 2012; Al-Gadhi et al., 1994; Bazire et al., 2005; Borowsky et al., 2008; Costa et al., 2014).

Saliency shows independency with other cognitive sign characteristics. For example, a study shows that although signs are noticeable, it does not necessarily comprehensible by road users (Akple & Biscoff, 2012). Saliency may also be influenced by driver's individual

characteristics. Borowsky et al. (2008) mentioned that novice drivers have a higher tendency to miss more traffic sign, as they need to be more attentive on driving task.

2.4.1.11 Standardization

Standardization is defined as the extent to which surface properties are consistent with all signs (Ben-Bassat & Shinar, 2006; Sanders & McCormick, 1993; Shinar et al., 2003). Standardization also means "strict adherence under all circumstances and in all places to the symbol used for a given object it represents" (Shinar et al., 2003, p. 1561). Standardization is regarded as one of the good ergonomic design principles of traffic sign along with meaningfulness, familiarity, and surface properties (Ward et al., 2004; Zakowska, 2001). Standardization is closely associated with other CSC, namely meaningfulness, and familiarity (Ben-Bassat & Shinar, 2006; Mukhopadhyay et al., 2013; Shinar & Vogelzang, 2013).

Having international standards is important to create consistency and minimize cultural differences (Ben-Bassat & Shinar, 2006). In line with the escalation of traffic volume, countries have implemented several initiatives to standardize their traffic control devices to accommodate the needs of international drivers and enhance traffic safety in general (Ben-Bassat & Shinar, 2006; S. Charlton & Baas, 2006; Makinde & Oluwasegunfunmi, 2014). Various institutions emerged to establish standardized signs, namely the British Standard Institution 1989, the International Standards Organization (ISO), 1982, 1994 as well as the American National Standard Institute (ANSI) Z535, 1998 (Laughery et al., 2002; Wogalter, 2004). For example, the Vienna Convention on Road Signs and Signals has been undersigned by 64 countries as at 2004 as standardization guideline for

their traffic signs (Costa et al., 2014). In 1968, 18 European countries agreed on standardization of traffic signs.

Despite the introduction of standard signage system, traffic signs are not globally standardized, as countries often modify traffic signs according to their cultural context and different degree of relevancy in a localized context of use. These modifications include the use of local language and new icons. Moreover, new signs were also introduced, resulting in gradual loss of uniformity (Shinar et al., 2003). Because of this unstandardized use of signs, it has been documented that international drivers have difficulties to comprehend local signs (Chan & Chan, 2011; Shinar et al., 2003; Shinar & Vogelzang, 2013; Zakowska, 2001). Therefore, traffic signs across countries should be standardized as much as possible to minimize this problem. However, standardization alone should be backed with easily comprehended surface properties and increased familiarity resulting from sign training (Ben-Bassat & Shinar, 2006; Dissanayake & Lu, 2001; Mukhopadhyay et al., 2013; Shinar & Vogelzang, 2013; Zakowska, 2001). Having local standards that are not used consistently will create confusion and can be dangerous (Ben-Bassat & Shinar, 2006).

2.4.1.12 Surface Properties

Surface properties of a sign are connected with a certain expected course of action (Bazire et al., 2005). In the case of traffic signs, warnings and information are received via the visual or surface properties. The term surface properties were described as the shape, colour, font, symbol/icon, text, and other possible visual attributes of a sign (Bazire et al., 2005; McDougald & Wogalter, 2011). The term “icon”, “symbol” and “pictogram” are often used interchangeably in the literature. An icon represents an image of a recognizable object, often depicts as a silhouette. Text signs are dependent on wording and typeface to relay

information rather than silhouettes. The combination of symbol and text is considered as more effective (Arphorn et al., 2003).

Surface properties are associated with familiarity, meaningfulness, and complexity (McDougall et al., 1999). A study noted that types of signs (symbol only, text only, symbol and text) significantly affected correct response rate and reaction time. The study also noted that adding text improves sign comprehension of signs, particularly unfamiliar signs (Shinar & Vogelzang, 2013). Surface properties can also affect saliency whereby distinctive colour can be used to increase saliency in a noisy background (McDougald & Wogalter, 2011), though the effect of colour will diminish as novice users become experts (McDougall et al., 1999). Colours used in signs can carry important meaning, thus have an important impact on users' cognition and behaviour (Elliot & Maier, 2014).

Signs can be categorised into three categories based on surface properties: symbol + text sign, symbol-only sign, and text-only sign. While the use of text is straightforward, the use of an icon is needed for relay complex meaning. Due to a mass of languages spoken worldwide, it is not feasible to use exclusively texts to convey a universal message. Individual differences in languages spoken and level of literacy can be a barrier in comprehending text-only signs (Lesch, 2008). To solve this problem, international signs (symbols) are adapted in place of texts. Symbols are used to provide a language-free means of communicating to the road users, thus have a higher probability of being comprehended by different groups of people varying in language competency or background (Costa et al., 2014; Lesch, 2008; Mukhopadhyay et al., 2013; Yuan et al., 2014).

2.4.1.13 Summary

Sub-section 2.4.1 has provided the definition for 12 CSC: Authoritativeness, Complexity, Comprehension, Familiarity, Learnability, Legibility, Meaningfulness, Memorability, Priming, Saliency, Standardization, and Surface Properties. All CSC relates to human cognitive processes and user's perception on the usability of sign. By studying these cognitive aspects of sign characterises, the usability of sign can be assessed. These characteristics were identified in the literature as shown in Sub-section 2.4.2 below.

2.4.2 Identifying CSC constituting sign usability

Sub-section 2.2.3 noted the existence of CSC as a component affecting sign usability. This sub-section lists articles containing the concept of at least one CSC. The process of identifying CSC is described in Section 3.4. Table 2.7 shows sources of CSC. The number of articles should not be assumed as an indicator of the importance of the CSC. To make a conclusive remark about the importance of certain CSC, more researches should be conducted. The abundance of articles shows that a specific CSC is widely recognised by researchers.

As in Table 2.7, Learnability was the least investigated characteristic. On the other hand, Comprehension and Meaningfulness are two of the most widely studied CSC as the indicator of sign usability. Sign comprehension was often studied as the prime indicator for sign usability relevant to CSC. Meaningfulness and Comprehension are two distinct concepts with a high degree of co-occurrence. There exist a strong association between these two characteristics. If a sign is perceived to be meaningful, users can attach a meaning to a sign, of which helps users to comprehend the sign (McDougall et al., 1999).

Table 2.7: CSC and Article Sources

Cognitive Sign Characteristics	Articles
Comprehension/ Meaningfulness	(Akple & Biscoff, 2012; Al-Gadhi et al., 1994; Arphorn et al., 2003; Bazire et al., 2006; Ben-Bassat & Shinar, 2006; Castro et al., 2012; Chan & Chan, 2011; Dissanayake & Lu, 2001; Duarte et al., 2014; Hancock et al., 2004; Ismail, 2012; Jian et al., 2011; Laughery et al., 2002; Lesch, 2003, 2004, 2008, Lesch et al., 2011, 2013; Makinde & Oluwasegunfunmi, 2014; Makinde & Opeyemi, 2012; McDougald & Wogalter, 2011; McDougall et al., 1999; Ng & Chan, 2011, 2013; Ng et al., 2011a; Ou & Liu, 2012; Podany et al., 2004; Purduski & Rys, 1999; Scialfa et al., 2008; Shinar et al., 2003; Shinar & Vogelzang, 2013; Talab et al., 2013; Ward et al., 2004; Wogalter, 2004; Wogalter et al., 2002; Yuan et al., 2014; Zakowska, 2001; Zamanian et al., 2013; Zhang & Chan, 2013)
Familiarity	(Ben-Bassat & Shinar, 2006; Chan & Chan, 2011; Hancock et al., 2004; Lesch et al., 2011; Liu, 2005; McDougall et al., 1999; Mukhopadhyay et al., 2013; Ng & Chan, 2011; Ou & Liu, 2012; Shinar & Vogelzang, 2013; Zakowska, 2001; Zamanian et al., 2013)
Surface Properties	(Costa et al., 2014; Koyuncu & Amado, 2008; Ng & Chan, 2013; Peschke et al., 2013; Scialfa et al., 2008; Shinar & Vogelzang, 2013; Zakowska, 2001)
Complexity	(Chan & Chan, 2011; Lesch et al., 2011, 2013; Ng & Chan, 2011; Ou & Liu, 2012; Wogalter et al., 2002; Yuan et al., 2014)
Legibility	(Ismail, 2012; Jamson et al., 2005; Moroni & Aymone, 2012; Ng & Chan, 2013)
Priming	(Bazire et al., 2006; Borowsky et al., 2008; Campbell & Cowley, 2007; Crundall & Underwood, 2001; Koyuncu & Amado, 2008)
Standardization	(Ben-Bassat & Shinar, 2006; Mukhopadhyay et al., 2013; Shinar & Vogelzang, 2013; Zakowska, 2001)
Saliency	(Al-Gadhi et al., 1994; Borowsky et al., 2008; Brémond & Deugnier, 2006; Campbell & Cowley, 2007; Laughery et al., 2002; Podany et al., 2004)
Authoritativeness	(McDougall et al., 1999; Podany et al., 2004)
Memorability	(Akple & Biscoff, 2012; Al-Gadhi et al., 1994; Lesch, 2003; Liu, 2005)
Learnability	(Dissanayake & Lu, 2001)

Sign comprehension should be regarded as a different sign characteristic because the perceived sign meaning does not necessarily the true meaning of the sign. This shows even if a sign is meaningful, the sign can still be subjected to misinterpretation. The magnitude of

co-occurrence between meaningful and comprehension is further discussed in Sub-section 2.4.4.

Table 2.8 shows the distribution of CSC based on year of publication in the span of 20 years (1994-2014). While this table does not show the full list of all available researchers on sign usability, it provides useful information on the pattern of CSC researches. Sign comprehension was treated as sole criteria for cognitive sign characteristics until the emergence of other CSC. The identification of other CSC does not stop the researchers to continue to study sign comprehension and meaningfulness.

Table 2.8: Distribution of CSC based on Year of Publication in 2 Decades

1994-1999	2000-2004	2005-2009	2010-2014
Meaningfulness (3) Memorability (1) Saliency (1) Familiarity (1)	Meaningfulness (9) Saliency (3) Surface properties (2) Familiarity (2) Authoritativeness (2) Priming (1) Learnability (1) Standardization (1) Complexity (1)	Meaningfulness (4) Priming (4) Conspicuity (2) Surface Properties (2) Memorability (2) Familiarity (2) Saliency (2) Standardization (1) Legibility (1)	Meaningfulness (19) Familiarity (5) Surface Properties (5) Complexity (4) Legibility (4) Standardization (2) Memorability (1)

2.4.3 CSC code occurrence

From Sub-section 2.4.2, 12 CSC have been identified: Authoritativeness, Complexity, Comprehension, Familiarity, Learnability, Legibility, Meaningfulness, Memorability, Priming, Saliency, Standardization, and Surface Properties. This sub-section investigates the number of times a concept (CSC) appears in document pool. The selection method of documents is described in Section 3.4. It is important to note that CSC code occurrences represent only code occurrence on the documents gathered, and the results should be

extrapolated beyond the document pool. There is a possibility that other relevant researches are not being included because of limitations in search method and search engine capability.

It is acknowledged that coded concept does not mean a concept was being investigated in the research. CSC code occurrence table provides insightful information on most mentioned CSC. Based on the number of code and relative code occurrence frequency, it is possible to know which CSCs are frequently discussed in sign usability researches. Relative code occurrence frequency has the sum of 1.

Table 2.9 Meaningfulness recorded the highest number of occurrence (979) and Surface Properties follows in a close second (966). Although sign meaningfulness and comprehension were regarded as the main indicator for sign usability, the abundance of code for surface properties is attributed to the scope of surface properties that includes sign colour, shape, text, and icon.

Table 2.9: Relative Code Occurrence Frequency of CSC

Cognitive sign characteristics	No. of code	Relative code occurrence frequency	Cumulative percentage
Meaningfulness	979	0.1985	19.85
Surface properties	966	0.1959	39.44
Comprehension	890	0.1805	57.48
Familiarity	484	0.0981	67.30
Authoritativeness	241	0.0489	72.18
Standardization	197	0.0399	76.18
Complexity	255	0.0517	81.35
Saliency	257	0.0521	86.56
Legibility	225	0.0456	91.12
Priming	202	0.0410	95.21
Memorability	131	0.0266	97.87
Learnability	105	0.0213	100.00
TOTAL	4932	1.0000	-

Figure 2.9 is the graphical representation of Table 2.9. Figure 2.9 provides summary for this subsection. It can be concluded that the meaningfulness of sign is the most investigated aspect and recorded the most mentions in sign usability researches. Learnability of a sign is still unpopular, but it shouldn't be disregarded.

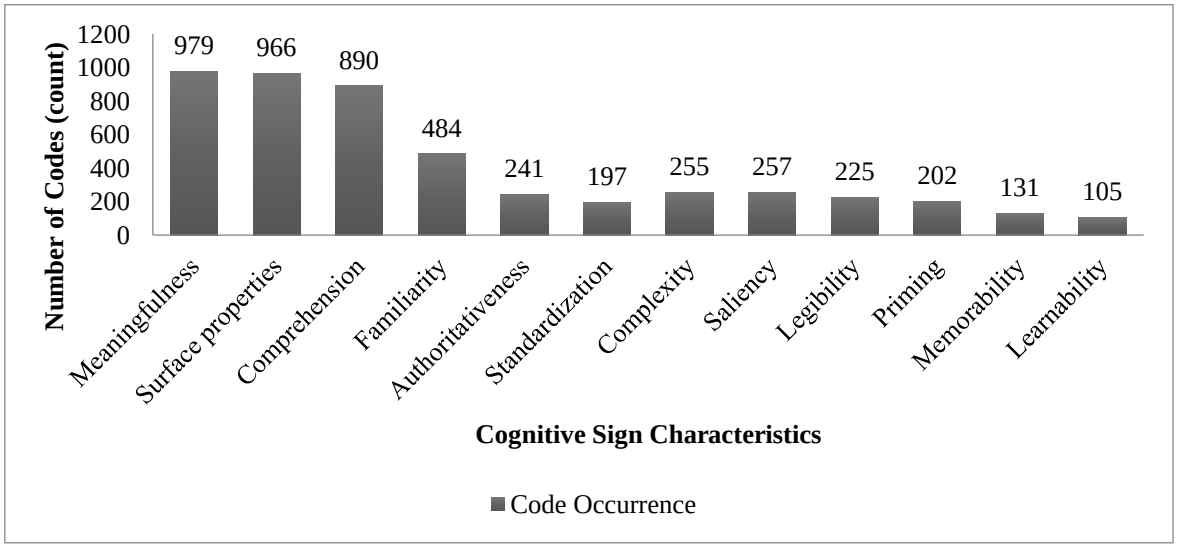


Figure 2.9: Number of code occurrence for CSC

Coding of CSC makes the relationship of between CSC becomes apparent. For example, a study about a Surface Properties (icon design) is based on how the icon gives meaning to the users (Meaningfulness), how easy it is to detect (Saliency), to what extent it can be easily read (Legibility) complexity, to which extent users are familiar with an icon (familiarity) and whether the icon design is simple or complex (Complexity).

2.4.4 CSC code co-occurrence

During coding of CSC, concepts are often found to be overlapping in a way that an event can be associated with more than one CSC. For example, icon comprehension is coded with Comprehension and Surface Properties, therefore Comprehension and Surface Properties are considered co-occur in this phenomenon. Associations between concepts as

noted by previous researches were described in the previous sub-section. This sub-section attempts to identify co-occurrence of concepts with the use of c-coefficient. C-coefficient shows the frequency of quotations in which two codes are co-occurring. By exploring these quotations, it gives sight on the association between CSC. It provides insight into contextual factors and how these factors shape the specific manifestation of a given phenomenon. C-coefficient is computed using ATLAS.ti software. The strength of association of those co-occurrences ranges between 0 and 1 (Contreras, 2011).

Table 2.10 is the extension of code occurrence in Sub-section 2.4.3. From the table, it was noted that Meaningfulness was associated with other concepts the most frequent. Priming, on the other hand, is the least co-occurring concept. Concepts that record a high co-occurrence probability are Complexity and Authoritativeness, as well as Meaningfulness and Comprehension. Co-occurrence of Complexity and Meaningfulness, Familiarity and Meaningfulness are also relatively high.

Table 2.10: C-coefficient Table of CSC

	AUT	CML	CMR	FAM	LRN	LGB	MNG	MMR	PRM	SLN	STD	SPR
AUT	0	0.24	0.05	0.06	0.01	0.03	0.07	0.01	0	0.03	0.03	0.05
CML	0.24	0	0.07	0.08	0.01	0.02	0.11	0	0	0.01	0.01	0.03
CMR	0.05	0.07	0	0.05	0.02	0.03	0.24	0.01	0.01	0.01	0.05	0.06
FAM	0.06	0.08	0.05	0	0.01	0.01	0.11	0	0	0.01	0.07	0.05
LRN	0.01	0.01	0.02	0.01	0	0.01	0.04	0.09	0.01	0	0.01	0.01
LGB	0.03	0.02	0.03	0.01	0.01	0	0.02	0	0.02	0.07	0.01	0.03
MNG	0.07	0.11	0.24	0.11	0.04	0.02	0	0.04	0.05	0.01	0.06	0.09
MMR	0.01	0	0.01	0	0.09	0	0.04	0	0	0.01	0	0
PRM	0	0	0.01	0	0.01	0.02	0.05	0	0	0.01	0.01	0.01
SLN	0.03	0.01	0.01	0.01	0	0.07	0.01	0.01	0.01	0	0.01	0.02
STD	0.03	0.01	0.05	0.07	0.01	0.01	0.06	0	0.01	0.01	0	0.03
SPR	0.05	0.03	0.06	0.05	0.01	0.03	0.09	0	0.01	0.02	0.03	0
TOTAL	0.57	0.59	0.6	0.46	0.21	0.25	0.84	0.17	0.13	0.21	0.28	0.4

* AUT= Authoritativeness; CML= Complexity; CMR= Comprehension; FAM= Familiarity; LRN= Learnability; LGB= Legibility; MNG= Meaningfulness; MMR= Memorability; PRM= Priming; SLN= Saliency; STD= Standardization; SPR= Surface Properties

2.5 Part Four: Usability Instruments

This sub-section describes the composition of generic usability instruments as well as System Usability Scale (SUS), which acts as the foundation of the instrument and reference for the calculation of the composite score. Some usability instruments also constitute for initial item pool for the new measure. Relevant items are subsequently filtered to be adapted.

2.5.1 Analysis of generic usability instruments

This sub-section collects eight of the most popular usability instruments and examines the composition of all instruments. Each item in the instrument was coded a concept that relates to the item. For example, Item "I am satisfied with the system" is assigned to concept "Satisfaction". The analysis includes three items from After Scenario Questionnaire (ASQ), 19 items from Post Study System Usability Questionnaire (PSSUQ), 19 items from Computer System Usability Questionnaire (CSUQ) (Lewis, 1995), 50 items from Software usability measurement inventory (SUMI) (Kirakowski & Corbett, 1993), 10 items from System Usability Scale (SUS) (Brooke, 1996, 2013), one item from Single Ease Question (SEQ) (Sauro & Dumas, 2009), 13 items from Standardized Universal Percentile Rank Questionnaire (SUPR-Q) (Sauro, 2004), and lastly four items from The Usability Metric for User Experience (UMUX) (Finstad, 2010b).

To reduce the items for further analysis, redundancy and relevancy analysis are performed (Ryu & Smith-jackson, 2005). Table 2.11 shows redundant items and concepts in the item pool. Each item was assigned to a concept of which it relates to. As in Table 2.11, there were eight common concepts emerging from generic usability instrument. Three concepts, which are Ease of Use, Satisfaction, and Support were the most recurring concept.

This shows a great emphasis on ease of use, user's satisfaction, and access to adequate support to use the system effectively.

Table 2.11: Redundant Concepts and Number of Item

Concepts	No. of items	Item Number (as listed in Appendix A)
Ease of use	19	1, 4, 15, 16, 23, 34, 35, 65, 80, 85, 88, 89, 90, 94, 102, 103, 104, 117, 118
Efficiency	10	2, 7, 8, 11, 26, 27, 30, 42,70,107
Error	7	12, 13, 31, 32, 45, 79, 119
Interface	4	19, 20, 38, 39
Learnability	8	10, 29, 46, 51, 76, 81, 98, 101
Satisfaction	21	1, 2, 3, 4, 9, 14, 15, 22, 23, 28, 39, 41, 43, 48, 53, 57, 63, 92, 105, 106, 119
Simplicity	4	5, 24, 93, 111
Support	21	3, 6, 14, 15, 17, 18, 25, 33, 34, 35, 36, 44, 49, 54, 59, 64, 74, 84, 95, 108, 112
Non-redundant	34	21, 37, 40, 47, 50, 52, 55, 56, 58, 60, 61, 62, 66, 67, 68, 69, 71, 72, 73, 75, 77, 78, 82, 83, 86, 87, 91, 96, 97, 99, 100, 109, 113, 116

Table 2.12 shows co-occurrence of concepts on item pool. This table shows the count instead of the magnitude of co-occurrence. From Table 2.12, satisfaction often overlaps with other concepts. This includes satisfaction on ease of use, satisfaction on efficiency, satisfaction on the user interface, and satisfaction on the system support.

Table 2.12: Co-occurrence of Concepts

	Ease of use	Efficiency	Error	Interface	Learnability	Satisfaction	Simplicity	Support
Ease of use	0	0	0	0	0	4	0	3
Efficiency	0	0	0	0	0	1	0	0
Error	0	0	0	0	0	0	0	0
Interface	0	0	0	0	0	1	0	0
Learnability	0	0	0	0	0	0	0	0
Satisfaction	4	1	0	1	0	0	0	3
Simplicity	0	0	0	0	0	0	0	0
Support	3	0	0	0	0	3	0	0
TOTAL	7	1	0	1	0	9	0	6

Concept co-occurrence of concept also observed between Ease of Use and Support, whereby the items asked about the ease of use for support function. Existing generic usability instruments are too narrow to be used to represent sign usability. The evident is apparent in Table 2.12, of which shows that these instruments measure concepts that are not relevant to sign usability such as user-system interaction, feedback and system support. This issue will be further explained in Sub-section 4.7.2 in Chapter 4.

Redundancy of items in initial item pool is summarized in Table 2.13. From this summary, it is apparent that usability instruments share similar aspects of usability. This analysis shows the need to generate items and not just adapting items from existing usability instrument.

Table 2.13: Item Pool Redundancy Analysis Summary

Initial item pool	Total redundant items	Items with concurrent concept	Items without concurrent concept	Non-redundant items
119	85	12	73	34

It is useful to investigate the frequency of qualifying words, subject words, and object words in usability instruments to understand the composition of the instruments. The results are presented in Table 2.14 and Figure 2.10 (graphical representation).

Table 2.14: Most used Qualifying, Subject, and Object Words

Types of words	Most used words
Qualifying	Use/Using (42), easy (19), complete (10), quickly (8), find/found (8), feel/felt (8), satisfied (7), able (6), learn (6), can (5), clear (5), help (5), provided (5), scenario (5), usability (5), work (5), like (4), needed (4), comfortable (4), get (3), simple (3), understand (3).
Subject or object	System (53), software (35), information (19), website (13), task(s) (11), messages (6), documentation (4), functions (4), interface (4), problems (4), questionnaire (4), capabilities (3), error (3), screen (3), things (3), time (3) user (3), way (3).

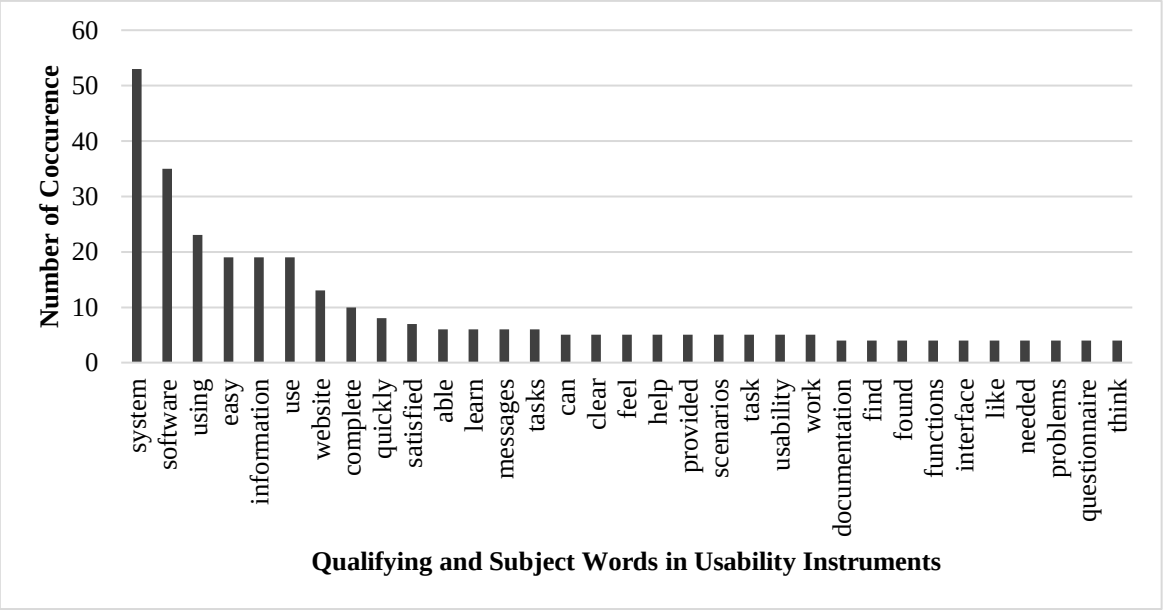


Figure 2.10: Frequency chart of qualifying words and subject words

The most abundant qualifying word was "use", indicating an emphasis on user experience in using and interacting with a system (42 quotations). Ease of use (19 quotations) is deemed another important attribute of a system. Subject words describe the subjects in usability testing instruments. Most usability instrument is found to be specifically about system or software (88 quotations).

This result shows the validity of supported earlier analysis that shows emphasis was placed on ease of use. It became more apparent that the core of usability instrument is the effectiveness of user-system interaction. It is a notable pattern that generic usability instruments are generally task-driven. This is justified by the heavy use of qualifying words such as "use", "easy", "complete", "quickly" and "able".

User's understanding is widely regarded as an important aspect of sign usability. However, the concept of understanding only recorded three quotations, which means it was not given priority by researchers. As previously mentioned, the use of sign does not involve

hands-on user-system interaction. This difference in usability aspects is a problem for researchers who want to adapt these instruments to assess sign usability. It is not recommended to merely rephrase and reword the items to accommodate the instrument for measuring sign usability.

In conclusion, the analysis of generic usability instruments confirmed has provided another justification for the problem statement in Sub-section 1.3.3 regarding the differences of usability aspects across products. This section shows there are many items not suitable for assessing sign usability; for example, items related to system error feedback, help and support system, as well as user interaction with the system interface. However, based on the cognitive sign characteristics as explained in Sub-section 2.4.1, there are several usability aspects that can be considered to assess sign usability. For example, ease of use, efficiency, learnability, and simplicity. The relevant items are retained, and irrelevant items are filtered out. This process of item filtering will be further discussed in Sub-section 3.5.1 and Sub-section 3.5.2.

2.5.2 System Usability Scale (SUS)

System Usability Scale (SUS) is one of the generic usability instruments being analysed in Sub-section 2.5.1. Due to its robustness, flexibility and unique interpretation method, it is used as the base for PSUS. SUS is a popular, remarkably simple, and reliable tool to measure the usability of diverse type of interfaces. SUS is a non-proprietary 10-items questionnaire developed by Brooke in 1996 to assess how people perceive the usability of a system (Brooke, 1996, 2013). For SUS, respondents are required to rate their degree of agreement or disagreement with a set of statements on a 5 or 7 point scale (Brooke, 1996). Typically, the positive and negative statements in SUS were arranged alternately. The

composite score of SUS is then calculated in a unique way, producing a single score which ranges from 0 to 100 (Brooke, 1996). Although principally developed for systems, SUS is flexible enough to be implemented for a wide range of interface technologies (Bangor, Kortum, & Miller, 2008). The use of SUS in usability testing has notably increased. It has been cited in more than 1,200 publications in 2013, not inclusive of unpublished usability tests (Brooke, 2013).

The original SUS was based on ten items, scale size of 5 and the alternation of positive and negative statements. Problems arise when SUS is modified and thus making it difficult to compute. Due to this variability, it can be challenging to transform scores modified SUS into a range from 0 to 100. The developer of SUS was cognizant of the problem, noting that the scoring of SUS can be complicated for some researchers and the use of both positive and negative items will make the scoring even more complicated (Brooke, 2013). SUS researchers failed to cater the need of other researchers that modify SUS depending on their context of use. This gap poses as a solid foundation and motivation to introduce a formula to facilitate the calculation of SUS. This can be particularly useful for researchers or practitioners who adopt a different version of SUS in the future.

SUS was often modified according to the context of use (Finstad, 2006; Ng et al., 2011b; Sauro, 2011). According to a research, approximately 90% of the SUS data reported were collected using modified versions of SUS (Bangor et al., 2008). Although SUS was originally 5-point Likert items, 7-point Likert item version of SUS was also observed (Bangor et al., 2009; Finstad, 2010a). Other than scale size, modified version of SUS may include a different number of items due to deletion and addition of new items.

Modified SUS has been used to assess the usability of safety signs (Ng et al., 2011b, 2011a). SUS modification was done in accordance with the functions of safety signs and context of us. This attempt to use modified SUS to assess sign usability also serves as a justification of why SUS is used as the base (version 0 of the instrument) for PSUS. Modifications of original SUS include word simplification, paraphrasing, and substitution of subject words. The purpose of using modified SUS for assessing sign usability was to supplement usability testing signs (Ng et al., 2011a). This modified SUS is adapted to form version 1 of PSUS. This will be discussed in Section 4.3.

2.5.3 Composite score of SUS

The composite score of SUS is not computed with mean or sum. The original SUS consists of both positive and negative statements and customarily arranged alternately (Ng et al., 2011a). The calculation of standard SUS requires the researcher to identify negative and positive statements because they are treated differently in the calculation. The calculation leads to the overall value of usability which ranges from 0 to 100 (Bangor et al., 2009, 2008; Brooke, 2013; Ng et al., 2011a). The higher the SUS score, the higher is the perceived usability. The range of 0 to 100 obtained corresponds to what is called visual analogue scale (VAS) whereby it defines exclusively the poles of a continuum (0 for lowest, 100 for highest) with no verbal label attached to each value. VAS can be regarded as equidistant and is on the level of an interval scale (Couper, 2006; Reips & Funke, 2008). Therefore, data corresponds with VAS can be analysed using a wide range of statistical procedures (Reips & Funke, 2008).

2.6 Chapter Summary

It has been established that generic usability instruments are too shallow to be adapted to assess sign usability. Therefore, instruments intended to assess sign usability need to be more inclusive to ensure good validity. To achieve this, aspects of sign usability needs to be identified. In 1996, a literature review was conducted to identify sign characteristics that constitute sign usability. This review identified several characteristics, which are concreteness, meaningfulness, complexity, familiarity and semantic distance (McDougall et al., 1999). These signs characteristics have been adapted in researches on sign usability for the past two decades. However, after two decades since the identification of the sign characteristics, dire need to revisit the sign characteristics resurfaced. The emergence of researches in sign usability provides new sign characteristics that need to be included, as they also constitute sign usability.

CHAPTER 3

METHODOLOGY

3.1 Overview

This chapter explains the process of instrument development. As shown in Table 3.1, this chapter is divided into several components: (1) Literature analysis methods, (2) development of items, (3) instrument validation, and (4) description of interpretation methods.

Table 3.1: Overview of Chapter Three

Chapter 3: Methodology			
Part one: Analysis of literature (Section 3.4)	Part two: Item development (Section 3.5)	Part three: Instrument validation (Section 3.6,3.7,3.8)	Part four: Interpretation methods (Section 3.9,3.10)
The process of Collecting literature for analysis and the identification of CSC that acts as justification and base for the new measure	Describe the process of item filtering, involving Redundancy Analysis, Relevancy Analysis, Readability tests and Multiple Responses Analysis	Section 3.6: Justifications for the nature participants and sample size by analysing sample characteristics of past researches Section 3.7: Procedure, instrument, and stimuli for item development and validation Section 3.8: Explains Rasch Analysis and Differential Item Functioning to analyse psychometric properties of PSUS	Section 3.9: Explains the development and testing of composite score formula for use in Perceived Sign Usability Scale (PSUS) Section 3.10: Suggests interpretation methods

Table 3.2 and Table 3.3 provides overview of the source of data, method and techniques as well as the list of tools and software used in each objective.

Table 3.2: Source of Data and Methods for Objective 1

Objectives 1: To develop an instrument to assess sign usability based on its cognitive sign characteristics			
Processes	**Source of data**	**Method/Technique/ Indication**	**Tools/ Software**
a) Developing a framework for the design and development of Perceived Sign Usability Scale. b) Identifying cognitive sign characteristics that constitute road signs usability c) Reviewing existing measures to assess sign usability	Secondary data from literature search	- Documentary analysis - Concept coding - Concurrent table - C-coefficient	- Atlas.ti - Microsoft Excel - Mendeley
d) Developing items for a new measure to assess road users' perception on traffic sign usability.	Secondary data from literature search, Feedbacks from participants	- Documentary analysis - Code co-occurrence - Relevancy & Redundancy analysis - Item review - Readability tests - Dichotomous multiple responses analysis - Cochran's Q test with Pairwise McNemar test	- Atlas.ti - Microsoft Excel - Readable.io - IBM SPSS
e) Establishing the quality of the instrument.	PSUS	- Item difficulty - Rasch and non-Rasch Reliability - Item discriminant - Separation indexes - Rasch fit statistics - Principal component analysis of Rasch residuals - DIF using CMH statistics - Item Characteristics Curve - Option Characteristics Curve - Test Characteristics Curve - Item map with Gaussian smoothing	- JMetrik - Microsoft Excel - IBM SPSS - JASP - Jamovi

Table 3.3: Source of Data and Methods for Objectives 2-4

Objectives	Source of data	Method/Technique/ Indication	Tools/ Software
2. To develop a formula to calculate composite score of the new instrument	Secondary data from literature search	- Modification of SUS formula - Treatment of negative and positive items - Derivation of PSUS constant	- Microsoft Excel - IBM SPSS - GNU PSPP
3. To suggest methods to interpret the data obtained with the instrument	Secondary data from literature search	- Cognitive mapping - Acceptability range - Dichotomous response - Categorization of sign	- Microsoft Excel
4. To assess usability of test signs	PSUS	- Cognitive mapping - Composite score - Classification of sign	- Microsoft Excel - IBM SPSS - Pixelmator - GIMP - Wix.com - Xmind7

3.2 Research Approach

This study conforms to the positivist paradigm that involves the representation of reality with a variable. This study attempts to capture human perception on sign usability in the form of a rating scale instrument. Quantitative analytical techniques were used to develop and analyse the instrument. The goal is to develop an instrument to be adapted in survey research on sign usability.

3.3 Instrument Development Flow

There are three approaches in designing a questionnaire in research. The first is to adapt ready-made instruments developed by other researchers. The second approach is through modification of existing questionnaire to match researcher's interests and context of use. Lastly is to develop new questionnaire if there is no known suitable questionnaire to

assess the constructs (Estabrooks & Wallin, 2004). This research attempts to develop sign usability instrument by utilizing modified items from existing instruments as well as the construction of new items.

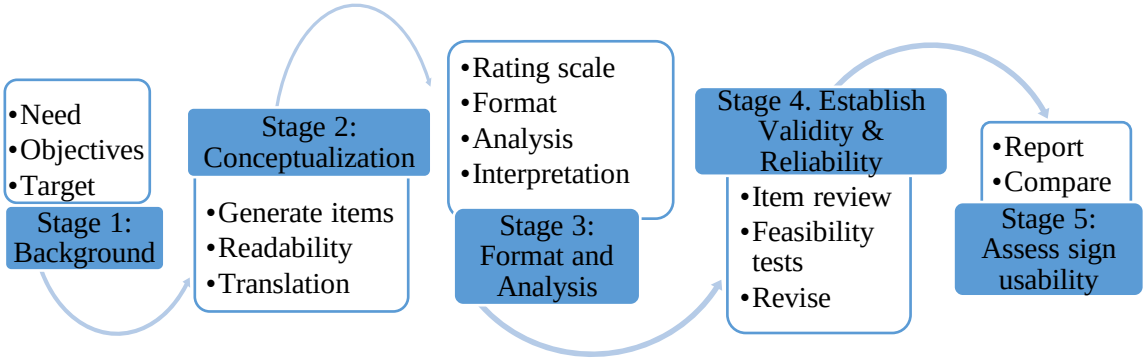


Figure 3.1: Development of instrument (Brancato et al., 2006; Radhakrishna, 2007)

A study explained that the initial stage of recommended practices in designing questionnaire involves a review of the literature, the specification of objectives, and exploration of concepts (Brancato et al., 2006). Once conceptual basis has been enumerated, a draft of the questionnaire can be constructed. The design and development framework of the instrument (Figure 1.3) follows the stages shown in Figure 3.1. The stages described are not strictly one-directional as the stages can be recursive and can co-occur with other stages. Stage recursion is necessary to address emerging issues and creates an instrument with convincing validity and reliability. The iterative process of developing an instrument allows the instrument to improve continually. Fine tuning and testing of the previous stage are crucial to ensure a fine result on current stage.

Stage one defines the purpose of the instrument. In this stage, an extensive literature search is conducted to understand the climate of the research area and identify related instruments. This stage is important because it does not only justify the need for a new

measure, it also helps the researcher to decide whether it is more appropriate to adapt existing instrument as it is, modify existing instrument based on the context of use, or to develop a new instrument. Establishing need of the measure avoids time wasting and unnecessary loss of energy and resources. If the existing instrument is enough to capture needed information, developing a new measure is not necessary. Furthermore, findings in this stage may be instrumental in stage two whereby new measure is built to bridge gaps and limitations of existing instruments. New concepts identified in this stage can be used as a guide to generate items.

After items are generated, the format of the questionnaire as well as data analysis technique must be defined. Stage four involves validity testing and revision of the items. The stages can consist of many steps and processes. Those processes may be repeated if necessary. After validity and reliability of the instrument have been established, advantages over existing measures need to be reported. Subsequently, suggestions for future improvements are also necessary. Lastly, the instrument is used to assess the usability of test signs. Aside from describing administration method, calculation and interpretation method, it is beneficial to administer the instrument in test scenario using test signs. This helps to facilitate and encourage future use by other practitioners.

3.4 Literature Analysis Methods

Reviewing and evaluating documents are conducted in a systematic manner with aim of understanding emerging themes and framework, as well as to give an objective account of what has been done in a subject area (Ryan, Coughlan, & Cronin, 2007). Secondary data are being analysed and interpreted to elicit meaning, gain understanding, as well as to develop empirical knowledge. This process involves various stages and interpretation

methods (Bowen, 2009). This iterative process incorporates elements of content and thematic analysis. Content analysis is the process of organizing information into certain categories depending on the research context. Thematic analysis is a form of pattern recognition to identify notable themes in the data (Fereday & Muir-Cochrane, 2006).

Shortcomings of the traditional narrative review prompt the increasing attempts to use other techniques in reviewing literature. Narrative literature reviews often describe emerging contradictions and similarities and depend heavily on researcher's opinion and judgment on the articles. A more systematic approach in literature review has become increasingly popular, taking advantage and incorporating statistical analysis on data obtained from the documents. This results in a more intuitive and comprehensiveness analysis (Mulrow, 1994; Page, 2008; Petticrew & Roberts, 2006). With this approach, a large quantity of data and document can be analysed and summarized into a compact analytical document (Bolderston, 2008). Systematic review employs a set of criteria to appraise the reliability and validity of the literature (Ham-Baloyi & Jordan, 2016).

The documentary analysis is important in the development of a new measure. Conducting documentary analysis in the beginning of the study is crucial as it allows the researcher to assess the need of developing a new instrument. It is important as the first step of scale development and validation as it provides useful justification whether a new scale is needed for the context of use. If a good measure exists, the cost of developing a new instrument could be avoided. However, development of new measure is crucial if it shows incremental validity, which includes theoretical or empirical advantage over existing measures. A new scale should represent the intended measure more accurately and efficiently. The product is an instrument with enhanced efficiency and ease of use for both

the participant and the researcher (Netemeyer, Bearden, & Subhash, 2003). The needs of a new measure of sign usability are addressed in Chapter 2. Analysis of document is also used to identify CSC in sign usability research.

3.4.1 Collection of sources

The initial stage of electronic search includes defining the type of research and laying out research problems. The scope of research is then defined to match the intended context of use within the field or expertise of the researcher, including locality, language, and other specific search variables. Time and database need to be addressed before starting an electronic document search. Once it has already established, the search for articles may be performed with relevant keywords or phrases (Page, 2008).

The understanding of different types of document is essential in a documentary analysis. There are two types of evidence in a documentary analysis, which are primary and secondary evidence. Typically, greater emphasis is placed on primary evidence instead of secondary evidence when conducting a literature search to minimize potential bias, misinterpretation, and inaccuracies from second-hand reports. Research journals are regarded as the prime example of primary evidence. It is also inclusive of original clinical and experimental researches and statistical reports. On the other hand, secondary evidences include review articles and researcher's report on original researches (Bolderston, 2008; Polgar & Thomas, 2008).

However, academic research should not be restricted to only one publication type. It should include a full range of relevant materials (Ridley, 2008). To minimize bias, various types of documents were searched. It includes books, journal articles, as well as grey literature. The term 'grey literature' refers to materials that are not commercially published

and may have limited access, such as institutional and government report, thesis/dissertation as well as proceedings (Page, 2008). The incorporation of grey literature enhances conclusiveness of the analysis as grey literature may contain valid and crucial information (Page, 2008; Ridley, 2008). For this reason, relevant grey literature was also included. Another reason to include grey literature is the lack of relevant studies on cognitive aspects of sign usability. As for sign usability studies, documents from government bodies are important. An example for important documents includes official traffic sign manual.

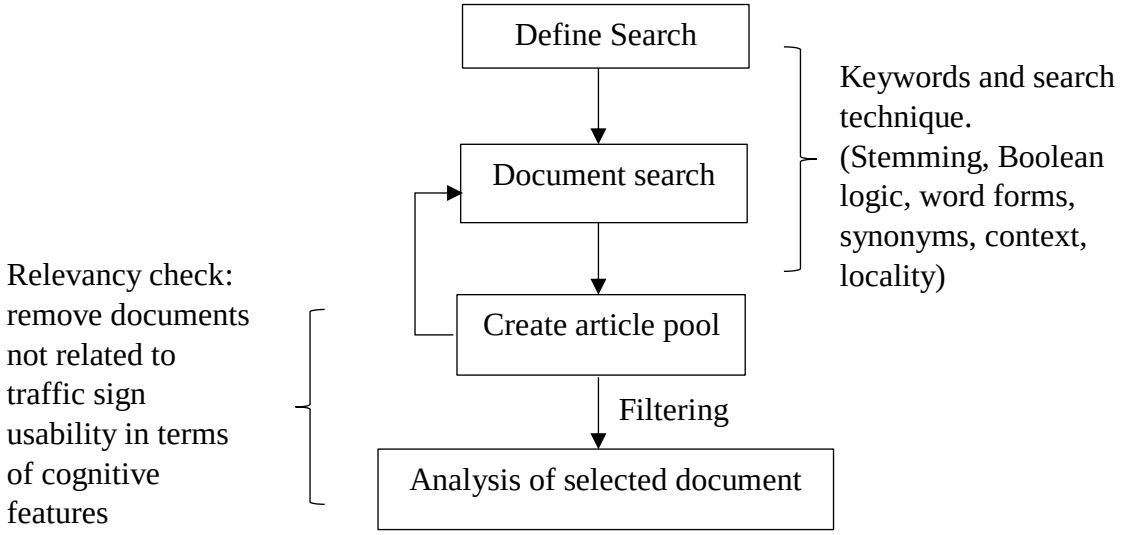


Figure 3.2: Documentary search flow, adapted from Ridley (2008) and Page (2008)

Figure 3.2 shows the process of document search prior to the analysis. This approach is adapted from Ridley (2008) and Page (2008). Different search technique will help to find as many potentially relevant articles as possible to include. A study describes two main search methods, which are protocol driven, and snowballing technique. Protocol-driven technique defines search strategies on the onset whereas snowballing technique emerges as the study unfolded (Greenhalgh & Peacock, 2005).

As in Figure 3.2, document search is facilitated using keywords. Using keywords and search protocol is notable in protocol-driven search method. In this study, main keywords searched include “road signs/traffic sign”, “comprehension/ understanding/ usability”, and “sign feature/characteristics”. For more comprehensive search using keywords, stemming technique using asterisk wildcard is adapted. The stemming technique is useful for searching for variations of a root word (Blachman & Jerry, 2007). Using author’s names of crucial studies to search is a good follow up searching technique (Page, 2008). Optimizing search strategy is crucial in documentary search (Uman, 2011). To facilitate search, search filter can be utilized to extract relevant articles from a database. When searching in a database, a search filter can be a single word or a phrase being entered in the search field, often with the combination of other parameters that can be set (Page, 2008).

Boolean logic was also adapted in advanced search of literature as a search filter. Boolean logic allows more sophisticated and detailed searches using logical operators such as AND, OR and NOT (Page, 2008; Ridley, 2008; Uman, 2011). To further refine the search, Boolean logic can be added to the search field to define the association between keywords and search fields (Page, 2008). It can be adapted in search engines and databases.

Aside from protocol-driven method, snowballing technique is also used to identify related literature. Snowball technique uses relevant studies to drive the study forward. This is one of the main advantages of snowballing as reference lists are easily assessable and examined within a research. When combined with researcher’s context of use, it can be very straightforward to identify other relevant documents (Wohlin, 2014). Snowballing technique identifies influential researchers that are often cited by researchers of the same research domain (Ridley, 2008).

Backward snowballing utilizes the citation in a research by looking at the reference list of a relevant paper to identify other relevant papers. On the other hand, the forward snowballing method identifies new papers based on other papers citing the paper (Wohlin, 2014). This can be done using online tools, namely Google Scholar. Snowballing is an iterative process. New papers identified in the process are in the queue for the next iteration until no highly relevant paper is identified (Jalali & Wohlin, 2012; Wohlin, 2014).

Online Public Access Catalogue (OPAC) of Universiti Malaysia Sarawak was primarily used. The database, digital library and publication search engines used include IEEE/IET Electronic Library (IEL), ACM Digital Library, ScienceDirect, JSTOR, EBSCOhost, Directory of Open Access Journals (DOAJ), Scopus, and Google Scholar. From the database search, ACM digital library and Google Scholar return more related articles. Hard copy materials were also retrieved from the library of Universiti Malaysia Sarawak. All sources were collected over the course of more than 4 years from 2010 to 2015. In searching for literature, alternative terms and concepts that address the same phenomenon or research area are addressed. The following variations are considered:

I. Synonyms and words that have close meaning (e.g. comprehension/understanding; Traffic sign/road sign)
II. Different word forms Singular/plural/ spelling forms (e.g. ‘sign’ ‘signs’ and ‘signage’)
III. Broader/ narrower terms (e.g. ‘Malaysia’, ‘Europe’, ‘Asia’, ‘USA’)
IV. All sources were managed, recorded, annotated and analysed with Mendeley, Microsoft Excel and Word, Foxit PDF as well as Atlas.ti software.

3.4.2 Coding

The coding process involves thorough reading and review of the article to perform coding and defining the category. The aim of this process is to uncover notable themes pertinent to a certain phenomenon being investigated. If a study comprises of various methods other than documentary analysis, predefined codes may be adapted in the coding process (Bowen, 2009; Fereday & Muir-Cochrane, 2006).

The stemming technique is also adapted in identifying sign characteristics in the body of articles. This technique is used to help the researcher to identify concepts that appear in the text. The use of this technique decreases the chance of a concept being overlooked when reading through a document. Every code occurrence in a text is screened to avoid wrong coding. Coding and identifying sign characteristics from the pool of selected articles happen simultaneously. When a new concept is found, it will be used as a code. The coding process is iterative until no new code is found (Thomas & Harden, 2008).

- MEANINGFULNESS:=meaning*|compatibility|concreteness|semantic|recogni*
- COMPREHENSION:=comprehen*|understand*|$MEANINGFULNESS|know|confus*| interpret*
- MEMORABILITY:=memora*|rememb*|recall|forget
- COMPLEXITY:=complex*|simpl*|complicat*|straightforward
- SALIENCY:=salien*|notice*|attent*|detect*|conspicui*|visib*
- LEGIBILITY:=legib*|read*|clear|
- FAMILIARITY:=familiar*|common|unfamiliar*
- STANDARDIZATION:=variation|standard*|consisten*
- AUTHORITATIVENESS:=authoritative*|compl*|obey*|ignore|disobey|violat*

- PRIMING:=priming|decision|action
- LEARNABILITY:=learn*|
- SURFACE PROPERTIES:=color|colour|text|icon|word*|icon|symbol
- USABILITY:=$MEANINGFULNESS|$COMPREHENSION|$MEMORABILITY|$COMPLEXITY|$SALIENCY|$LEGIBILITY|$FAMILIARITY|$STANDARDIZATION|$AUTHORITATIVENESS|$PRIMING|&LEARNABILITY|&SURFACE PROPERTIES

Twelve different codes are used in identifying different concepts. These concepts are extracted from the literature analysis. Sign usability includes all the concepts. After the process of collection, articles are filtered to include only cognitive aspects of sign usability studies, as explained in Sub-section 2.2.3.

3.5 Item Development

This section describes the process of generating initial item pool. Creating item pool is crucial in establishing content validity in the early stage. Content validity is threatened if (a) item reflecting the domain is omitted, (b) irrelevant items not intended to measure the construct is included, (c) composite score represents disproportion weight on domains, and (d) difficulties administering and responding to the instrument (Netemeyer et al., 2003).

Shortlisting items to be tested started with gathering items for initial item pool. Item pool is gathered by compiling items from various usability instruments. Since the focus of generic usability questionnaire is the usability of software products, there is a need to reassess the existing standardized usability instruments if practitioners intend to use it other types of products and interface (Ryu & Smith-jackson, 2005).

Initial item pool of 119 items was gathered from eight generic and popular usability instruments. The composition instruments for initial item pool include three items from After Scenario Questionnaire (ASQ), 20 items from Post Study System Usability Questionnaire (PSSUQ), 19 items from Computer System Usability Questionnaire (CSUQ) (Lewis, 1995), 50 items from Software usability measurement inventory (SUMI) (Kirakowski & Corbett, 1993), 10 items from System Usability Scale (SUS) (Brooke, 1996, 2013), one item from Single Ease Question (SEQ) (Sauro & Dumas, 2009), 13 items from Standardized Universal Percentile Rank Questionnaire (SUPR-Q) (Sauro, 2004), and lastly four items from The Usability Metric for User Experience (UMUX) (Finstad, 2010b). To reduce the items for further analysis, redundancy and relevancy analysis is performed (Ryu & Smith-jackson, 2005).

3.5.1 Redundancy analysis

To eliminate item redundancy based on researcher's context of use, redundancy analysis is conducted (Ryu & Smith-jackson, 2005). Items that have the same meaning from various instruments can cause unwanted redundancy in the item pool. The process of redundancy analysis is facilitated with the use of Atlas.ti software. Duplicated concepts are identified and coded. Repeated items and items with the same concept are grouped together and subsequently reduced. An effective way to identify concept is to identify keywords of an item and compare it with another item in the item pool.

Researchers must be careful of wording redundancy in developing a questionnaire, as some redundancies are unwarranted. In the final scale, redundancy causes a spike in internal consistency without substantial contribution to the content validity (Netemeyer et al., 2003). In redundancy analysis, same worded items and items measuring the same

usability aspect is grouped together. Non-redundant items and redundant items that have been grouped together will be further analysed in relevancy analysis.

Researchers should be careful when eliminating items that share similar concepts. In some cases, where an instrument consists of several constructs, constructs can share common components, thus making content overlap possible. In this scenario, the elimination of overlapping items can compromise content validity, as it may not represent the entire scope of a construct. Researchers should avoid overlapping items unless it is distinct enough, or else it will be prone to misinterpretation and creating misleading infatuation of item correlation (Spector, 1992).

3.5.2 Relevancy analysis

Different instruments have a varying context of use of which researchers must take into account in conducting usability testing. Some questionnaires are specifically designed for administration after completing a task. Some instruments are designed for websites; therefore, some items might not be relevant for another context of use. Relevancy analysis removes items that deemed not suitable for sign usability. The difference between sign usability and usability of other system/product is further explained in Sub-section 4.7.2 in Chapter 4. Items removed contains at least one of the following: help and documentation, completing tasks, productiveness, error message, user interface, system response, and other software specific attributes, as well aesthetics. Understanding item pool is beneficial to understand the characteristics of initial pool and aspects of Usability that is often tested in usability testing.

3.5.3 Readability Tests

Readability is defined as the extent to which a text can be easily comprehended (Bailin & Grafstein, 2001). A Readability formula allows the researcher to evaluate the readability of the text without asking the participants (Heydari & Riazi, 2012). Readability formulas are mathematical equations derived by regression analysis to provide an estimation of reading difficulty based on the text. This is achieved by investigating the linguistic characteristics such complexity of vocabulary, word length and sentence length (Bailin & Grafstein, 2001; Friedman & Hoffman-Goetz, 2006; Laughlin, 1969). In the case of instrument development, it gives insight on the readability of instrument before conducting feasibility test. This is possible because readability measures do not consider participants' prior knowledge to estimate text readability (Bailin & Grafstein, 2001). In post feasibility study, this can act as cross-validation of readability testing involving test participants.

Flesch-Kincaid Reading Ease, Flesch-Kincaid Grade Level, Gunning FOG Index (Frequency of Gobbledygook), SMOG Index (Simple Measure of Gobbledygook), are among the popular readability tests (Svider et al., 2013; Walsh, Volsko, & Faarc, 2008). Flesch Reading Ease Scale ranges from 0 (unreadable) to 100 (very easy to read). Flesch-Kincaid Grade Level is a modified version of the Flesch-Kincaid Reading Ease that generates a grade level rather than a scale (Friedman & Hoffman-Goetz, 2006). Alternatively, Simple Measure of Gobbledygook (SMOG) index, Gunning-Fog Index, and Coleman-Liau Index also provide an estimation of the target grade level based on the US education system. Therefore, grade level should be interpreted with caution in countries having different grade level system. Moreover, using short texts in readability analysis can affect the reliability of some formula. To solve this, multiple readability tests were used and the correlation between tests is reported to ensure trustworthy analysis.

There are several benefits of using multiple readability scores. A study suggested the calculation of readability estimate by computing average of multiple readability scores. This provides the most reliable estimate, especially in computerized readability test (Mailloux, Johnson, Fisher, & Pettibone, 1995). Based on this suggestion, multiple readability tests were implemented in this study. The use of multiple readability indexes is common. A study uses multiple readability tests (Coleman-Liau index, Flesch-Kincaid Grade Level, Flesch Reading Ease, Gunning-Fog Index, and SMOG index) to perform a comparative analysis of readability tests (Agarwal, Chaudhari, Hansberry, Tomei, & Prestigiacomo, 2013). Comparative analysis of readability tests results in a more holistic analysis of readability. In this study, mean for items and different test scores are calculated. This method of implementing readability tests would also provide lower and upper boundary of results, thus supplying the researcher with information on how much confidence can be placed in drawing a conclusion from the scores (Mailloux et al., 1995).

3.5.4 Multiple Responses Analysis (MRA)

When utilizing single-response questions (multiple-choice question, rating scales, etcetera) for a survey, researchers often overlook the possibility that they can be a smorgasbord of answers. Unlike a single response question, a multiple-response question allows the respondent to select zero, one or more answers. This is regarded as the main advantage of MRA as it captures a flexible amount of pre-formatted responses (Santos, 2000). In this study, MRA is implemented to evaluate whether the items can be associated with sign characteristics that have been identified. In using single-response question, problems arise when there is a possibility that an item can be linked to a smorgasbord of concepts. If collectively exhaustive categories are possible, the use of single response question will result in loss of information as it forces the respondents to disregard other

responses by selecting best-of-fit answer (Plašil & Vlach, 2006). The purpose of Multiple Responses Analysis (MRA) is to obtain relative frequency to make a deductive investigation. The diversity of MRA data structure means it is necessary to perform data transformation into a different structure (such as dichotomous dummy coding) to ease tabulating process (Jann, 2005).

In researcher's perspective, every subset of MRA is viewed as one single question that utilizes numeric dichotomous dummy coding (Santos, 2000). In this case, if a concept is selected, it will be numerically coded "1". Meanwhile, if a concept is not selected, it will be coded "0". This procedure can be tedious with a large set of options and items, as researcher needs to examine each option as an individual question. This study comprises of 20 items and 12 options (CSC). The distribution of CSC is examined to identify most significant CSC associated with an item. Consequently, non-significant CSC is dropped while retaining only CSC with relatively high frequencies for further investigation. CSC with relatively isomorphic data tabulation will be further investigated with Cochran's Q test with McNemar as the post-hoc test.

3.5.4.1 Cochran's Q test

Dichotomous Multiple Responses Analysis is implemented prior to this application of Cochran's Q calculation. This analysis is to find characteristics that are viewed to have the possibility of having equal responses proportions. Cochran's Q test with post hoc analysis of multiple McNemar is adapted based on the nature of the data. Data obtained from dichotomous multiple responses data is paired dichotomous data with the independence of observation. Cochran's Q test is also suitable for analysis involving small sample.

Table 3.4 shows the calculation of Cochran's Q on data table of MRA. Suppose k is the number of characteristics tested where k≥2 involving N participants. In this case, k=3 is selected based on the result of Multiple Responses Analysis.

H_0: The proportions of responses are equal across all characteristics.

$H_{0:}\ \pi_1 = \pi_2 = \cdots = \pi_k$

H_1: The proportions of responses are not equal across all characteristics.

$H_{1:}\ \pi_a \neq \pi_b$ for at least one pair π_a and π_b , with $a \neq b$ and $1 \leq a, b \leq k$.

Table 3.4: Calculation on Data Table

Characteristic 1	Characteristic 2	Characteristic 3	$\sum_{j=1}^{k} X_{ij}$	$\left(\sum_{j=1}^{k} X_{ij}\right)^2$
X_{11}	X_{12}	X_{13}	$X_{11}+X_{12}+X_{13}$	$(X_{11}+X_{12}+X_{13})^2$
X_{21}	X_{22}	X_{23}	$X_{21}+X_{22}+X_{23}$	$(X_{21}+X_{22}+X_{23})^2$
X_{31}	X_{32}	X_{33}	$X_{31}+X_{32}+X_{33}$	$(X_{31}+X_{32}+X_{33})^2$
⋮	⋮	⋮	⋮	⋮
$\sum_{i=1}^{N} X_{ij}$	$\sum_{i=1}^{N} X_{ij}$	$\sum_{i=1}^{N} X_{ij}$	Sum: $\sum_{i=1}^{N}\left(\sum_{j=1}^{k} X_{ij}\right)$	Sum: $\sum_{i=1}^{N}\left(\sum_{j=1}^{k} X_{ij}\right)^2$

$$Q = \frac{(k-1)\left[k\left(\sum_{j=1}^{k}\left(\sum_{i=1}^{N} X_{ij}\right)^2\right)-\left[\sum_{i=1}^{N}\left(\sum_{j=1}^{k} X_{ij}\right)\right]^2\right]}{k\left[\sum_{i=1}^{N}\left(\sum_{j=1}^{k} X_{ij}\right)\right]-\left[\sum_{i=1}^{N}\left(\sum_{j=1}^{k} X_{ij}\right)^2\right]} \quad \ldots \textbf{Equation 3.1}$$

Equation 3.1 shows the formula for Cochran's Q test. Table 3.4 helps in the understanding of the computation method of Cochran's Q test. After obtaining Q value, p-value can be calculated as $p = Pr(Q > X^2_{1-\alpha,k-1})$, where $X^2_{1-\alpha,k-1}$ is the value of the (1-α) quantile of the chi-square distribution based on the degree of freedom (df), that is k-1. If

Cochran's test is significant, post hoc analysis will be executed by implementing pairwise McNemar analysis with Bonferroni correction.

3.6 Justifications for the Nature Participants and Sample Size

The knowledge on the nature of participants in sign usability researches is important in instrument design process as the knowledge acts as deciding factor when selecting appropriate participants and the most suitable data collection method. This section investigates sample characteristics and sample size in sign usability researches to determine the suitable sample and data collection approach for this study, in line with good practices in current sign usability researches.

3.6.1 Determining the nature of participants

Most of the studies on sign usability uses educated young road users (University/college student) as shown in Figure 3.3. This figure shows the group of participants used in sign usability researches from 1994 to 2014 as in Section 2.3. From 38 sign usability researches, 55% of studies involved college/university students as test subjects/participants.

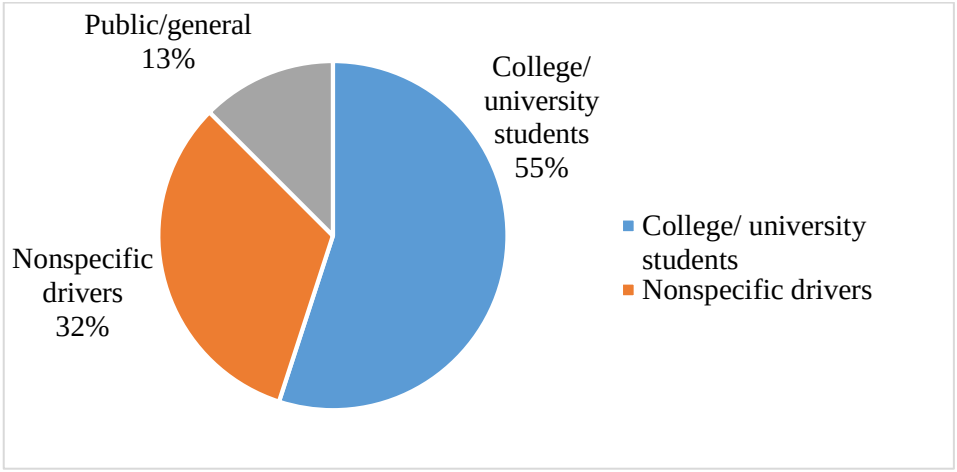


Figure 3.3: Participants in past sign usability studies

The possible reason for prioritizing the inclusion of the participation of students in higher education stemmed from the assumption that young educated road users will perform better in sign comprehension tests. Based on these findings, it can be safely assumed that employing young, educated adults in sign usability studies will produce the upper bound of sign usability scale, that is the overestimation of usability (highest possible). The evident of this is clear, as explained in Sub-section 2.3.3, of which justified that younger road users in higher education scored higher than any other professions in sign comprehension test. In this case, if a sign scored a low usability score, it can be expected that the result will be worse for other samples. This overestimation of sign usability score is desirable when the purpose of the study is to detect problematic signs (signs with usability issues).

The frequent involvement of young drivers in sign usability studies is justified as alarming accident rate from the Ministry of Transport reveals that young drivers are the most vulnerable group (Road Safety Department of Malaysia, 2018). Road Safety Department of Malaysia compiled road accident deaths from 2007 to 2016. Based on the report (updated April 2018), the vulnerable age group is between 16 to 25 years of age as they recorded the highest accident rate, as shown in Figure 3.4. The graph in Figure 3.4 shows a constant pattern of the high-risk age group. Young road users (16 to 25 years old) accounted for 26.9% to 33.6% of total accident-related deaths from 2007 to 2016. The number of road accident-related deaths steadily decreasing for road users who are 30 years and older.

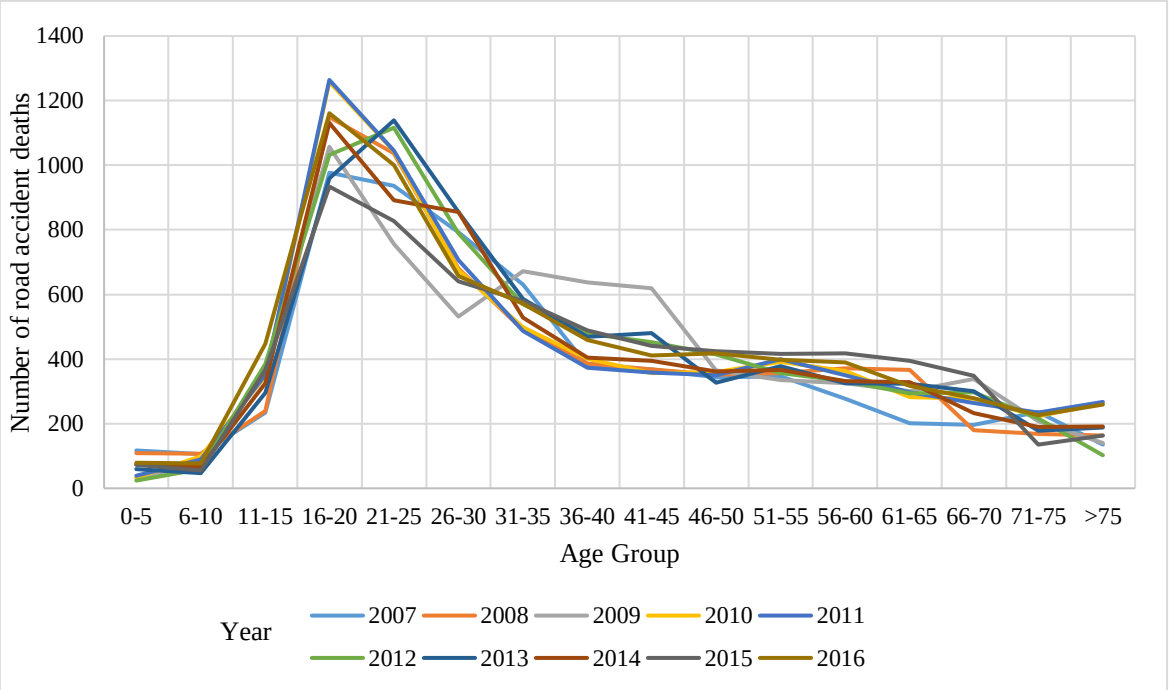


Figure 3.4: Number of road accident deaths by age group

Homogeneous non-probabilistic samples were collected at different stages throughout this study due to several factors. Aside from the identification of aforesaid sample attributes and vulnerability of young drivers, the choice of sampling method is driven by several issues related to the problem in sample randomization. Randomization is infeasible as there is no definite boundary as to what extent road users are defined in term of the use of traffic signs. To determine if randomization is possible, the population of road users must be determined. This is a tricky process as the number of road users is not readily available because the use of traffic signs is not exclusive to users and passengers of motorized vehicles. Non-motorized road users are also vulnerable to risk factors. Road Safety Department of Malaysia recorded that 5277 pedestrians and 1622 cyclists were involved in fatal road accident from 2008 to 2013 (Road Safety Department of Malaysia, 2018). Although in many cases signs are used to warn drivers and motorists about potential dangers and regulations, some signs are intended for road users who are not using motorized

vehicles such as pedestrians and cyclists. Due to this, laying out an accurate list of all traffic signs users is not possible.

Randomized, probability sampling is however recommended for studies beyond instrument development when researcher intent to draw generalizations pertaining to a specific population. Randomization is possible when all members of the population are known, and researcher can impose a limitation on test signs used for the study. An example of this scenario is a study on workers' comprehension of safety sign used in a specified construction area. In this case, there is a non-zero probability for each member of the population to be selected in the study.

In terms of sample size used in sign usability research, it is useful to visualize the trend and pattern. Figure 3.5 shows the distribution of sample size in sign usability studies.

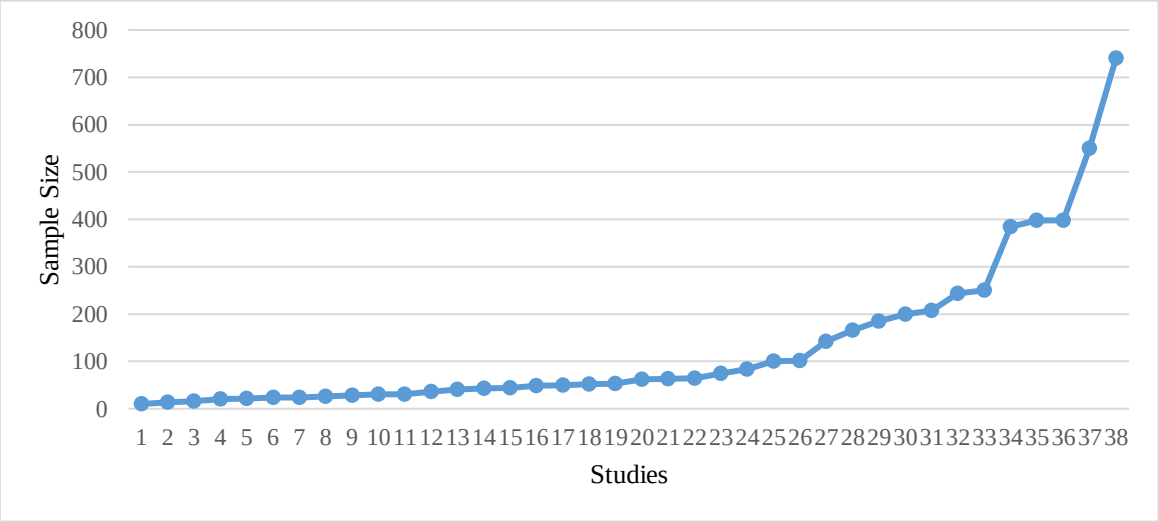


Figure 3.5: Distribution of sample size in sign usability studies

The sample size of sign usability study is relative low, from Figure 3.5, about 78.9% of sign usability studies used fewer than 200 participants. In the case of sign usability based on perception, the sample size requirement is 50 (Talab et al., 2013). Sign usability studies

do not attempt to generalize the result, but rather to identify the effectiveness of signs based on various signs characteristics. Another reason for low sample size is the attempt to control certain unwanted variables and impose delimitations, therefore sign usability studies were often conducted in a controlled laboratory setting (Brémond & Deugnier, 2006; Campbell & Cowley, 2007; Koyuncu & Amado, 2008; Liu, 2005). For example, the use of driving simulator (Al-Gadhi et al., 1994; Castro et al., 2012), and the use of eye-fixation tracking system (Borowsky et al., 2008; Costa et al., 2014).

The goal of usability instrument is to find underlying usability problems. It focuses on the interface or the product itself, and not the user. Usability studies generally have small sample size because the first few participants can easily identify critical design problems. The finding of this analysis is consistent with the previous report on sample size in usability studies. In an analysis of sample size in usability studies, it was discovered that the median number of users per usability test was 10. The number ranges from four to 296 with 64% of the tests employed eight and 12 users. The Analysis shows 80% of usability studies involved no more than 20 participants (Sauro & Lewis, 2009). Another study reviewed dozens of datasets used in usability studies publications, namely in human-computer interaction. It was noted that the average number of participants per study was 32 users with a standard deviation of 29. The minimum number of participants is six and the largest number of participant in a study amounted to 181 (Hornbæk & Law, 2007).

3.6.2 Sample size requirement for Cochran's Q test

The involvement of more than 16 test participants is appropriate for Cochran's Q test (Lewis-Beck, Bryman, & Liao, 2003). In this study, 68 students who have undergone human

factors related courses (Cognition and Design and Cognitive Ergonomics) were asked to relate items with identified cognitive sign characteristics.

3.6.3 Sample size requirement for Rasch

A paper identified sample size of less than 50 as small sample and more than 100 as large sample size (Chen et al., 2014). Rasch analysis gives better analysis with even smaller sample size but of sufficient stability (Hancock & Mueller, 2010; Linacre, 1994). Rasch analysis can handle a sample size of 25-30 to generate a sound 95% Confidence Interval statistic and 50-60 for a 99% Confidence Interval. A sample of 50 well-targeted examinees is conservative for obtaining useful, stable estimates. 30 participants are enough for well-designed pilot studies.

Even if the scale is poorly targeted, the sample size of minimum 144 is sufficient for ± 0.05 logit item calibration with 95 % confidence level (Linacre, 1994). However, Rasch analysis based on a small sample can be misleading as it can contribute to erroneous item-level decisions. Using small sample in Rasch analysis may lead to higher number of misfit items as compared to analysis using large sample size (more than 100 samples) (Chen et al., 2014). This finding is in line with another study that observed more stable estimates in samples that were of size 100 and greater (Revicki, Chen, & Gelhorn, 2013). This finding serves as a guideline for determining the sample size for this study.

For this study, 13 item reviewers were involved in item review, and 68 students who have basics in cognitive ergonomics were involved in Multiple Responses Analysis. For Rasch analysis and assessment of test signs, 205 participants (age ranges from 20-27) were involved. Overall, there were 286 participants involved in this study. For Rasch analysis,

sample size of 250, the person measures calculated would be stable within ± ½ logit with 99% confidence. Even with poor targeting, this sample size would yield reliable results.

3.7 Procedure and Instrumentation

This study can be divided into three parts: (1) development of items, (2) establishment of instrument validation and (3) suggestions for interpretation method. Development of items in this study involves primary and secondary sources. The technique to extract information from secondary sources has already discussed. Procedure and instrumentation section describe procedures and instruments for primary sources. The overall number of participants in this study is 286 participants, inclusive of all stages. For data collection conducted on campus, the faculty was notified before carrying out the collection. In all stages, participants were briefed about the research as well as benefits and risks in involving themselves in the research. All stages generally took no more than 30 minutes of participants' time. Participants are assured that all information gathered will be kept confidential and will not be used for any other purpose other than for this study. Participants were assured that every effort would be taken by the researcher to preserve their confidentiality. There will be no identifying questions and risks are like any experience when disclosing demographic information to others. There was no cost to the participants in involving in the study other than their invaluable time and effort. None of the participants were paid for participating in this study. By completing the tasks, it is an indication that the participants agreed to participate and have understood the risks.

3.7.1 Item feedback and translation of instrument

Before testing the instrument for the subsequent feasibility test, the instrument was subjected to iterative peer reviews process to increase validity and readability of the

instrument. The review process was iteration process until there is an agreement, and no conflicting views reached. A total of 13 participants participated voluntarily in this stage. The review group includes three lecturers in the field of human factors/ergonomics, as well as two experienced teachers who have taught Malaysian University English Test (MUET) or secondary English in Malaysian public schools. The teachers have obtained their Master's degree and pursuing their Ph.D. degree. The rest of the review group participants are three Ph.D. students in the field of Cognitive Sciences, two Master students in the field of ergonomics, one international Master students and two Master students who are native Malay speaker. All participants have a good command of English. Participants were approached individually to avoid the effect of conformity. Every feedback for each item is recorded. This will be further discussed in Chapter 4. Participants were asked to comment on each item to identify confusing items and other issues on items. Participants were also asked to review the translation.

Translation is needed because in an unsupervised web-based survey, unmodified SUS is likely to cause misunderstanding or misinterpretation to the participants, thus creating unwanted bias (Finstad, 2006; Lewis & Sauro, 2009). Therefore, for a more effective execution of the usability testing, translation is needed, especially for a multi racial country such as Malaysia. In a multilingual country, it is important to have a language that is generally understood by the public. In Sarawak (East Malaysia) for example, Mis (2010) estimated that there are about 30-40 languages spoken and the language used by an individual varied across different contexts. However, the national language, standard Malay is well understood by most the population.

3.7.2 Instrument validation

3.7.2.1 Multiple Responses Analysis & Cochran's Q tests

There are several local government agencies dealing with road safety and traffic-related issues: The Road Safety Department, Ministry of Public Works and the Road Transport Department. After visits to those agencies' office in Kuching, Sarawak and Kota Kinabalu, Sabah in 2015, it was informed that there was no expert on traffic sign usability based on cognitive sign characteristics available. This prompted the use of MRA to evaluate item association with cognitive sign characteristics identified in the study (as defined in Sub-section 2.4.1). Participants in this analysis were 68 undergraduate students who have studied human factors courses: Cognition and Design as well as Cognitive Ergonomics. Participants were given a list of items and a list of sign characteristics. They were asked to judge if an item represents the sign characteristics listed. If they think an item does not represent any characteristics, they can choose not to assign any. Participants are free to assign as many sign characteristics to an item if they think the item represents the sign characteristics. This approach is Multiple Responses Analysis whereby participants are free to not to assign any or to assign as many sign characteristics to an item. This process is conducted in a classroom to maintain focus. Data from this are further analysed using Cochran's Q test with pairwise McNemar tests as post hoc analysis. 70 sets of the instrument were distributed, and all 68 students responded, and two sets of instruments returned blank as leftovers.

3.7.2.2 Rasch Analysis

Rasch Analysis was performed to evaluate the quality of the instrument. Overall of 250 questionnaires distributed and 205 completed. Two sets of questionnaires discarded because of too many incomplete responses. The data collection was conducted in a computer

lab because of equipment needed for the study and to allow the participant to respond to the items with less distraction. Driving task and stimuli are complex; therefore, to control nuisance parameters and aggregate behaviour, a controlled laboratory setting is needed (Brémond & Deugnier, 2006). Advantages of using laboratory studies have long been recognized as an ideal method for data collection in signing studies. Laboratory setting is chosen to eliminate problems dealing with environmental variables (weather, ambience/ lighting condition, and traffic conditions) (Katz et al., 2012). This allows more focus and accuracy in investigating cognitive sign characteristics. Laboratory studies also reduce the time required for data collection and provide researchers with additional control over the experiment. With this method, however, it is not possible to obtain random participants. Rasch analysis is further explained in section 3.8.

3.7.2.3 Stimuli

Item review process revealed time required to respond to the questionnaire on three signs was about 20 minutes. Therefore, three types traffic signs (symbol only, text only, and symbol + text sign) were used in this analysis. From the process of item review, it was noted that time taken for the respondents to respond to one sign was around 10 to 15 minutes. Long surveys can cause measurement error in survey research because of survey fatigue (Egleston, Miller, & Meropol, 2011; Porter, Whitcomb, & Weitzer, 2004). To avoid survey fatigue, three test signs (each for one type of sign) were used. Three test signs should take about 30 to 45 minutes to complete.

There are four categories of traffic signs in Malaysia: regulatory (prohibitive and mandatory), warning, temporary, as well as guide signs (Public Works Department Malaysia, 1985b, 1985a). Signs for test signs in this study were selected randomly from a

list of standard regulatory and warning traffic sign listed recognized by the Ministry of Works, Malaysia (Public Works Department Malaysia, 1985b, 1985a). The reason for selecting only regulatory and warning traffic sign is the fact that these signs plays a more prominent role in warning drivers of potential hazards on the road and to prevent road accidents (Makinde & Oluwasegunfunmi, 2014; Mitchell, 2010).

From the list of regulatory and warning sign, each sign is categorized into three types: symbol only, text only, and symbol + text sign. A sign can contain both symbol and text, text only or just symbol/icon without any texts. The perception of text and symbols can be different (as explained in Sub-section 2.2.2: Human perception in relation to traffic signs). Therefore, it is important to categorize signs into separate group based on their content (symbol only, text only, and symbol + text sign). Each regulatory and warning sign is assigned a code by the ministry, and it is used to randomly select a sign from each category to be included in this study.

All traffic signs were printed on paper and projected on a white projector screen. Participants can also view the traffic sign with a web link given to them with the option of zooming in and out on the picture using DELL OptiPlex 745 Desktop display. This redundancy in display reduces the physical and environmental effect on sign usability such as sign placement, weather, day/night condition, maintenance, luminance, and distance. All pictures were edited using Pixelmator and GNU Image Manipulation Program (GIMP) software. For consistency, there were four colours used in the signs. RGB colour code for traffic signs is set as: red RGB (246, 30, 40); black RGB (65,63,64); dark blue RGB (36, 36, 90) and white RGB (250, 251, 246). There might be variation in colour due to monitor, printer and projector setting as well as lighting condition.



Figure 3.6: Test signs (from left: RP.06, RP.10, and RP.05)

Figure 3.6 shows test signs used to evaluate the instrument with Rasch Analysis. From the left, the signs are symbol + text sign (coded as RP.06 by the Ministry of Works), symbol sign (code RP.10), and text only sign (code RP.05). Participants can choose to answer in either English or Malay. They were also allowed to use mixed language to bridge language gap in describing the meaning of signs. Participants were given three tasks to complete individually and without discussion. They were being assured that there were no right or wrong answers. Before starting, participants were required to read and understand the informed consent and commit to volunteer: in the study. Participants were required to (1) state the meaning of traffic sign, (2) answer to the Likert-type items based on their level of agreement/ disagreement, and (3) answer demographic questions. The duration of the data collection was about 20 minutes.

In administrating the instrument, it is assumed that the participants will answer truthfully and were not influenced by peers. In sign comprehension test, there is a possibility that participants will browse the internet to seek for the correct answer or ask others about the meaning of traffic sign. To avoid this problem, the interaction between participants during the test is discouraged. Participants have been briefed that the instrument is about their perception on traffic sign; therefore, there is no correct or wrong answer. They are also

made aware that all responses are anonymous and confidential. There will be no identifying question. It was also assumed that the participants have read all statements and responded accordingly. This can be done by providing ample time and comfort during the administration of the instrument.

3.8 The Rasch Model

Instruments intended to quantify latent traits are usually developed using the basis of Classical Test Theory, which is considered traditional psychometric method. Rasch unidimensional measurement model emerges as a replacement for classical methods, at times also being implemented alongside classical methods. This model can be used to assess scale quality and its psychometric properties (Hendriks, Fyfe, Styles, Rachel Skinner, & Merriman, 2012).

Rasch analysis is a powerful method to evaluate construct validity (Baghaei, 2008). Predominantly used in educational research, Rasch analysis is gaining popularity in social sciences. Rasch measurement model has become increasingly popular in various fields, as it is regarded as a modern and robust approach of designing, analysing, and scoring rating scales. The use of Rasch model has expanded beyond health and educational sciences. It has also been implemented in other fields utilizing rating scales to measure variables such as attributes, aptitude, and ability. This claim is backed by the increase of publications implementing Rasch analysis to assess the quality of instrument (Hagquist & Andrich, 2004). The outcome is favourable, as Rasch Analysis provides valid internally measure that can be extrapolated to the population resulting from the independency on the test sample (Boone, Staver, & Yale, 2014; Granger, 2008; Shohov, 2004; Sick, 2010). To examine an item that falls on the latent trait, it should not matter who answered the item.

To assess instrument dimensionality, two available approaches are often used. Rasch analysis is based on modern test theory approach. Factor analyses, on the other hand, are based on a classical test theory approach. Rasch analysis has several advantages over factor analysis in assessing the quality of an instrument (Apple & Neff, 2012; Belvedere & De Morton, 2010). The limitations of classical methods prompted researchers to develop new and modern psychometric methods. Rasch measurement is a progression of Classical Test Theory (CTT) with improved underlying theory underpinning the treatment of rating scales. Therefore, regarded as one of the new methods to evaluate and analyse rating scales. This new measure tackles flaw and attempts to improve CTT's underlying theory, of which has a weakness in delineating a conclusion about test performance and measurement of the participants (Hobart & Cano, 2009).

Rasch models are a class of probabilistic model that explains the response of a person to a set of items (Ayele, Zewotir, & Mwambi, 2014). It models the relationship between person ability, item difficulty and the probability of a response (Teo, 2011). In Rasch theory, it is assumed that items have a different level of difficulty. Traits of participants influence the probability of responding in a certain direction for a particular item. (Curtin, Browne, Staines, & Perry, 2015). The measurement of the latent construct is based on this probability (Apple & Neff, 2012). The implementation of Rasch measurement model provides a solution to the issue of instrument validity as it provides robust statistics and probing analysis on the quality of an instrument (Bond & Fox, 2015).

The distinction between traditional statistical models and Rasch model is apparent, in the sense that unlike traditional model, Rasch model is based on pre-constructed theory prior to the data. The data are compared with the model, to identify misfit between the data and

the model. This should be done without the inclusion of additional parameters. On contrary, traditional statistical models are explanatory in nature, as it is used to describe or explain the data (Hagquist & Andrich, 2004). Rasch analysis uses modern psychometric approaches that convert categorical data into interval-based measures (logit) by utilizing logarithm function (Lamoureux, Pallant, Pesudovs, Hassell, & Keeffe, 2006). Rasch converts ordinal scale into using logarithm functions (logit). It provides estimates of person and item estimates as well as item difficulty based on the person ability. However, this is not available in SEM. SEM assumes that the ordinal Likert-scale categories are divided into equal parts that align with its strength (Koh, Yen, & Lan, 2012).

Data obtained from questionnaires are often regarded as interval and hence analysed as if they were continuous data. However, the true nature of such data is ordinal. There are controversies in interpreting Likert scales, on whether they can be treated as interval data and subsequently used in parametric tests (Sullivan & Artino, 2013). Likert scales have a rank order; therefore, they should be treated as ordinal data (Jamieson, 2004). However, data with an ordinal level of measurement can only be correctly described with mode, median, and percentage, although in most cases, a more detailed analysis is desired (Aziz, Masodi, & Zaharim, 2013). Mistreatment of Likert-type categorical data can result in erroneous analyses (Bond & Fox, 2015).

Different items measuring the same construct have a different level of endorsement, but items have been treated as though they contribute equally to the measurement. This has been observed in the implementation of exploratory factor analysis and correlational tests for Likert-type items. Unlike traditional method, Rasch analysis provides reliability analysis in terms of measurement, that indicates the quality of the items to measure a sample

population, as well as person reliability that represents response consistency (Apple, 2013; Apple & Neff, 2012). This is viewed as a way out of the limitation of using a conventional method that relies heavily on Cronbach's alpha to show the quality of items representing a conceptualized construct (Smith, 2001).

Rasch analysis can be used to evaluate several quality aspects of a measurement, including rating scale functioning, internal scale validity, person-response validity, person-separation reliability, and differential item functioning (DIF) (Lerdal et al., 2016). This shows Rasch analysis is a comprehensive technique to assess the quality of a measurement.

This study uses five-scale Likert-type response anchors (strongly disagree, disagree, neutral, disagree & strongly disagree). Polytomous items such as Likert-type items require a special case of Rasch model. The need to cater for polytomous items gives rise to various extensions of Rasch models such as Rating Scale Model and Partial Credit Model. This study adopts Rating Scale Model because all items share the same response anchors. Andrich's Rating Scale Model is a special instance of Partial Credit Model that applies a common threshold across all items (Andrich, 1978). Using Andrich's Rating Scale Model is appropriate for Likert-type items as it will improve the estimation of thresholds (Meyer, 2014).

The probability of the response k is for Likert-type item j with h=1, 2, 3, …, m_j ordinal categories, it can be calculated as in Equation 3.2:

$$P_{jk}(\theta) = \frac{exp\sum_{v=1}^{k}(\theta - b_j - \tau_v)}{\sum_{h=1}^{m_j} exp\sum_{v=1}^{k}(\theta - b_j - \tau_v)} \quad \ldots \textbf{Equation 3.2}$$

Where θ = person ability; b_j = item difficulty; τ_v = threshold parameter

The difference between Partial Credit Model and Rating Scale Model lies in the step parameter. The step parameter for the Partial Credit model is $b_{jv} = b_j + \tau_{jv}$ whereas for Rating Scale Model, there is a missing *j* subscript in the threshold parameter ($b_{jv} = b_j + \tau_v$) (Meyer, 2014). The *j* subscripts in the threshold parameter of Partial Credit Model indicates a different threshold parameter is used for each item.

3.8.1 Rasch fit statistics

Fit statistics in regression analysis have a different function as compared to fit statistics in Rasch analysis. Fit statistics in regression analysis are used to identify a model that considered fitting in the data. On the other hand, Rasch analysis is a predetermined model of which the fit statistics act as measurement quality determinant to filter items that failed to adhere to Rasch model specifications (Smith, 1996).

Fit statistics help to determine whether the item estimations may be held as a meaningful quantitative summary of the observations (whether each item contributes to the measurement of only one construct (Bond & Fox, 2015). Although poor fit does not mean that the Rasch measures (parameter estimates) are not additive (appropriate). There are two mean square fit statistics used in Rasch analysis, which are Weighted (infit) Mean Squares (WMS) and Unweighted (outfit) Mean Squares (UMS) that range from zero to infinity and having an expected value of 1 (Bond & Fox, 2015; Meyer, 2014). The Rasch model forces its estimates to be additive. Therefore, the value of WMS value more than 2.0 suggests a deviation from unidimensionality in the data, not in the measures. This value indicates unpredictability and model underfit. Items with a high value of WMS are detected first and removed from the subsequent iteration of calculation before eliminating items with low WMS value (Wright & Linacre, 1994). This iterative process of removing underfit and

overfit items continues until there is no deviation detection. At the end of the analysis, all remaining items yield WMS value close to 1.0 (Khan et al., 2013).

Weighted (infit) Mean Squares fit statistic is a inlier-sensitive fit statistic, as it is affected by erroneous data in the interior of the data distribution of which is sensitive to the responses pattern of items answered by the participant. Unweighted (outfit) Mean Squares fit statistic is sensitive to outliers, in this case, referring to unexpected item responses to items that are very difficult or very easy based on the person ability. Generally, this problem is easier to detect and to counter. The degree of randomness in the measurement is shown with mean-square fit statistics, also with the expected value of 1.0. A low value indicates excessively predictable observation (data overfit) and a high value indicates unpredictable observation (data underfit) (Linacre, 2002). Since fit statistics of Rasch have the expected value of 1.0, fit statistic of 1.30 indicates 30% more variation as compared to predicted Rasch model, thus deviates from the unidimensional construct. Similarly, a fit statistic value of 0.70 would indicate 30% less variation (considered redundant) as predicted value of the Rasch model. (Ohrbach, 2008; Smith, Rush, Fallowfield, Velikova, & Sharpe, 2008). Misfitting of item and person responses may be attributable to item bias (Wolfe & Smith, 2007a).

There are no absolute rules in defining a reasonable range for fit statistics as it varies based on the context of use and type of tests. Weighted (infit) Mean Squares and unweighted (outfit) Mean Squares values that range from 0.6 to 1.4 is ideal for rating scales although the range of 0.5 to 1.5 is already considered productive for measurement (Bond & Fox, 2015; Linacre, 2002; Wright & Linacre, 1994). Items with poor fit statistics should be considered for removal from the instrument because not only it deviates from the assumption of unidimensionality but also not providing information beyond what was given by the rest of

items on the scale (Tang, Wong, Chiu, Lum, & Ungvari, 2005). Aside from WMS and UMS, standardized fit statistics are also used to assess the instrument. It indicates the extent to which the data fit the model. It has the expected value of 0.0 and acceptable range of -1.9 to 1.9. Items with standardized fit statistics of -1.9 to 1.9 have reasonable predictability (Linacre, 2002).

Rasch fit statistics are used to detect whether the data fits the unidimensionality of Rasch model. Misfit items are explaining other underlying construct and violate the unidimensionality assumption. If all items fall within the ideal range depending on context of use, they are considered unidimensional (Anshel, Weatherby, Kang, & Watson, 2009; Baghaei, 2008; Hibbard, Stockard, Mahoney, & Tusler, 2004; Kendel et al., 2010; Linacre, 1998; Miller, Webster, & MacIntosh, 2002; Tang et al., 2005; Tang, Wong, Chiu, & Ungvari, 2007; Teo, 2011). Misfit items are identified and excluded for the next iteration of fit statistics calculation. A misfit item is an indication of construct-irrelevant variance (Baghaei, 2008). This process identifies and removes items to improve unidimensionality of the instrument. There is no absolute rule to eliminate misfit items. It is beneficial to examine why an item is a misfit to the model. The researcher has to make a decision whether to modify or to exclude the misfit item (Kendel et al., 2010; R. M. Smith, 1996). In some cases, misfit items can result from incorrect data entry and other mistakes; therefore, the researcher should investigate the possibility of errors before eliminating a misfit item.

3.8.2 Assessment of reliability and separation index

Reliability is not regarded as an indicator for instrument quality (the extent to which an instrument is a good measure for the intended construct), but an indicator of relative reproducibility (the extent to which an instrument can be repeated and producing dependable

results). Since total variance in the measure comprises of true variance and error variance, the reliability of the measure describes the proportion of true variance in the observed variance (Clauser & Linacre, 1999; Linacre, 1997; Wright, 1996). Alternately, it refers the absence of error variance (mean-square error) by dividing true (observed) variance with the adjusted variance.

Traditionally, researchers adapted Cronbach's alpha as the indicator for instrument reliability. The use of Cronbach's alpha as an indicator for internal consistency if often criticized because increasing the number of items can increase alpha coefficient. Besides, the inclusion of considerably many low inter-correlation items can also produce a relatively high value of alpha coefficient (Spector, 1992). It was also noted that in some cases conventional True-Score estimators such as KR-20 and Cronbach's alpha coefficient do not summarize the effectiveness of measurement (Linacre, 1996). Using Rasch measurement model provides a solution for aforesaid controversies as it enables researchers to investigate the reliability of their instrument in a more comprehensive way.

Rasch measurement model provides more reliability indicators as compared to Cronbach's alpha (Aziz et al., 2013). Rasch Person reliability is analogous to Cronbach's alpha coefficient in classical reliability assessment (Eklund, Erlandsson, Persson, & Hagell, 2009; Kook & Varni, 2008; Lo et al., 2015; Petrillo, Cano, McLeod, & Coon, 2015; Soh, Barker, Morello, Dalton, & Brand, 2016). Rasch Person reliability is based on the person location estimates along the measurement continuum (McCreary et al., 2013). It yields a more accurate estimate of internal consistency in skewed distribution as compared to Cronbach's alpha (Kurtaiş et al., 2011). On the other hand, Rasch item reliability does not have a traditional equivalent. Wide item difficulty range and large sample size contribute to

high item reliability. Thus, Rasch item reliability is used to evaluate whether the sample size is sufficient to locate the items along the latent variable line with precision (Linacre, 2006). All reliability value including Person and Item Reliability should be more than 0.8 to be considered very good. A value of less than 0.6 is deemed not acceptable. A high value indicates that the instrument is consistent with the unidimensionality assumption of Rasch model (Ohrbach, 2008).

Reliability coefficient is restricted to the range of 0 to 1. Therefore, it is useful to express reliability coefficient as a separation coefficient that ranges from 0 to infinity by dividing true (adjusted) standard deviation by error standard deviation (root mean-square error) (Wright, 1996). Separation index describes how many different groups can be distinguished by the measurement (Gothwal, Wright, Lamoureux, & Pesudovs, 2010; Wilson, 2005; Wright, 1996). These Rasch separation indexes (item/person) serve as analytical indices to monitor a variable's continuing utility and thus evaluate whether the development of a variable is successful (Stone & Wright, 1999). The accepted criterion for the separation indices is 3.0 (Bond & Fox, 2015). However, the cut-off point for separation index is not absolute. For example, another guide suggested that person separation index should be more than 2 and item separation should be more than 3 (Linacre, 2012). Meanwhile, another report describes the range for person separation index as 1.5- acceptable, 2.0- good, and 3.0- excellent (Duncan, Bode, Lai, & Perera, 2003). A high value for Separation Index (person or item) indicates that persons (or items) who are estimated with high scoring indeed have a higher probability to have higher score compared to persons (or items) with estimated low scores.

3.8.3 Unidimensionality and PCAR

Construct validity of an instrument can be detected with unidimensionality (Ariffin et al., 2010). Unidimensionality is desirable because it provides an evidence that the instrument is measuring a single underlying construct (Vaughan, Mulcahy, & McLaughlin, 2014). A common mistake committed by researchers when assessing unidimensionality is that mathematical indicators of unidimensionality are regarded as absolute and must be strictly adhered to, resulting in loss of content validity with the removal of crucial items that record slight deviation from acceptability range. This aggressive approach and strict reliance on indicators are pernicious because there is no absolute measure of unidimensionality. In fact, unidimensionality is an argument and the decision whether an instrument is unidimensional should ultimately derive beyond the data and driven by theoretical basis (Hagell, 2014; Hobart & Cano, 2009).

Rasch analysis can help to assess the instrument by analysing the data based on the assumption that it is unidimensional. It uses fit statistics to describe how well the data fit the assumption of unidimensionality of Rasch model (Linacre, 2011). Furthermore, Principal Components Analysis of Rasch Residuals (PCAR) is utilized to further investigate the degree of unidimensionality. In PCAR, the correlation between the items and the Rasch residuals is examined. Failure to identify a notable pattern in residual is ideal after extracting Rasch factor. There should be no significant difference between these two polarities of residuals.

Unidimensionality of a measure can be detected using various methods. Methods for assessing unidimensionality from PCAR include whether person estimates derived from subsets of items are different. This is conducted by first categorizing items into subsets based

on the positive and negative residuals of factor loadings on the first Rasch factor, then finding the difference of the person estimate between subsets using paired samples t-test (Ahmad, Oranye, & Danilov, 2017; Dencker, Sunnerhagen, Taft, & Lundgren-Nilsson, 2015; Mulcahy & Vaughan, 2015; Pallant, Miller, & Tennant, 2006; E. V Smith, 2002; Vaughan et al., 2014; Zoanetti, Griffin, Beaves, & Wallace, 2009). If there is no difference found between subsets person estimate, this indicates the measure is unidimensional (Pallant et al., 2006). This approach is a more sensitive test of unidimensionality using PCAR residuals (Dencker et al., 2015).

In another test of unidimensionality using PCA of Rasch residuals, the eigenvalue of first contrast is examined. The eigenvalue of less than 3 is considered good. Eigenvalue more than 3 indicate a violation of unidimensionality (Eakman, 2012; Galli, Chiesi, & Primi, 2008; Linacre, 2012). Some authors also note that the eigenvalue for the first component should be three times larger than the second component (Khan et al., 2013). The abundance of method assessing unidimensionality using Rasch analysis gives more flexibility and enable the research to perform a cross check to gain more confidence about their measurement.

The combination of fit statistics and PCAR, Rasch analysis provide more support to claim generalizability and unidimensionality of a measure (Wolfe & Smith, 2007b). Due to differences in calculations and underpinning theories, different statistical approaches for instrument validation may yield a different result (Lamoureux et al., 2006); therefore, instrument developed using classical test theories may give a different result compared to the instrument developed by using modern test theories.

3.8.4 Differential Item Functioning

Detection of Differential Item Functioning (DIF) refers to the analysis to examine if an item tends to yield different result based on certain characteristics or group of participants. DIF is a form of measurement bias that may occur for certain items (Bond & Fox, 2015; Dencker et al., 2015; Linacre, 2012). In this study, Mantel-Haenszel (MH) procedure is used to detect DIF in the items. The Mantel-Haenszel procedure is a widely used statistic and is considered industry standard in computing DIF (Badia, Prieto, & Linacre, 2002; Karami, 2012). MH is a preferable statistic to compute DIF because it is non-parametric and relies less on model assumptions (Wiberg, 2007, p. 30).

DIF of a polytomous item is classified into one of three categories: AA, BB, or CC depending on the degree to observed DIF (negligible, medium, or large). Plus (+) or minus (-) sign is added to the classification to indicate whether an item favours focal (+) or reference (-) group (Cunningham et al., 2011; Meyer, 2014). An item will be assigned category AA if Mantel-Haenszel's chi-square is not statistically significant or the effect size has the absolute value of less than or equal to 0.17. Category BB is assigned if Mantel's chi-square is statistically significant. The absolute value of effect size also must be less than or equal to 0.25. Finally, if Mantel-Haenszel's chi-square is statistically significant and the effect size is big (more than 0.25), CC classification will be assigned to the item. Items in this category are not desirable because there is a possible bias. However, items that show signs of bias are not immediately deleted. Before deletion, they should be carefully examined and reviewed to detect plausible reason that item may be biased towards certain groups (Eklund et al., 2009; Hagquist & Andrich, 2004; Karami, 2012; Khan et al., 2013).

3.8.5 Rasch item-person mapping

Rasch item-person mapping helps to visualize which item is most difficult (or easy) for the participants to endorse. Items are placed on a hierarchical logit scale based on the item difficulty from the easiest to the hardest item. Value 0 is set as the average ability to endorse an item. Difficult items are located at the lower half the map (below item mean) and vice versa (Blackman, Bradman, & Graetz, 2015). Rasch item-person can be used to identify if there is Construct under-representation by examining at the gaps on the map (Baghaei, 2008).

3.9 Development of Composite Score Formula for PSUS

When the reliability and validity of the instrument has been established, the next challenge is to calculate the composite score of the instrument to draw conclusion about sign usability. Resulting from the controversial composite score methods as explained in the problem statement (Sub-section 1.3.5), this study attempts to find a more appropriate and intuitive way to compute the composite score of the instrument. The concept of the composite score calculation is explained in Sub-section 3.9.1. PSUS formula can be found in Sub-section 3.9.2. Lastly, Sub-section 3.9.3 details the testing for the formula using simulated data.

3.9.1 Concept and mechanism

Abuse of Likert Scales, which centred on the computation of composite score, was discussed in Sub-section 1.3.5. This study attempts to develop a universal mathematical formula to calculate the composite score of Likert-type items of varying scale size, negative/positive polarity of statements as well as the number of items. This formula will be used to calculate the composite score of Perceived Sign Usability Scale (PSUS). The aim of

this formula is to convert responses on the instrument into a 0-100 scale. By using proposed formula, regardless of the number of items, scale size or point range, this formula will always yield 0-100 scale. Scoring Likert scale responses into 0-100 has already being practiced by researchers to present their findings, as it is a better way to handle categorical data obtained using Likert Scale.

The range of 0 to 100 obtained corresponds to what is called Visual Analog Scale (VAS) whereby it defines exclusively the poles of a continuum (0 for lowest, 100 for highest) with no verbal label attached to each value. VAS data are equidistant and is on the level of an interval scale (Couper, 2006; Reips & Funke, 2008). Therefore, data corresponds with VAS can be analysed using a wide range of statistical procedures (Reips & Funke, 2008). Furthermore, by having a scale of 0 to 100, not only it produces solid consistency across instruments, it also allows the result to be effortlessly comprehended in a short space of time, especially by those who have no prior knowledge and experience in usability testing (Bangor et al., 2008; Brooke, 2013).

In the field of ergonomics, prominent use of this composite score calculation method is demonstrated in the computation of System Usability Scale (SUS). However, there was no formula provided by Brooke (1996, 2013). The developer provided a brief description on scaling procedure based on the 10 items and scale size of 5. The original SUS consists of both positive and negative statements and customarily arranged in an alternate manner (Ng et al., 2011a). The calculation of standard SUS requires the researcher to identify negative and positive statements because they are treated differently in the calculation. The calculation leads to the overall value of usability which ranges from 0 to 100 (Bangor et al.,

2009, 2008; Brooke, 2013; Ng et al., 2011a). The higher the SUS score, the higher is the perceived usability.

3.9.2 Development

As with any instrument, negative and positive worded items need to be identified. It is a common practice to identify negative coded items and reverse coding them (Pallant, 2010). The formula developed in this study is more complex as compared to the calculation of SUS because it considers the variation in the scale in term of scale size (response anchors/options), number of items, and polarity of statements (positive or negatively worded). The increased complexity is the result of added robustness of the formula to cater the need of modified instruments. In this study, instead of reverse coding, the highest scale position is subtracted with the responses on negative-worded items, then added together. This operation is represented by $\sum_{k\in N}(s_h - x)$. For positive-worded items, the responses are subtracted with the lowest value on the scale and then summed. This action is represented by $\sum_{k\in P}(x - s_l)$. If the participants gave high value for negative-worded items and vice versa (indicating disagreement), the formula will yield low scale value. The summation of these two operations is multiplied by a coefficient that can be calculated with the formula $100/(s\_h - s\_l)i$.

$$\text{Scale} = \left[\sum_{k\in N}(s_h - x) + \sum_{k\in P}(x - s_l)\right]\left(\frac{100}{(s_h - s_l)i}\right) \quad \ldots \textbf{Equation 3.3}$$

Where
k = Item
x = Score position in each item.
i = Number of items.
s_h= Highest scale position
s_l = Lowest scale position
N = Negative worded items.
P = Positive worded items

Given that Likert scale is adapted in an instrument with the smallest scale position is the negative response (e.g. extremely disagree) and the highest scale position is the positive response (e.g. extremely agree). By using the proposed formula in Equation 3.3, regardless of the number of items, scale size or point range, this formula will always yield 0-100 scale. Table 3.5 shows the value of $100/(s_h - s_l)i$ for different scale size and number of items. The value of $(s_h - s_l)i$ can be calculated by calculating the difference between highest and lowest possible scale position and multiply the value by the number of items. For instruments other than the example shown below, the value of $100/(s_h - s_l)i$ can be calculated manually. Refer to Appendix F for a more comprehensive table for $100/(s_h - s_l)i$ value based on the number of items and scale size. A short example is as Table 3.5.

Table 3.5: Value for $100/(s_h - s_l)i$ based on No. of Items and Scale Size

No. of Item	Scale Size								
	2	3	4	5	6	7	8	9	10
1	100.0000	50.0000	33.3333	25.0000	20.0000	16.6667	14.2857	12.5000	11.1111
2	50.0000	25.0000	16.6667	12.5000	10.0000	8.3333	7.1429	6.2500	5.5556
3	33.3333	16.6667	11.1111	8.3333	6.6667	5.5556	4.7619	4.1667	3.7037
4	25.0000	12.5000	8.3333	6.2500	5.0000	4.1667	3.5714	3.1250	2.7778
5	20.0000	10.0000	6.6667	5.0000	4.0000	3.3333	2.8571	2.5000	2.2222
6	16.6667	8.3333	5.5556	4.1667	3.3333	2.7778	2.3810	2.0833	1.8519
7	14.2857	7.1429	4.7619	3.5714	2.8571	2.3810	2.0408	1.7857	1.5873
8	12.5000	6.2500	4.1667	3.1250	2.5000	2.0833	1.7857	1.5625	1.3889
9	11.1111	5.5556	3.7037	2.7778	2.2222	1.8519	1.5873	1.3889	1.2346
10	10.0000	5.0000	3.3333	2.5000	2.0000	1.6667	1.4286	1.2500	1.1111
11	9.0909	4.5455	3.0303	2.2727	1.8182	1.5152	1.2987	1.1364	1.0101
12	8.3333	4.1667	2.7778	2.0833	1.6667	1.3889	1.1905	1.0417	0.9259
13	7.6923	3.8462	2.5641	1.9231	1.5385	1.2821	1.0989	0.9615	0.8547
14	7.1429	3.5714	2.3810	1.7857	1.4286	1.1905	1.0204	0.8929	0.7937
15	6.6667	3.3333	2.2222	1.6667	1.3333	1.1111	0.9524	0.8333	0.7407
16	6.2500	3.1250	2.0833	1.5625	1.2500	1.0417	0.8929	0.7813	0.6944
17	5.8824	2.9412	1.9608	1.4706	1.1765	0.9804	0.8403	0.7353	0.6536
18	5.5556	2.7778	1.8519	1.3889	1.1111	0.9259	0.7937	0.6944	0.6173
19	5.2632	2.6316	1.7544	1.3158	1.0526	0.8772	0.7519	0.6579	0.5848
20	5.0000	2.5000	1.6667	1.2500	1.0000	0.8333	0.7143	0.6250	0.5556

The calculation can be performed in statistical software. For example, for an instrument that has nine items (four positive, five negative) with 5 scale size ranges from 1-5, the calculation can be performed with the command: SUM ((5-N1), (5-N2), (5-N3), (5-N4), (5-N5), (P1-1), (P2-1), (P3-1), (P4-1))*2.7778. The value $100/(s_h - s_l)i$ can be calculated or extracted from the table above. In this case, by referring to the Table 3.5, the value of $100/(s_h - s_l)i$ for an instrument that has nine items, and 7 scale size is 2.7778. Scale size refers to the number of response anchors that can be selected. It can also be calculated with the formula $(s_h - s_l) + 1$. N1 refers to the first negatively worded items in the instrument and so forth. Similarly, P1 refers to the first positively worded items and etcetera. The command was tested in two statistical packages: IBM SPSS v22 and GNU PSPPIRE v0.10.

Given that an instrument has 5-point anchors with 1- Strongly Disagree, 2- Disagree, 3- Neutral, 4- Agree, 5- Strongly Agree. Participants rate their level of agreement or disagreement with nine items (five negative statements and four positive statements). The example of calculation as shown in Table 3.6 is based on responses on the instrument as answered by a single participant

Table 3.6: Operation on Item Responses

Item	Scale position	Calculation	Difference
1 (negative)	2	(s_h-x)	5-2=3
2 (positive)	5	(x-s_l)	5-1=4
3 (negative)	3	(s_h-x)	5-3=2
4 (positive)	4	(x-s_l)	4-1=3
5 (negative)	2	(s_h-x)	5-2=3
6 (positive)	5	(x-s_l)	5-1=4
7 (negative)	1	(s_h-x)	5-1=4
8 (positive)	5	(x-s_l)	5-1=4
9 (negative)	2	(s_h-x)	5-2=3
		Sum:	30

$$\text{Scale} = \left[\sum_{k\in N}(s_h - x) + \sum_{k\in P}(x - s_l)\right]\left(\frac{100}{(s_h - s_l)i}\right)$$

$$= \quad 30 \times 2.7778$$

$$= \quad 83.33 \approx 83$$

From the formula, $\sum_{k\in N}(s_h - x) + \sum_{k\in P}(x - s_l)$ yield the value of 30 (as in Table 3.6). The value is then multiplied by $100/(\mathrm{s_h} - \mathrm{s_l})\mathrm{i}$ which is 2.7778. From the example, the composite score is 83. The instrument in the example uses 5-point Likert scales. However, this formula can accommodate different scale size and number of items if researchers need to make amendments based on context of use. Different number of positive and negative statements as well as lowest anchor on the scale (0 or 1) would also yield correct results. Moreover, the absence of either positive or negative statements in the questionnaire will not affect the reliability of the formula. In this study, this formula is used to calculate sign usability based on user's perception on sign characteristics. As the calculation is analogous to the approach in System Usability Scale.

The summation of $(s_h - x)$ for negative items and $(x - s_l)$ for positive items will be equivalent to $(s_h - s_l)i$ for the construct that scores a perfect score (highest possible score). However, for the construct with the lowest score, summation would yield 0. In any case, the value calculated will always in the range of 0 to 100. Example, for a construct that has the highest possible score (5 scale size, 1-5 point, nine items):

$$Scale = \left[\sum_{k\in N}(s_h - x) + \sum_{k\in P}(x - s_l)\right]\left(\frac{100}{(s_h - s_l)i}\right)$$

$$= \quad [(5\text{-}1) + (5\text{-}1) + (5\text{-}1) + (5\text{-}1) + (5\text{-}1) + (5\text{-}1) + (5\text{-}1) + (5\text{-}1) + (5\text{-}1)]\ (2.7778)$$

$$= \quad 36 \times 2.7778$$

$$= \quad 100.0008 \approx 100$$

Therefore, for a construct that has the lowest possible score

$$\text{Scale} = \left[\sum_{k\in N}(s_h - x) + \sum_{k\in P}(x - s_l)\right]\left(\frac{100}{(s_h - s_l)i}\right)$$

$$= [(5\text{-}5) + (5\text{-}5) + (5\text{-}5) + (5\text{-}5) + (5\text{-}5) + (1\text{-}1) + (1\text{-}1) + (1\text{-}1) + (1\text{-}1)]\ (2.7778)$$

$$= 0 \times 2.7778$$

$$= 0$$

3.9.3 Formula testing and analysis

PSUS uses a formula to calculate the composite score for responses provided by every participant. Probing testing and investigation of the formula found that PSUS formula can also be used to replace the conventional composite score calculation methods such as mean and sum computation. This sub-section examines the use of the formula with two different data sets to assess if the formula produces consistent results.

This analysis compares the difference of *p*-value of common statistical tests using three different methods to calculate composite score. Both parametric and non-parametric tests are used to compare the result. By using gender as categorical data, independent t-test and Man-Whitney U test was adapted to observe the difference between the three methods. This analysis is performed with two different data sets:

I. Data set A – Smaller sample size (df=53), 9 items, 5 scale-size, no missing value, simulated data

II. Data set B – Larger sample size (df=203), 20 items, 5 scale-size, missing value present

As in Table 3.7, Scale, Mean, and Sum yield the same value of t-test results $t(53)=-0.431$, $p=0.668$ as shown in Table 3.7 for data set A. There is no difference in *p-value* and *t-value* for calculation between 0-100 scale, mean as well as sum. However for data set B,

Scale, Mean, and Sum yield slightly different results: Scale $t(203)= -0.346$, $p=0.730$, Mean $t(203)= -0.261$, $p=0.794$, and Sum $t(203)= -0.576$, $p=0.566$.

Table 3.7: Independent T-test Results Comparison for All Composite Scores

Data set		T-test for Equality of Means				
		t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference
A	Scale	-0.431	53	0.668	-1.08933	2.52510
	Mean	-0.431	53	0.668	-0.04357	0.10100
	Sum	-0.431	53	0.668	-0.39216	0.90903
B	Scale	-0.346	203	0.730	-0.85177	2.46213
	Mean	-0.261	203	0.794	-0.02582	0.09880
	Sum	-0.576	203	0.566	-1.14620	1.99132

Mann-Whitney U test statistics comparison as in Table 3.8 shows all composite scores give the exact result for data set A, Scale (*Mdn*=55.6), Mean (*Mdn*= 3.2), Sum (*Mdn*=29), *U*=356, *p*=0.993. For data set B, the results are slightly different. From both parametric and non-parametric tests, all results are non-significant. In this case, the use of parametric and non-parametric did not affect the result (significant/non-significant).

Table 3.8: Mann-Whitney U Test Comparison for All Composite Scores

	Data set A			Data set B		
	Scale	Mean	Sum	Scale	Mean	Sum
Mann-Whitney U	356.500	356.500	356.500	3518.500	3562.500	3385.000
Wilcoxon W	951.500	951.500	951.500	4553.500	4597.500	4420.000
Z	-0.009	-0.009	-0.009	-0.232	-0.107	-0.612
Asymp. Sig. (2-tailed)	0.993	0.993	0.993	0.817	0.915	0.541

Grouping Variable: gender

To support this claim, correlational analyses were performed to examine the relationship between three different calculations of composite scores. Table 3.9 shows the result of this correlation analysis. Table 3.9 examines the relationship between three different approaches to calculate the composite score (using sum, the formula, as well as mean). For data set A, Spearman rho and Pearson r coefficient yield perfect correlation $r, r_s = 1$, $p<0.01$

for three different methods of composite score computation. For data set B, the correlation coefficients range from 0.988-0.999, which is very close to perfect correlation of 1.

Table 3.9: Spearman and Pearson Comparison for all Composite Scores

Data	Correlation	Composite Score	Coefficient, r/r_s		
			Scale	Mean	Sum
A	Spearman's rho	Scale	-	1.000**	1.000**
		Mean	1.000**	-	1.000**
		Sum	1.000**	1.000**	-
	Pearson's r	Scale	-	1.000**	1.000**
		Mean	1.000**	-	1.000**
		Sum	1.000**	1.000**	-
B	Spearman's rho	Scale	-	0.998**	0.993**
		Mean	0.988**	-	0.988**
		Sum	0.993**	0.988**	-
	Pearson's r	Scale	-	0.999**	0.994**
		Mean	0.999**	-	0.990**
		Sum	0.994**	0.990**	-

** Correlation is significant at 0.01 level (2-tailed)

At this point, the question arises of what could have contributed to this slight difference in statistics. The simulated data set A was altered to examine the alterations include scale size, the number of items, sample size, and the presence of missing value. The change in scale size, number of items, and sample size do not affect the result. The presence of missing values, on the other hand, affects statistics such as t and p value for all composite score methods. This proves that although the scale is affected by missing value, this problem is not unique to PSUS scale.

In the absence of missing values, the analysis for t-test and Mann-Whitney U test produce the same result for all composite score calculation method. Since the controversies of using mean and sum is well discussed in previous chapters, the use of the formula is a better alternative to calculate the composite score for Likert-type items. This result proved that the formula can also be used to combine the score of Likert-type items measuring the

same construct and subsequently used in various statistical analyses. The method for implementation was also described to aid practitioners to apply this formula. In addition, the scale yielded from this formula is also more intuitive to be interpreted.

3.10 Suggestion for Interpretation of Results

This section details some suggestions to interpret results obtained from PSUS. PSUS consists of open-ended sign comprehension test as well a set of Likert-type items to measure sign usability based on cognitive sign characteristics. This section is divided into two parts : Interpretation for open-ended sign comprehension test (3.10.1) as well as Interpretation for (3.10.2).

3.10.1 Interpretation for open-ended sign comprehension test

3.10.1.1 Coding of sign comprehension

Open-ended sign comprehension requires the participants to recall the meaning rather than to recognize the sign meaning from the options provided. Open-ended sign comprehension test is perceived to be more reliable than recognition technique such as multiple-choice question; therefore, this approach was adapted in several studies (Bazire et al., 2005; Ben-Bassat & Shinar, 2006; Podany et al., 2004; Ward et al., 2004; Zakowska, 2001). In this study, open-ended comprehension test is adapted alongside Likert-type items.

To quantify and interpret the level of sign comprehension, participants' answers were coded. As observed in a study on the effectiveness of sign training, participant responses were coded into one of two categories: incorrect (coded as 0) or correct (coded as 1) (Ward et al., 2004). This approach is simple, but its validity is questionable when dealing with partially correct answers. A study explains partially correct as incomplete; for example,

when presented with “no left turn” sign, the participant answered “no turn” instead of “no left turn” (Zakowska, 2001).

To solve this, several studies opted for four categories of accuracy: Correct and complete (coded as +2), partially correct (coded as +1), incorrect (0), and opposite meaning (coded as -2) (Bazire et al., 2005; Ben-Bassat & Shinar, 2006; Podany et al., 2004; Ward et al., 2004; Zakowska, 2001). These studies reported that approach of coding sign comprehension result is straightforward with less than 5% of answers require more than one coder, in the form of unanimous group decision. This study uses just three categories of accuracy (correct, partially correct, and incorrect) because opposite meaning can be regarded as incorrect answer. Past studies have not justified the need to assign a different code to opposite meaning.

3.10.1.2 Cognitive mapping of perceived meaning

Cognitive mapping of perceived sign meaning stemmed from the abundance of efforts to utilize open-ended sign comprehension test that requires the participants to provide sign meaning verbally or in written form. Generally, researchers report frequency or percentage of correct answers. This study extends the data representation by incorporating not only correct responses but also wrong responses including the number of responses, relative responses frequency as well as the relative person frequency. A study noted that extra information gained about confusions and misinterpretations people make can provide vital information in subsequent redesign work (Wolff & Wogalter, 1998).

Computation of relative frequency for an open-ended question is recommended to identify recurring response patterns (Evans & Rooney, 2014; Owen & Brooks, 2008; Spönemann, 2015). Sign meaning is assigned a relative response value ranges from 0-1.

High value indicates increased likelihood that an answer is favourable by the participants. By including both correct and wrong responses, it allows the researcher to identify common misinterpretation of a sign. It also allows the identification of signs that are well understood and signs that are confusing to the users. That aside, researcher can easily pinpoint which meaning is commonly associated with a sign. However, the actual value of the relative frequency should be interpreted with care, as it can be easily misinterpreted. The comparison of the relative frequency should be examined instead of the actual value (Spönemann, 2015). For example, a sign meaning that has the relative frequency value of 0.7 is more frequently stated by the participants than a sign meaning that has the relative frequency value of 0.5.

3.10.2 Interpretation for Likert-type items

This sub-section shows suggested methods (as in Sub-section 3.10.2.1 and 3.10.2.2) to interpret the composite score of Likert-type items in PSUS.

3.10.2.1 Acceptability range and grade scale

The computation of the mean score of every item may lead to a dangerous assumption on the usability of the system. Therefore, the PSUS formula yields a single usability score. The misinterpretation of the scale is common. The score of 70 does not represent a product is 20% more usable than a product that scores 50. A high score is a mere indication that the product is perceived to be usable, and a low score is an indication of usability failure. To aid with the interpretation, Bangor et al. (2009, 2008) suggested a guideline for the interpretation of the single usability score.

The scale calculated with PSUS can be interpreted individually, or collectively using predefined acceptability range and grade scale. However, more research needed to interpret the scale. A comprehensive study by Bangor et al. (2008) collected over 3,500 SUS data

over more than a decade used to evaluate different products and systems. A follow up study was conducted by Bangor et al. (2009) by administering SUS to 964 participants using the same type of interfaces and diverse samples (diverse age and balanced gender) as reported by Bangor et al. (2008). The studies include various types of user interface, including Interactive Voice Response (IVR) systems, Graphical User Interfaces (GUI), hardware, websites, cell phones, and televisions. Participants were asked to provide the adjective rating after they responded to SUS items. Figure 3.7 shows the mean score for each adjective rating in both studies.

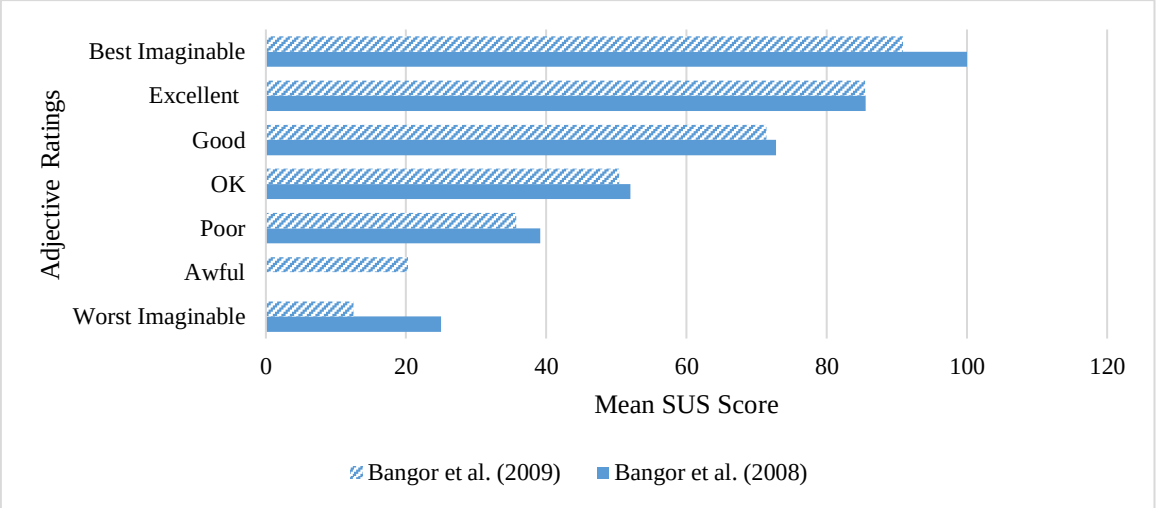


Figure 3.7: Mean SUS score for adjective ratings (Bangor et al., 2009, 2008)

From the results in Figure 3.7, it was concluded that it is possible to interpret the individual SUS score as shown in Figure 3.8 (Bangor et al., 2009, 2008). The introduction of the adjective rating scale in the interpretation of composite scale (as in Figure 3.8) is eminently useful for human factors practitioners to explain the result more intuitively non-human factors practitioners (Bangor et al., 2009, 2008).

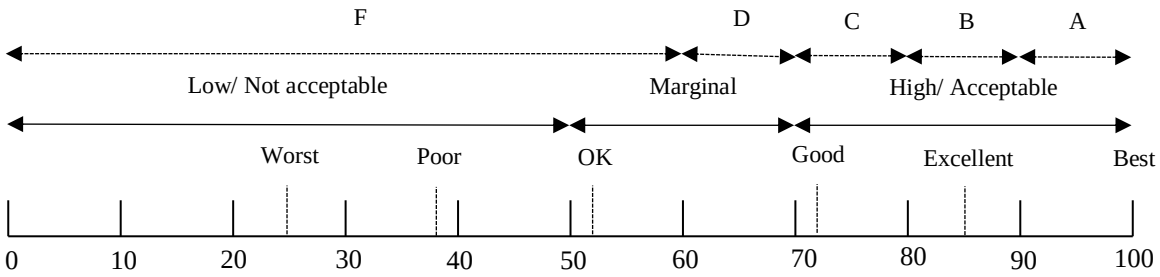


Figure 3.8: Grade scale and acceptability range (Bangor et al., 2009, 2008)

By collecting SUS data, researchers can make inference about usability based on scale. Usability of a product can be classified based on its grade scale, acceptability range and 6-anchor categorization from worst imaginable to best imaginable. Grade scale assigns a grade to reflect the usability of a product. Grade F assigned to SUS scores below 60, indicating low usability. Grade D is assigned to scores between 60 and 70, grade C for scores from 70 to 80, and grade B for scores from 80 to 90. The highest grade is grade A that will be assigned to score more than 90. Acceptability range is simpler, in a way that it categorizes score into three categories: low/ not acceptable, marginal, and high/ acceptable. This study suggests the use of SUS acceptability range because of its simplicity and there is no viable advantage of grading traffic signs. It is also ambiguous how different grade and adjective rating differ.

3.10.2.2 Dichotomous response

Another suggestion to interpret the individual scores is to reduce the score to dichotomous low/high value (i.e. disagree/agree, dissatisfied/satisfied) based on the calculation of midpoint. Should researchers wish to reduce the scores into two categories (low satisfaction/high satisfaction; acceptable/not acceptable etc.), midpoint acts as a dividing point for lower and higher end of the scale (example as in Equation 3.4). Such

reduction will categorize individual scores into dichotomous categorical variables. However, a different context of use may require a different dividing point. For a 5-scale instrument with 9 items, the midpoint of the scale is calculated as below:

$$midpoint = \left[s_h - \left(\frac{s_h + s_l}{2}\right)\right]\left(\frac{100}{(s_h - s_l)i}\right)i \quad \ldots \textbf{Equation 3.4}$$
$$= \left[5 - \left(\frac{5+1}{2}\right)\right]\left(\frac{100}{(5-1)9}\right)9 \approx 50$$

This method corresponds to the "forced choice" method whereby the subjects are forced to pick a side and the "neutral" or "undecided" option is removed (Allen & Seaman, 2007). Low acceptability value of SUS ranges from 0-50 was observed from the collection of SUS data for a decade. Figure 3.9 shows the dichotomous range based on SUS score (Bangor et al., 2009, 2008). This is useful to separate critically problematic or confusing signs and to view the distribution of respondent.

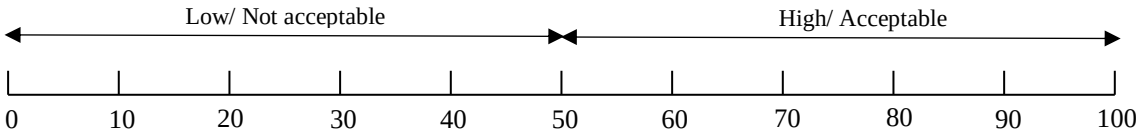


Figure 3.9: Dichotomous range of PSUS

3.10.3 Classification of sign

By using results from open-ended sign comprehension test (as in Sub-section 3.10.1.1), and acceptability range (as in Sub-section 3.10.2.1), study attempts to categorize signs into different classes of usability. As explained in Sub-section 1.3.3, sign usability is different than any other interfaces; therefore, it must be analysed differently. It has been established that comprehension is the most important aspect of sign usability (Ward et al.,

2004). Therefore, researchers must always prioritize sign comprehension before the composite score of

The interpretation of PSUS Acceptability range is adapted from a study based on various SUS data (Bangor et al., 2009, 2008). PSUS Acceptability range, when combined with sign comprehension, allows the classification of sign based on sign usability. As in Table 3.10, sign can be classified into nine classes in descending usability: Extremely low usability (class 9), Low Usability (class 8), Misinterpreted sign (class 7), Somewhat low usability (class 6), Below average usability (class 5), Intended meaning partially conveyed (class 4), Effective but low satisfaction (class 3), Acceptable usability (class 2), Very high usability (class 1).

Table 3.10: PSUS Classification of Sign

		PSUS Acceptability range		
		Low	**Moderate**	**High**
Comprehension	**Correct**	Effective but low usability (3)	Acceptable usability (2)	Very high usability (1)
	Partial	Somewhat low usability (6)	Below average usability (5)	Intended meaning partially conveyed (4)
	Incorrect	Extremely low usability (9)	Low usability (8)	Misinterpreted (7)

Shinar et al. (2003) , Shinar & Vogelzang (2013), and Zakowska (2001) realized the problem of partially understood signs; therefore, they make distinction between fully comprehended signs, partially comprehended signs, and wrongly comprehended signs. Table 3.11 shows proposed nine sign usability classes.

Table 3.11: Sign Usability Class and Description (Ascending)

Class	Classification	Description
9	Extremely low usability	Lowest sign usability class. Signs in this class scored low PSUS score and sign comprehension. The use of signs in this class should be avoided. Sign redesign is recommended. Criteria: PSUS ≤50; incorrect sign comprehension
8	Low usability	Generally regarded as low usability, signs in this usability class recorded average PSUS score and incorrect correct comprehension. Criteria: 50< PSUS≤70; incorrect sign comprehension
7	Misinterpreted	A sign is regarded as misinterpret in two circumstances: • Most participants perceive the sign to have a certain meaning which is not the true meaning of the sign (Houldin, 2011). • When participants perceive the signs to mean the opposite of its true meaning (Ben-Bassat & Shinar, 2006; Mukhopadhyay et al., 2013). In both cases, participants are confident that they have provided the true meaning of the sign, but in truth they have provided the incorrect meaning of the sign. Criteria: PSUS >70; incorrect sign comprehension
6	Somewhat low usability	Signs in this usability class recorded low PSUS score and partially correct comprehension. Criteria: PSUS≤50; partially correct sign comprehension
5	Below average usability	Signs in this usability class recorded average PSUS score and partially correct comprehension. Criteria: 50< PSUS≤70; partially correct sign comprehension
4	Intended meaning partially conveyed	When the intended meaning of a sign is partially conveyed but the sign recorded high usability score, individual characteristics (such as cultural differences and locality) can be the cause of imprecise comprehension.(Shinar et al., 2003). Criteria: PSUS >70; partially correct sign comprehension
3	Effective but low usability	In some cases, the measure for sign effectiveness is the extent to which users can identify the correct meaning of the sign (Jian et al., 2011; Yuan et al., 2014). Signs in this usability class recorded low PSUS score but high rate of correct comprehension. Criteria: PSUS ≤50; correct sign comprehension
2	Acceptable usability	When a sign records a high rate of correct sign comprehension and average PSUS score, it is considered acceptable. There is no major issue on the use of signs in this class. Criteria: 50< PSUS≤70; correct sign comprehension
1	Very high usability	This usability class is the highest class of sig usability, of which records high PSUS score (more than 70) and high percentage of correct sign comprehension. Criteria: PSUS >70; correct sign comprehension

This method of classification will add imperative sophistication to the analysis, which allows practitioners to perceive sign usability will greater flexibility. Since the analysis of comprehension score and PSUS acceptability range is independent, it is possible to impart the usability of sign that records varying polarity in terms of comprehension level and acceptability range. For example, the importance of added sophistication is conspicuous when the acceptability range is high but has a different level of comprehension. There are three possible classifications that can be assigned to a sign when the PSUS acceptability is high: comprehensible, intended meaning partially conveyed, or misinterpreted. This approach is also more relevant as compared to conventional methods that rely on the composite score. Practitioners using merely composite score of rating scale in assessment risk losing invaluable information.

3.11 Chapter Summary

The development of PSUS stemmed from emerging need and limitations of existing methods. To establish the underlying concept of sign usability based on cognitive characteristics, analysis of literature was conducted by utilizing several tools, and search method. The articles were coded to find notable themes and concepts to be integrated into the new measure. An item pool consisting of existing usability instruments was filtered to find relevant items and concepts. This was conducted with redundancy and relevancy analysis. This chapter also described readability tests, of which serve as an added justification of readability of the instrument.

The establishment of measure quality was conducted with various techniques. In the early stage, Multiple Responses Analysis with Cochran's Q and McNemar was used to assess the relevancy of items in relation to CSC identified. Instrument development using

Rasch measurement model allows for reliability assessment and other quality assessment, which includes item polarity, unidimensionality, and detection of item bias.

The design of instrument also includes the development of a formula to calculate its composite score. Development of formula includes describing the underlying concept and some examples of calculation. The mechanism of calculation is discussed. In addition to PSUS formula, this chapter also provides justification for instrument assessment methods, selection of participants, test format as well as test-administering method. The chapter concludes by explicating interpretation approaches for the PSUS instrument.

CHAPTER 4

RESULTS

4.1 Overview

This chapter contains four parts: (1) Item development, (2) Rasch analysis (3), Assessment of test signs using PSUS, and (4) Discussion. Table 4.1 shows the overview of the chapter.

Table 4.1: Overview of Chapter Four

Chapter 4: Results			
Part one: Item development (Section 4.3)	Part two: Rasch analysis (Section 4.5 3.5)	Part three: Assessment of test signs (Section 4.6)	Part four: Discussion (Section 4.7)
Results of item development process such as item feedback, readability test and item justification. Also shows results from various methods described in Chapter 3, including MRA, Cochran's Q, and McNemar.	Results from Rasch analysis to ensure reliability and validity of the instrument	Assess the usability of test signs to demonstration of use of composite score calculation and interpretation methods	Provides major findings, linking findings with the problem statements and discusses issues pertaining to sign usability.

The first part of this chapter reports the development of a new usability instrument based on findings on cognitive sign characteristics identified in Chapter 2. The development of a framework on cognitive sign characteristics (CSC) that constitute sign usability gives an insight on cognitive components that need to be addressed when generating item pool. This knowledge, along with the understanding of generic usability instruments served as a

tool to shortlist and to generate items for the instrument. This includes results from group review and participants feedback on initial items.

Methods and tools to develop the instrument were discussed in Chapter 3, inclusive of justifications and validation technique. This chapter provides results from various methods described in Chapter 3, including MRA, Cochran's Q, and McNemar as the post-hoc test. This is followed by results from Rasch analysis to ensure reliability and validity of the instrument. The final part of this chapter is the demonstration of use and interpretation methods of the new instrument. This gives future research an idea how to use the instrument and how to interpret the result based on proposed methods. Presentation of results is followed by discussion section. Discussion section discusses the difference of sign usability and usability of other system/product, CSC, and interpretation of scale. This chapter answers to objectives as shown in Table 4.2.

Table 4.2: Objectives and Results

Objectives	Sections
- To develop items for a new measure to assess road users' perception on traffic sign usability.	Item feedback, readability tests and item justification are described in Section 4.3
- To assess the reliability and validity of the instrument.	Reliability and validity of the instrument are assessed from Section 4.3 to Section 4.5
- To determine the best methods to interpret the data obtained with the instrument. - To assess the usability of test signs using PSUS.	Interpretation method is further described and implemented in Section 4.6. This section also assesses the usability of three test signs.

4.2 Demographic Information

A total of 284 participants participated in this study. Table 4.3 shows the participants in various stages. The table shows the frequency and percentage of participants based in each

category. For item review and translation, 13 item reviewers were involved. 68 students who have basics in cognitive ergonomics were involved in Multiple Responses Analysis. There were no identifying/ demographic questions asked because it was not necessary for this stage. Finally, 205 participants were involved in Rasch analysis. Participants' age ranges from 20-27.

Table 4.3: Demographics of Participants

Phase	Demographic Information		Observed	Percent	Valid Percent
Item review & translation	Gender	Male	4	30.8	30.8
		Female	9	69.2	69.2
	Occupation	Masters student	5	38.5	38.5
		PhD Student	5	38.5	38.5
		Lecturer	3	23.0	23.0
	Nationality	Malaysian	12	92.3	92.3
		Non-Malaysian	1	7.7	7.7
Rasch	Gender	Male	45	22.0	22.0
		Female	160	78.0	78.0
	Age	20-21	70	34.1	34.8
		22-23	113	55.1	56.2
		24-25	16	7.8	7.9
		26-27	2	0.9	1.0
	Race	Malay	99	48.3	48.3
		Chinese	50	24.4	24.4
		Native S&S	47	22.9	22.9
		Indian	4	2.0	2.0
		Others	5	2.4	2.4
	Mother tongue	Malay	109	53.2	53.2
		Chinese	53	25.9	25.9
		Dayak	37	18.0	18.0
		Native S&S	5	2.4	2.4
		Tamil	1	0.48	0.48
	License status	None	78	38.6	39.6
		Learner's License	23	11.2	11.7
		Probation License	35	17.1	17.8
		Normal License	58	28.3	29.4
		Others	3	0.01	0.02

4.3 Item Development

This section explains the process of item development, prior to reliability and validity assessment. There are three sub-sections: Sub-section 4.3.1 - Item feedback, Sub-section 4.3.3 - Readability tests and Sub-section 4.3.2 - Item rationales and justification. Figure 4.1 is the graphic for item development from initial item pool to the final version of PSUS. It shows the action taken in each version and the number of item in each process.

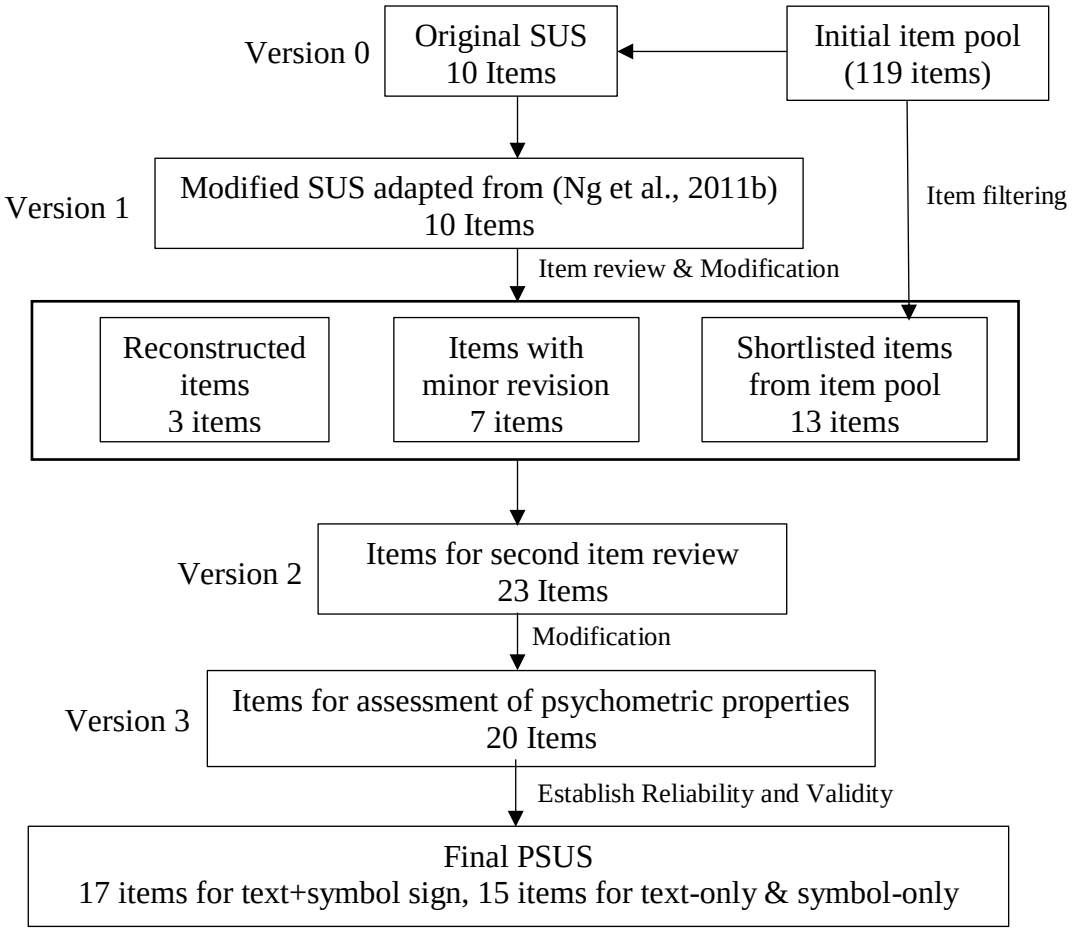


Figure 4.1: Item development from version 0 to the final version

The Initial item pool consists of 119 items, of which includes SUS instrument that serves as the base of this instrument. This decision is made because of the existence of sign usability instrument developed by Ng et al. (2011b), of which was the modified version of

the SUS. The modified SUS was revised and reconstructed based on reviews (Sub-section 4.3.1). Based on the knowledge about cognitive sign characteristics (CSC) and the analysis on item pool, 13 items were added or adapted from generic usability instruments. The rationales and justification for each item are explained in Sub-section 4.3.2. When item reviewers reached an anonymous decision and there were no further changes on the instrument, readability tests were conducted to examine the readability of the instrument (Sub-section 4.3.3).

4.3.1 Item feedback

In the past, SUS has been modified to assess sign usability. Due to the existence of modified SUS for sign usability, it is crucial to examine the suitability and limitation of modified SUS before adding other items. The process of instrument development is iterative; therefore, suitable items will be brought forward to be included in the next version of the instrument. For the next step, items are modified or removed based on the results of item review. Additional items are introduced, and the inclusion is justified based on findings from analysis of literature and CSC identified. Ng et al. (2011b) modified SUS items to assess safety sign usability. To accommodate the context of use in this study, items are reviewed and modified accordingly.

The first version of the instrument as in Table 4.4, of which is a modified version of SUS for assessing sign usability. The table compares original SUS items and a modified version of SUS.

Table 4.4: Modification of SUS

Item Number	Original SUS (Brooke, 1996)	Modified SUS adapted from (Ng et al., 2011b)
1	I think that I would like to use this system frequently	-
2	I found the system unnecessarily complex	I found the road signage unnecessarily complex
3	I thought the system was easy to use	I thought the road signage was easy to understand
4	I think that I would need the support of a technical person to be able to use this system	I think that I would need support to be able to understand this road signage
5	I found that the various functions in this system were well integrated	I found that the function of this road signage is well understood by people
6	I thought that there was too much inconsistency in this system	I thought there was too much inconsistency in this road signage
7	I would imagine that most people would learn to use this system very quickly	I would imagine that most people would learn this signage very quickly
8	I found the system very cumbersome to use	I found this signage very awkward to understand
9	I felt very confident using the system	I felt very confident in my understanding of this road signage
10	I needed to learn a lot of things before I could get going with this system	I needed to learn a lot of things before I could understand this road signage

The first item of SUS is omitted because it is deemed not suitable in the context of traffic signs. Unlike conventional consumer products or websites, the use of road signs is involuntary. Road signs are traffic control devices (TCDs) presented to the road users on the road with the function of controlling the flow of traffic, provide directional guidance, reassure drivers about their location, indicate sites of local services, indicate the rules that the user must abide, and warn of upcoming changes or hazards (Mitchell, 2010). Road users do not decide the location of road signs or when they will see one while they are on the road. Therefore, the first item on the SUS is irrelevant in this context. Practitioners can either omit the first item of the SUS or use a different alternate statement to replace it. However, to replace the original statement, it must reflect the usability aspect of one specific sign while

avoiding overlapping with other statements. A study modified the first item to *"I thought that I would like to pay attention to safety signs at workplace and other public places"* (Ng et al., 2011b). However, this statement is a general statement and does not target a specific sign, rendering it invalid for an instrument measuring a targeted sign. The instrument intends to assess the usability of a specific sign presented to the road user and not road users' attitude towards traffic signs in general. For this reason, both original and modified SUS were omitted in version 1.

A study proposed two guidelines for practitioners to intend to implement SUS, involving simplification of terms and the use of product-specific terms appropriate for a given test (Lewis & Sauro, 2009). For example, the word "cumbersome" should be changed to "awkward". This substitution is widely recommended by researchers (Bangor et al., 2008; Finstad, 2006) and the word "system" can be changed to "product" depending on test scenario or interface.

Table 4.5: Researcher's Notes and Action on Item Feedback (version 1)

Item	**Researcher's notes and action on item feedback**
I found the road signage unnecessarily complex	- The term "unnecessarily complex" is ambiguous and does not convey a clear meaning. - The term "unnecessarily" might be redundant in this context. - Use alternative phrase for "road signage". Commonly suggested term is "traffic sign" or "sign" for all items. - Suggest the omission of "I found" as it is unnecessary. - Simplify to "This sign is complicated".
I thought the road signage was easy to understand	- The item is relevant. - Suggest the omission of "I thought" as it is unnecessary. - Suggest changing to "The meaning of this sign is easy to understand". - Paraphrase to "This sign is easy to understand".
I think that I would need support to be able to understand this road signage	- Change "support" to "help". - Restructure the sentence - Suggest the omission of "I think" as it is unnecessary. - I need help in understanding this sign

Table 4.5 Continued

	- Paraphrase to "I need help in understanding this sign for the first time".
I found that the function of this road signage is well understood by people	- Restructure the sentence. - Suggest the omission of "I found" as it is unnecessary. - The word "people" should be changed to "I". - Item redundancy as it represents the extent to which a sign is easily comprehended.
I thought there was too much inconsistency in this road signage	- Restructure the sentence. - Suggest the omission of "I thought" as it is unnecessary. - What does it mean by "inconsistencies"? - Suggest changing to "I have seen many variations of this sign".
I would imagine that most people would learn this signage very quickly	- Restructure the sentence. - Suggest the omission of "I would imagine" as it is unnecessary. - The phrase "most people" should be changed to "I". - Paraphrase to "I would learn this sign quickly".
I found this signage very awkward to understand	- What does it mean by "awkward to understand?" - Suggest the omission of "I found" as it is unnecessary. - Suggest simplification or paraphrasing. - The word "awkward" does not have a clear meaning. - Based on context, suggest changing it to "tricky to understand" or "difficult to understand". - "Awkward to understand" may indicate that a sign can be easily confused with another sign. - Might overlap with other items on comprehension.
I felt very confident in my understanding of this road signage	- Item seems relevant and straightforward. - What if the participant is confident but the understanding is wrong? - Paraphrase to "I am confident that I understand the meaning of this sign".
I needed to learn a lot of things before I could understand this road signage	- Some resemblance with "I think that I would need support to be able to understand this road signage" - Consider for removal or paraphrasing. - Deemed unsuitable for traffic sign context because users do not need to execute complex operation/steps to use a traffic sign.

* Responses are grouped and summarized into bullet points by the researcher.

Responses are grouped and summarized into bullet points as in Table 4.5. Based on feedback, three items were found problematic and needed to be reconsidered. Item "I found that the function of this road signage is well understood by people", "I found this signage very awkward to understand" and "I needed to learn a lot of things before I could understand

this road signage" were reconstructed for the next version of the instrument. There were seven items were retained with modifications. Table 4.6 shows researcher's notes and action on item feedback for the set of items after the addition of new items and modification of items. This results in the second version of the instrument.

Table 4.6: Researcher's Notes and Action on Item Feedback (version 2)

Item	Researcher's notes and action on item feedback*
This sign is easy to read	- Simple, clear and easy to understand
This sign does not convey simple meaning	- Suggest changing "simple" to "clear"
The meaning of this sign is easy to understand	- The word "meaning" is not required. Better to use "This sign is easy to understand.
I need help in understanding this sign	- Adding the phrase "for the first time" to indicate learning the sign for the first time. This is to differentiate this item from item "This sign is easy to understand".
The shape of this sign is relevant	- Rather than using the word "relevant", should consider the word "suitable"
There are many inconsistencies in this sign	- What does it mean by "inconsistencies"? - Suggest changing the item to "I have seen many variations of this sign"
Most people would learn this road sign very quickly	- The phrase "most people" should be changed to "I"
The meaning of this sign is awkward to understand	- What does it mean my "awkward to understand"? The word "awkward" does not have a clear meaning. - Based on context, suggest changing the item to "tricky to understand" - The phrase "the meaning of" is not required. - The item has the resemblance to item "This sign is easy to understand".
I am very confident in my understanding of this sign	- Paraphrase to "I am confident that I understand the meaning of this sign"
It is not easy to remember the meaning of this sign	- Suggest changing "not easy to forget" to "I easily forget"
The meaning of this sign is straightforward	- Seems overlapping with previous items ("The meaning of this sign is awkward to understand" and "This sign does not convey simple meaning")
The symbol design in this sign is inappropriate	- What is symbol design? - Based on the context of use, replace "symbol" with "icon" as "icon" is clearer and easier to understand.
The colour in this sign is appropriate	- Better to use the word "suitable" rather than "appropriate".

*Responses are grouped and summarized into bullet points

After modification of items is performed, a new version of the instrument (version 2) is conceived and was reviewed again. Feedbacks on the second version of the instrument are shown in Table 4.6. For this version, there are ten items were regarded as having a clear meaning and easy to understand. The items are: “This sign is complicated”, “The text in this sign conveys a clear meaning”, “The icon in this sign conveys a clear meaning”, “This sign is easy to read”, “This sign can easily capture my attention”, “I know what I should do when I see this sign”, “The shape of this sign is not suitable”, “The colour contrast of this sign is poor”, “I am not familiar with this sign”, and “The text in this sign is easy to read”. Feedbacks for the rest of the items are summarized in Table 4.6.

4.3.2 Item rationales and justification

Rationales and justifications for each item that will subsequently be tested in readability tests (Sub-section 4.3.3), Multiple Responses Analysis (Section 4.4), and Rasch analysis (Section 4.5) are as follow:

I. Item: I am confident that I understand the meaning of this sign.

   Sources/rationales: This item corresponds to Static Comprehension as described in a study, referring to the comprehension of still-image (Charlton, 2006). This item gives information about road user’s confidence on a sign. In many cases, although road users are confident about their comprehension, the meaning given is incorrect. This item was adapted from (Brooke (1996) and Ng, Lo, & Chan, 2011a, 2011b). This item identifies if the meaning of the sign was a guess (correct response but low rating), or false interpretation (wrong response but high rating).

II. Item: I need help in understanding this sign for the first time.

Sources/rationales: An effective sign is supposed to be self-explained and relatable to road users' knowledge so that it can be comprehended with minimal help and training (Bazire et al., 2005). A study acknowledged that road users should be able to understand a sign with a brief glance (Podany et al., 2004). The item was adapted from Podany et al. (2004), Brooke (1996), and Ng et al. (2011a, 2011b).

III. Item: I am familiar with this sign.

Sources/rationales: The item was adapted from usability instrument by Kirakowski & Corbett (1993). Sign familiarity is an important aspect of sign usability (McDougall et al., 1999).

IV. Item: I easily forget the meaning of this sign.

Sources/rationales: The item was adapted from usability instrument by Kirakowski & Corbett (1993). Sign usability can be improved if sign meaning can be easily remembered (Hancock et al., 2004). If a sign adheres to design features, its meaning can be easily remembered as it resembles the object it represents in the real world (McDougall et al., 1999; Ng & Chan, 2007a).

V. Item: I would learn this sign quickly.

Sources/rationales: Learnability is an important aspect of usability instruments. This item was adapted from usability instrument by Brooke (1996), Kirakowski & Corbett (1993), Lewis (1995) and Ng et al. (2011a, 2011b).

VI. Item: This sign can be easily confused with another sign.

Sources/rationales: Numerous signs can be easily confused with other signs. This was cited as one of the reasons for low sign comprehension (Al-Madani & Al-Janahi, 2002; Lee, Chuang, & Young, 2013).

VII. Item: This sign can easily capture my attention.

Sources/rationales: The item was adapted from several studies that modified SUS instrument to be used to assess traffic and safety sign usability (Brooke, 1996; Chan & Ng, 2012; Ng et al., 2011b). A sign needs to capture attention easily because in certain circumstances users may be exposed to a sign for a brief duration. In that short period of time, the intended meaning will need to be effectively delivered (Wogalter et al., 2002). This item was adapted from (Podany et al., 2004).

VIII. Item: This sign is complicated.

Sources/rationales: Simple signs were observed to have higher usability than complex signs (Wogalter et al., 2002).

IX. Item: This sign is easy to understand.

Sources/rationales: The item was adapted from several studies that modified SUS instrument to be used to assess traffic and safety sign usability (Brooke, 1996; Chan & Ng, 2012; Ng et al., 2011b). The item was observed in a usability questionnaire (Lewis, 1992, 1995).

X. Item: I have seen many variations of this sign.

Sources/rationales: The item was adapted from Brooke (1996), Ng et al. (2011a, 2011b). Effectiveness and consistency of sign are maintained by adopting the same symbol, text and shape for the same sign meaning (Ben-Bassat & Shinar, 2006).

XI. Item: This sign is easy to read.

Sources/rationales: An effective sign should be identifiable within a short period. In a study, it was demonstrated that the reason of why signs caught their attention is because the sign is easy to read (Kurniawan & Zaphiris, 2001)

XII. Item: (7 items) The colours used in this sign are not suitable; The colour contrast of this sign is good; The shape of this sign is not suitable; The icon in this sign conveys a clear meaning; The icon design of this sign is not suitable; The text in this sign conveys a clear meaning; The text in this sign is difficult to read.

Sources/rationales: A study used the term Surface Properties to represent tangible design attributes of a sign that includes colour, shape, icon, and text (Bazire et al., 2005). Colours should be suitable in a way that it has good contrast and conforms to users' mental representation of what the colour means (Funkhouser et al., 2008; McDougall et al., 1999). This is important because different shape/icon and colour represent a certain meaning that may help in comprehension. Using shape and colours that are not suitable lead to misinterpretation of the sign. (Bazire et al., 2005; Brézillon et al., 2004). Icon design should not deviate far from the user's mental model (Mcdougall & Curry, 2001). The resemblance to real world objects may ease understanding. Suitable icon conforms to symbol usability characteristics (McDougall et al., 1999).

XIII. Item: I know what I should do when I see this sign.

Sources/rationales: The item was adapted from Kirakowski & Corbett (1993). A sign conveys a certain message so that the drivers can execute in accordance with the sign meaning (Jamson et al., 2005). If drivers know what to do when seeing a sign, they have been primed to exhibit certain behaviour appropriate to the situation. It is important to ask whether the driver knows what the sign allows or prohibits. A study noted that participants have less confident about expected behaviour even when a sign is perceived to be easily readable and understandable (Houldin, 2011).

XIV. Item: It is likely that I will ignore this sign.

Sources/rationales: A study noted that a sign is considered to be effective if the users comply with the sign when they notice it (Podany et al., 2004). Other studies noticed that some signs have a high tendency to be ignored (Peschke et al., 2013; Waterson, Pilcher, Evans, & Moore, 2012). This is more apparent for signs that have a lot of written text, particularly small texts (Waterson et al., 2012). This item was adapted from (Podany et al., 2004).

4.3.3 Readability tests

After modification of version 2, version 3 of the instrument is as listed in Table 4.7. Items in Table 4.7 were tested with various readability scales. The items are assessed with several readability tests as listed in Sub-section 3.5.3 (Chapter 3) to examine their readability grade levels and readability score. Based on participant's feedback on items, it was proven that the items are generally well understood.

Table 4.7: The Perceived Sign Usability Scale (Version 3)

No.	Item
1	I am confident that I understand the meaning of this sign.
2	I need help in understanding this sign for the first time.
3	I am familiar with this sign.
4	I easily forget the meaning of this sign.
5	I would learn this sign quickly.
6	This sign can be easily confused with another sign.
7	This sign can easily capture my attention.
8	This sign is complicated.
9	This sign is easy to understand.
10	I have seen many variations of this sign.
11	This sign is easy to read.
12	The colours used in this sign are not suitable.
13	The colour contrast of this sign is good.
14	The shape of this sign is not suitable.
15	I know what I should do when I see this sign.
16	It is likely that I will ignore this sign.
17	The icon in this sign conveys a clear meaning.
18	The icon design of this sign is not suitable.
19	The text in this sign conveys a clear meaning.
20	The text in this sign is difficult to read.

Readability scores are used to assess readability and complexity of English texts. Readability grade level and readability scores were calculated using readability calculator developed by Child & Colmer (2011). Average grade level is determined by several readability formulae: Flesch Kincaid Grade Level, Gunning Fog Score, Coleman-Liau Index, Simple Measure of Gobbledygook (SMOG) Index, and Automated Readability Index (ARI). All readability tests based on grade level show how many years of education needed to understand a text easily on the first read. It is based on the U.S grade level. For example, grade level 5 shows 10-11-year-old students in the U.S can easily comprehend the text.

There are two reasons of why readability analysis was conducted. Firstly, is to provide evidence and explanation of why the items are easily understood by participants. Secondly, in a case whereby the items are being implemented in English-speaking countries,

it is intuitive to assess the level of English comprehension needed to fully understand the items. Readability tests are based on American grading system. It is a concern that researcher and academics tend to use unfamiliar wordings or field jargons in the instrument; therefore, readability tests are useful as they provide an alternate indicator of how easily the items can be comprehended. Furthermore, it also provides useful statistics such as average character per word, average words per sentence, average syllables per word, and the percentage of complex words.

The average of item statistics summary for readability test is shown in Table 4.8. It is demonstrated that the items were written in simple English with 4.1 average characters per word, 8.1 average words per sentence. From the table, it was noted that the percentage of complex words is low (9.94%).

Table 4.8: Test Statistics for 20 Items (Version 3)

No. of items	20
No. of character	668
No. of syllable	224
Average character per word	4.1
Average words per sentence	8.1
Average syllables per word	1.4
Percent of complex words	9.94%
Average grade level (US)	5 (Should be easily understood by 10–11-year-olds.)

Readability grade levels and readability scores of 20 items are shown in Table 4.9. Readability scores were calculated using Readable.io tool. From the table, it was shown that all items can be easily comprehended.

Table 4.9: Readability Grade Levels and Readability Scores (Version 3)

Item	Readability Grade Levels						Readability Scores		
	Flesch Kincaid Grade Level	Gunning Fog Score	Coleman-Liau Index	SMOG Index	ARI	Average grade level	Flesch Kincaid Reading Ease	Spache Score	New Dale-Chall Score
1	5.9	11.7	9.3	11.3	4.2	8.5	72.6	4.2	3.4
2	3.7	8	9.3	9	4.2	6.8	88	2.7	2
3	4.5	9.1	6.7	9	-0.4	5.8	73.8	2.8	2.9
4	5.2	8.2	8.5	9	2	5.8	71.8	3.7	4.3
5	0.5	2.4	9.7	3.3	2	3.0	102	2.8	0.3
6	7.6	16.9	11.7	13.2	5	8.9	56.7	3.6	4
7	7.4	14.2	13.6	11.3	5.6	8.7	54.7	3.8	4.9
8	6.6	11.6	15	9	5.3	9.5	54.7	3,2	4.1
9	4.5	9.1	9.7	9	2	6.9	73.8	2.8	0.3
10	5.2	8.2	8.2	9	2	6.6	71.8	2.7	2.4
11	0.5	2.4	3.8	3.3	-2.7	1.3	102	1.4	0.3
12	8.2	8.2	12.9	9	5.5	7.1	50.7	3.7	2.4
13	2.3	3.2	7.7	3.3	1.4	3.6	93	2.7	2.4
14	2.3	8.2	7	9	0.8	5.5	93	2.7	2.4
15	0.5	4.4	2.4	3.3	-1.4	1.8	111.1	2	0,5
16	2.3	3.6	5.8	3.3	0.3	3.0	94.3	3.6	2.2
17	3.7	3.6	8.4	3.3	2.4	4.6	84.9	5.4	5.7
18	8.2	8.2	12.1	9	4.9	8.5	50.7	4.7	6.3
19	2.3	3.6	8.4	3.3	2.4	4.0	94.3	5.4	5.7
20	2.3	8	6.4	9	0.9	5.3	94.3	3.6	4
All	4	7.4	8.6	8.5	2.1	5.2	81	2.8	3
Mean	4.2	7.6	8.8	7.4	2.3	5.8	79.4	3.4	4.8

Individual readability score should not be misused as the reliability of some readability tests may be compromised for short text. To compensate this matter and to increase the reliability of readability score, correlation of FKGL, GFS, CLI, SMOG, and ARI was assessed to examine the consistency between scores. As shown in Table 4.10, all readability scores record significant relationship. Readability scores yield consistently high estimations of correlation (0.570 – 0.933) with the average of 0.79. This shows that the readability scores are consistent.

Table 4.10: Correlation of Readability Tests

		FKGL	GFS	CLI	SMOG	ARI
GFS	Pearson Correlation	.775				
	Sig. (2-tailed)	.000				
CLI	Pearson Correlation	.811	.617			
	Sig. (2-tailed)	.000	.004			
SMOG	Pearson Correlation	.771	.944	.570		
	Sig. (2-tailed)	.000	.000	.009		
ARI	Pearson Correlation	.825	.642	.933	.609	
	Sig. (2-tailed)	.000	.002	.000	.004	
AGL	Pearson Correlation	.903	.880	.833	.882	.851
	Sig. (2-tailed)	.000	.000	.000	.000	.000

In conclusion, English-speaking countries should be able to use the instrument without simplification as it records average grade level of 5. Results show average 10–11-year-old children in the United States can comprehend the items well. These results reaffirm the notion that all items are easy to understand. Furthermore, the results facilitate the adaptation of the items in the context of English-speaking countries.

4.4 Multiple Responses Analysis (MRA)

The set of items as in Sub-section 4.3.2 and 4.3.3 is used in the Multiple Responses Analysis. MRA involves the categorization of items based on CSC. As items are constructed based on the CSC identified, it is important to ensure the items reflects at least one of the CSC. Items that are not related to any of the CSC listed will be removed.

4.4.1 Item categorization based on CSC

To justify that items are relevant to the identified sign characteristics, Multiple Responses Analysis was conducted. MRA allows an item to be assigned any number of CSC. Table 4.11 shows the list of items used in this analysis and list of CSC. Table 4.11 shows the list of items in this stage as well as the list of CSC.

Table 4.11: List of Items and CSC

Item	Items	Cognitive sign characteristics
1	I am confident that I understand the meaning of this sign.	Authoritativeness
2	I need help in understanding this sign for the first time.	Complexity Comprehension
3	I am familiar with this sign.	Familiarity
4	I easily forget the meaning of this sign.	Learnability
5	I would learn this sign quickly.	Legibility
6	This sign can be easily confused with another sign.	Meaningfulness
7	This sign can easily capture my attention.	Memorability
8	This sign is complicated.	Priming
9	This sign is easy to understand.	Saliency
10	I have seen many variations of this sign.	Standardization
11	This sign is easy to read.	Surface
12	The colours used in this sign are not suitable	
13	The colour contrast of this sign is good.	
14	The shape of this sign is not suitable.	
15	I know what I should do when I see this sign.	
16	I will likely ignore this sign.	
17	The icon in this sign conveys a clear meaning.	
18	The icon design of this sign is not suitable.	
19	The text in this sign conveys a clear meaning.	
20	The text in this sign is difficult to read.	

Results of MRA for all items are summarized in Table 4.12 to Table 4.16, not inclusive of results from Cochran's Q and McNemar tests, of which will be discussed in Sub-section 4.4.2. Table 4.12 to Table 4.16 shows the count for each CSC for the respective item, as well as relative participants and responses percentage of the CSC. From Table 4.12 to Table 4.16, Relative Participant Percentage (RPP) is the percentage of responses on a characteristic in relation to the total number of participants. Meanwhile, Relative Responses Percentage (RRP) refers to the percentage of responses on a characteristic in relation to the total number of responses assigned to an item.

Table 4.12: MRA of Item 1 to 4

Characteristics	Item 1			Item 2			Item 3			Item 4		
	Count	RPP	RRP	Count	RPP	RRP	Count	RPP	RRP	Count	RPP	RRP
Authoritativeness	1	1.47	0.58	1	1.47	0.88	0	0.00	0.00	7	10.29	7.22
Complexity	1	1.47	0.58	24	35.29	21.05	0	0.00	0.00	24	35.29	24.74
Comprehension	54	79.41	31.40	14	20.59	12.28	6	8.82	5.17	2	2.94	2.06
Familiarity	8	11.76	4.65	4	5.88	3.51	52	76.47	44.83	2	2.94	2.06
Learnability	20	29.41	11.63	22	32.35	19.30	4	5.88	3.45	5	7.35	5.15
Legibility	13	19.12	7.56	2	2.94	1.75	5	7.35	4.31	0	0.00	0.00
Meaningfulness	33	48.53	19.19	5	7.35	4.39	1	1.47	0.86	5	7.35	5.15
Memorability	7	10.29	4.07	3	4.41	2.63	12	17.65	10.34	36	52.94	37.11
Priming	12	17.65	6.98	21	30.88	18.42	1	1.47	0.86	4	5.88	4.12
Saliency	10	14.71	5.81	4	5.88	3.51	11	16.18	9.48	1	1.47	1.03
Standardization	4	5.88	2.33	4	5.88	3.51	17	25.00	14.66	9	13.24	9.28
Surface Properties	9	13.24	5.23	10	14.71	8.77	7	10.29	6.03	2	2.94	2.06
Categorization	Comprehension			Complexity, Priming, Learnability			Familiarity			Complexity, Memorability		

Table 4.13: MRA of Item 5 to 8

Characteristics	Item 5			Item 6			Item 7			Item 8		
	Count	RPP	RRP	Count	RPP	RRP	Count	RPP	RRP	Count	RPP	RRP
Authoritativeness	0	0.00	0.00	2	2.94	1.85	0	0.00	0.00	4	5.88	4.30
Complexity	4	5.88	3.05	26	38.24	24.07	0	0.00	0.00	65	95.59	69.89
Comprehension	13	19.12	9.92	2	2.94	1.85	1	1.47	1.09	3	4.41	3.23
Familiarity	5	7.35	3.82	6	8.82	5.56	2	2.94	2.17	0	0.00	0.00
Learnability	53	77.94	40.46	1	1.47	0.93	2	2.94	2.17	0	0.00	0.00
Legibility	15	22.06	11.45	2	2.94	1.85	8	11.76	8.70	0	0.00	0.00
Meaningfulness	7	10.29	5.34	3	4.41	2.78	1	1.47	1.09	0	0.00	0.00
Memorability	8	11.76	6.11	3	4.41	2.78	1	1.47	1.09	2	2.94	2.15
Priming	2	2.94	1.53	6	8.82	5.56	0	0.00	0.00	5	7.35	5.38
Saliency	10	14.71	7.63	2	2.94	1.85	60	88.24	65.22	2	2.94	2.15
Standardization	3	4.41	2.29	42	61.76	38.89	1	1.47	1.09	5	7.35	5.38
Surface Properties	11	16.18	8.40	13	19.12	12.04	16	23.53	17.39	7	10.29	7.53
Categorization	Learnability			Standardization			Saliency			Complexity		

Table 4.14: MRA of Item 9 to 12

Characteristics	Item 9			Item 10			Item 11			Item 12		
	Count	RPP	RRP	Count	RPP	RRP	Count	RPP	RRP	Count	RPP	RRP
Authoritativeness	0	0.00	0.00	1	1.47	1.14	0	0.00	0.00	0	0.00	0.00
Complexity	1	1.49	0.79	6	8.82	6.82	0	0.00	0.00	8	11.76	9.88
Comprehension	40	59.70	31.50	1	1.47	1.14	7	10.29	6.25	0	0.00	0.00
Familiarity	4	5.97	3.15	10	14.71	11.36	4	5.88	3.57	1	1.47	1.23
Learnability	21	31.34	16.54	3	4.41	3.41	10	14.71	8.93	0	0.00	0.00
Legibility	15	22.39	11.81	1	1.47	1.14	59	86.76	52.68	3	4.41	3.70
Meaningfulness	27	40.30	21.26	2	2.94	2.27	11	16.18	9.82	0	0.00	0.00
Memorability	2	2.99	1.57	3	4.41	3.41	2	2.94	1.79	0	0.00	0.00
Priming	7	10.45	5.51	1	1.47	1.14	2	2.94	1.79	3	4.41	3.70
Saliency	3	4.48	2.36	1	1.47	1.14	7	10.29	6.25	4	5.88	4.94
Standardization	3	4.48	2.36	44	64.71	50.00	1	1.47	0.89	3	4.41	3.70
Surface Properties	4	5.97	3.15	15	22.06	17.05	9	13.24	8.04	59	86.76	72.84
Categorization	Comprehension, Meaningfulness			Standardization			Legibility			Surface Properties		

Table 4.15: MRA of Item 13 to 16

Characteristics	Item 13			Item 14			Item 15			Item 16		
	Count	RPP	RRP	Count	RPP	RRP	Count	RPP	RRP	Count	RPP	RRP
Authoritativeness	1	1.47	1.01	1	1.47	1.25	10	14.71	8.62	51	75.00	52.58
Complexity	1	1.47	1.01	9	13.24	11.25	1	1.47	0.86	13	19.12	13.40
Comprehension	0	0.00	0.00	0	0.00	0.00	20	29.41	17.24	0	0.00	0.00
Familiarity	1	1.47	1.01	0	0.00	0.00	4	5.88	3.45	1	1.47	1.03
Learnability	1	1.47	1.01	0	0.00	0.00	9	13.24	7.76	0	0.00	0.00
Legibility	11	16.18	11.11	2	2.94	2.50	6	8.82	5.17	1	1.47	1.03
Meaningfulness	0	0.00	0.00	0	0.00	0.00	8	11.76	6.90	2	2.94	2.06
Memorability	2	2.94	2.02	0	0.00	0.00	2	2.94	1.72	1	1.47	1.03
Priming	1	1.47	1.01	5	7.35	6.25	45	66.18	38.79	19	27.94	19.59
Saliency	20	29.41	20.20	3	4.41	3.75	5	7.35	4.31	6	8.82	6.19
Standardization	0	0.00	0.00	3	4.41	3.75	3	4.41	2.59	2	2.94	2.06
Surface Properties	61	89.71	61.62	57	83.82	71.25	3	4.41	2.59	1	1.47	1.03
Categorization	Surface Properties			Surface Properties			Priming			Authoritativeness		

Table 4.16: MRA of Item 17 to 20

Characteristics	Item 17			Item 18			Item 19			Item 20		
	Count	RPP	RRP	Count	RPP	RRP	Count	RPP	RRP	Count	RPP	RRP
Authoritativeness	1	1.47	0.75	2	2.94	2.17	0	0.00	0.00	1	1.47	1.00
Complexity	0	0.00	0.00	17	25.00	18.48	2	2.94	1.44	33	48.53	33.00
Comprehension	17	25.00	12.69	3	4.41	3.26	20	29.41	14.39	2	2.94	2.00
Familiarity	0	0.00	0.00	0	0.00	0.00	2	2.94	1.44	0	0.00	0.00
Learnability	14	20.59	10.45	1	1.47	1.09	15	22.06	10.79	1	1.47	1.00
Legibility	17	25.00	12.69	0	0.00	0.00	21	30.88	15.11	26	38.24	26.00
Meaningfulness	43	63.24	32.09	3	4.41	3.26	41	60.29	29.50	3	4.41	3.00
Memorability	5	7.35	3.73	0	0.00	0.00	5	7.35	3.60	0	0.00	0.00
Priming	1	1.47	0.75	7	10.29	7.61	2	2.94	1.44	7	10.29	7.00
Saliency	8	11.76	5.97	2	2.94	2.17	5	7.35	3.60	0	0.00	0.00
Standardization	3	4.41	2.24	5	7.35	5.43	2	2.94	1.44	4	5.88	4.00
Surface Properties	25	36.76	18.66	52	76.47	56.52	24	35.29	17.27	23	33.82	23.00
Categorization	Meaningfulness			Surface Properties			Meaningfulness			Legibility, Surface properties, Complexity		

CSC assigned to each of the items are as follow: From Table 4.12, item 1 (Comprehension), item 2 (Complexity, Priming, Learnability), item 3 (Familiarity), item 4 (Complexity); from Table 4.13, item 5 (Learnability), item 6 (Standardization), item 7 (Saliency), item 8 (Complexity); from Table 4.14, item 9 (Comprehension, Meaningfulness), item 10 (Standardization), item 11 (Legibility), item 12 (Surface Properties); from Table 4.15, item 13 (Surface Properties). Item 14 (Surface Properties), item 15 (Priming), item 16 (Authoritativeness); from Table 4.16, item 17 (Meaningfulness), item 18 (Surface Properties), item 19 (Meaningfulness) and lastly item 20 (Legibility, Surface properties, Complexity). Items that need post hoc analysis will be further analysed with Cochran's Q. If the result of Cochran's Q is non-significant for at least one pair, McNemar tests performed to examine the difference in the proportion.

Table 4.17 shows total responses on an item (Responses), the number of participants who did not assign any characteristic to an item (Unassigned) and the number of participants

assigned at least one characteristic to an item (Assigned). From the result, all items were regarded as relevant to at least one of the CSC.

Table 4.17: Responses on Each Item for MRA

	Item																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Total Responses	172	114	116	97	131	108	92	93	127	88	112	81	99	80	116	97	134	92	139	144
Not Assigned	1	0	1	1	0	0	0	0	1	1	3	3	0	3	3	1	0	1	1	2
Assigned	67	68	67	67	68	68	68	68	67	67	65	65	68	65	65	67	68	67	67	66

Figure 4.2 illustrates the number of categorization for all items as in Table 4.17. Figure 4.3 illustrates the weight (count) of CSC. It was noted that Authoritativeness has the lowest count (83), meanwhile, surface properties recorded the most count (432). A larger weight of Surface properties may be resulting from different aspects of surface properties such as text, icon, shape, and colour. It was also noted that four items (item 2, 4, 9 and 20) were associated with more than one characteristic.

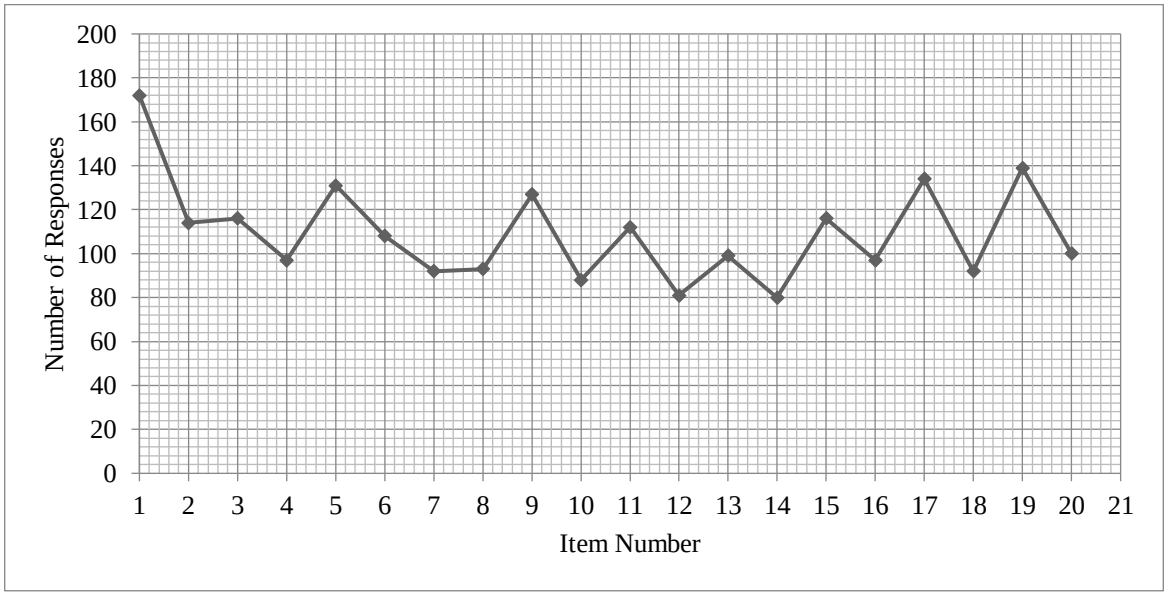


Figure 4.2: Number of categorization for all items

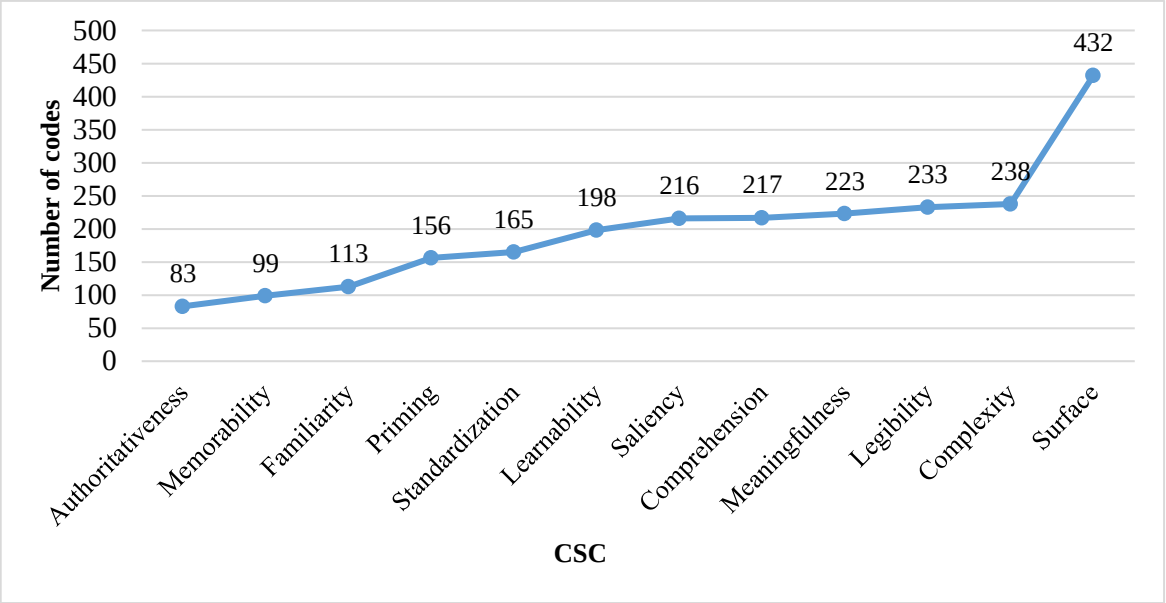


Figure 4.3: Weight of characteristics

4.4.2 Analysis of item – Cochran's Q and pairwise McNemar

If the proportion of responses on CSC is not distinguishable, Cochran's Q and pairwise McNemar analysis were adapted to assess the difference between proportions of CSC responses. This helps to identify if an item is deemed related to at least one CSC. From the result of MRA, 14 items recorded obvious differences between proportions (significant difference between the highest and the second highest count). The six remaining items (item 2, 4, 6, 9, 19, and 20) were further examined in sub-sections below.

4.4.2.1 Item 2: I need help in understanding this sign for the first time.

Based on Table 4.18, the percentage of responses based on respondents ranges from 30.88% to 35.29% and percentage based on responses ranges from 18.42% to 21.05%. For example, 35.30% of all people participated in the study selected Complexity as a characteristic related to given statement. It accounts for 21.10% of all answers selected.

Cochran's Q test was performed to investigate if there is a difference in the proportion of three characteristics assigned to item 2.

Table 4.18: Complexity, Priming, and Learnability for Item 2

Characteristics	N	Mean	Std. Dev.	Count 0	Percent (0)	Count 1	Percent (1)
Complexity	68	0.3529	0.48144	44	64.71	24	35.29
Priming	68	0.3088	0.46544	47	69.12	21	30.88
Learnability	68	0.3235	0.47130	46	67.65	22	32.35

Descriptive statistics in Table 4.18 presents a descriptive analysis of three characteristics of which the participants regarded as related to the item "I need help in understanding this sign for the first time". The proportions of responses range from 0.309 to 0.353 for the dichotomous value of 0 (not selected) and 1 (selected). Cochran's Q test did not indicate any difference between proportion of responses on three characteristics – Complexity, Decision, and Learnability, $X^2(2) = 0.259, p = 0.878$. Therefore, the result indicates there is no difference in the proportion of responses.

4.4.2.2 Item 4: I easily forget the meaning of this sign

This sub-section analyses the proportion of CSC for Item 4, "I easily forget the meaning of this sign". Descriptive statistics in Table 4.19 presents a descriptive analysis of two characteristics, Complexity and Memorability for the item "I easily forget the meaning of this sign".

Table 4.19: Complexity and Memorability for Item 4

Characteristics	N	Mean	Std. Dev.	Count 0	Percent (0)	Count 1	Percent (1)
Complexity	68	0.3529	0.48144	44	64.7	24	35.3
Memorability	68	0.5294	0.50285	32	47.1	36	52.9

The proportion of Complexity is 0.3529, and the proportion of Memorability is 0.5294. Since there was no third characteristic, McNemar test was implemented directly to investigate if the proportion of Memorability is big enough to be isolated from Complexity

P(Memorability)>P(Complexity). Table 4.20 shows the result of McNemar analysis. Results indicate the proportion of Memorability is not significantly larger p(0.063)>0.05. Therefore, it is concluded that the proportion of Complexity and Memorability is not statistically different for item 4. This result means item 4 relates to two CSC, which are Complexity and Memorability.

Table 4.20: McNemar Test for Item 4

Pair	Proportions		Difference (Unsigned)	McNemar Test (1-tailed)	For Discordant Cells:		
	P_A	P_B			A=1 B=0	A=0 B=1	Odds Ratio (larger/smaller)
Complexity (A) Memorability (B)	0.3529	0.5294	0.1765	0.0632	20	32	1.6

4.4.2.3 Item 6: This sign can be easily confused with another sign.

This sub-section analyses the proportion of CSC for Item 6, "This sign can be easily confused with another sign". Descriptive statistics in Table 4.21 presents a descriptive analysis of two characteristics, Complexity and Standardization for the item "This sign can be easily confused with another sign".

Table 4.21: Complexity and Standardization for Item 6

Characteristics	N	Mean	Std. Deviation	Count 0	Percent (0)	Count 1	Percent (1)
Complexity	68	0.3824	0.48958	42	61.8	26	38.2
Standardization	68	0.6176	0.48958	26	38.2	42	61.8

The proportion of Complexity is 0.3824. Meanwhile, the proportion of Memorability is 0.0.6176. McNemar test was implemented to assess if the proportion of Standardization is significantly larger than Complexity, P(Standardization)>P(Complexity).

Table 4.22: McNemar Test for Item 6

Pair	Proportions		Difference (Unsigned)	McNemar Test (1-tailed)	For Discordant Cells:		
	P_A	P_B			A=1 B=0	A=0 B=1	Odds Ratio (larger/smaller)
Complexity (A) Standardization (B)	0.6176	0.3824	0.2352	0.00976	29	13	2.2308

Table 4.22 shows the result of McNemar analysis. Results indicate the proportion of Standardization is significantly larger than the proportion of Complexity p(0.010)<0.05 (p-value rounded to three decimals). Based on this result, it can be proved that only one characteristic (Standardization) is associated with the item.

4.4.2.4 Item 9: This sign is easy to understand

This sub-section analyses the proportion of CSC for Item 9, "This sign is easy to understand". Descriptive statistics in Table 4.23 presents a descriptive analysis of three characteristics (Meaningfulness, Comprehension, and Learnability) for the item "This sign is easy to understand".

Table 4.23: Meaningfulness, Comprehension, and Learnability for Item 9

Characteristics	N	Mean	Std. Deviation	Count 0	Percent (0)	Count 1	Percent (1)
Meaningfulness	68	.3971	.49293	41	60.3	27	39.7
Comprehension	68	.5882	.49581	28	41.2	40	58.8
Learnability	68	.3088	.46544	47	69.1	21	30.9

As in Table 4.23, proportions of responses in the three characteristics range from 0.309 to 0.589. Cochran's Q test indicates there is a significant difference between proportion of responses on three characteristics –Meaningfulness, Understanding, and Learnability $X^2(2) = 10.481, p = 0.005$. The proportion of the responses on the three characteristics is not equal, hence indicates the need to perform post-hoc test by implementing multiple McNemar tests with Bonferroni Correction. Bonferroni correction is

applied to reduce the possibility of type 1 error. With Bonferroni Correction, α is defined by 0.5/3=0.017. With one-tailed analysis, analysis is performed to assess if larger proportions are significant than the smaller proportions P(Comprehension)> P(Meaningfulness)> P(Learnability).

Table 4.24: McNemar Test for Item 9

Pair	Proportions		Difference (Unsigned)	McNemar Test (1-tailed)	For Discordant Cells:		
	P_A	P_B			A=1 B=0	A=0 B=1	Odds Ratio (larger/smaller)
Comprehension (A) Learnability (B)	0.5882	0.3088	0.2794	0.001689	29	10	2.9
Comprehension (A) Meaningfulness (B)	0.5882	0.3971	0.1911	0.023516	25	12	2.0833
Learnability (A) Meaningfulness (B)	0.3088	0.3971	0.0883	0.188543	13	19	1.4615

Table 4.24 shows the unsigned difference between proportions of various pairs, ranging from 0.0883 to 0.2794. Multiple McNemar result shows the proportion of Comprehension is significantly larger than Learnability p(0.002)<0.017 but shows not significantly larger than Meaningfulness p(0.024)>0.017. Meanwhile, the proportion of Meaningfulness is not significantly larger than Learnability. Based on the results, two characteristics (Comprehension and Meaningfulness) are deemed related to the item.

4.4.2.5 Item 19: The text in this sign conveys a clear meaning

This sub-section analyses the proportion of CSC for Item 19, "The text in this sign conveys a clear meaning". Descriptive statistics in Table 4.25 presents a descriptive analysis of three characteristics (Legibility, Surface properties, and Meaningfulness) for the item "The text in this sign conveys a clear meaning".

Table 4.25: Legibility, Surface Properties and Meaningfulness of Item 19

Characteristics	N	Mean	Std. Deviation	Count 0	Percent (0)	Count 1	Percent (1)
Legibility	68	.3088	.4654	47	69.1	21	30.9
Surface	68	.3529	.4814	44	64.7	24	35.3
Meaningfulness	68	.6029	.4929	27	39.7	41	60.3

As shown in Table 4.25, the proportion of responses of the three characteristics (Legibility, Surface, and Meaningfulness) ranges from 0.3088 to 0.6029. Cochran's Q test indicates there is a significant difference between proportion of responses on three characteristics – Legibility, Surface, Meaningfulness $X^2(2) = 13.170, p = 0.001$. Based on this result, it can be concluded that the proportion of the responses on the three characteristics is not equal, hence indicates the need to perform post-hoc test by implementing multiple McNemar tests with Bonferroni Correction. With Bonferroni Correction, α is defined by 0.5/3=0.017. With one-tailed analysis, analysis is performed to assess if larger proportions are significant than the smaller proportions P(Meaningfulness)>P(Surface Properties) >P(Legibility).

Table 4.26: McNemar Test for Item 19

Pair	Proportions		Difference (Unsigned)	McNemar Test (1-tailed)	For Discordant Cells:		
	P_A	P_B			A=1 B=0	A=0 B=1	Odds Ratio (larger/smaller)
Legibility (A) Surface (B)	0.3088	0.3529	0.0441	0.345019	11	14	1.2727
Legibility (A) Meaningfulness (B)	0.3088	0.6029	0.2941	0.001111	10	30	3
Surface (A) Meaningfulness (B)	0.3529	0.6029	0.25	0.005754	12	29	2.4167

As in Table 4.26, Multiple McNemar result shows the proportion of Meaningfulness is significantly larger than Legibility p(0.001)<0.017 and Surface Properties p(0.006)<0.017. Meanwhile, the proportion of Surface Properties is not significantly larger than Legibility

p(0.345)>0.017. Based on the results, there is only one characteristic (Meaningfulness) can be assigned to the item.

4.4.2.6 Item 20: The text in this sign is difficult to read

This sub-section analyses the proportion of CSC for Item 19, "The text in this sign conveys a clear meaning". Descriptive statistics in Table 4.26 presents a descriptive analysis of three characteristics (Legibility, Surface properties, and Meaningfulness) for the item "The text in this sign conveys a clear meaning".

Table 4.27: Legibility, Surface Properties, and Complexity of Item 20

Characteristics	N	Mean	Std. Dev.	Count 0	Percent (0)	Count 1	Percent (1)
Legibility	68	.3824	.48958	42	61.8	26	38.2
Surface	68	.3382	.47663	45	66.2	23	33.8
Complexity	68	.4853	.50350	35	51.5	33	48.5

As shown in Table 4.27, proportions of responses in the three characteristics range from 0.338 to 0.485. Cochran's Q test indicates there is no significant difference between proportion of responses on three characteristics – Legibility, Surface properties, and Complexity $X^2(2) = 2.724, p = 0.256$. The proportion of the responses on the three characteristics is equal, and therefore there is no difference in proportion of responses of the three characteristics on item 20.

4.4.3 Summary for MRA, Cochran's Q and McNemar tests

The purpose of Multiple Responses Analysis, Cochran's Q and McNemar tests were to assign one or more CSC to an item based on the proportion responses of 68 participants that were asked to associate CSC with any number of items. The responses were recorded and transformed into dichotomous variables using dummy codes (0- not selected, 1- selected). Table 4.28 shows the list of items and the CSC that was associated with the item.

In conclusion, each CSC was assigned to at least one item. Item 2, 4, 9, and 20 were associated with more than one CSC.

Table 4.28: CSC assigned to each item

No.	Item	Cognitive Sign Characteristics
1	I am confident that I understand the meaning of this sign.	Comprehension
2	I need help in understanding this sign for the first time.	Complexity, Priming, Learnability
3	I am familiar with this sign.	Familiarity
4	I easily forget the meaning of this sign.	Complexity, memorability
5	I would learn this sign quickly.	Learnability
6	This sign can be easily confused with another sign.	Standardization
7	This sign can easily capture my attention.	Saliency
8	This sign is complicated.	Complexity
9	This sign is easy to understand.	Comprehension, Meaningfulness
10	I have seen many variations of this sign.	Standardization
11	This sign is easy to read.	Legibility
12	The colours used in this sign are not suitable.	Surface Properties
13	The colour contrast of this sign is good.	Surface Properties
14	The shape of this sign is not suitable.	Surface Properties
15	I know what I should do when I see this sign.	Priming
16	It is likely that I will ignore this sign.	Authoritativeness
17	The icon in this sign conveys a clear meaning.	Meaningfulness
18	The icon design of this sign is not suitable.	Surface Properties
19	The text in this sign conveys a clear meaning.	Meaningfulness
20	The text in this sign is difficult to read.	Legibility, Surface properties, Complexity

4.5 Rasch Analysis

This section assesses the psychometric properties of the instrument using Rasch analysis for rating scales as explained in Sub-section 3.7.2.2. Rasch analysis was conducted using JMetrik software (version 4) developed by Meyer (2014). The assessment of instrument validity and reliability with Rasch Analysis involves the assessment of three randomly selected signs from the manual of traffic sign by The Ministry of Works, Malaysia

(Public Works Department Malaysia, 1985c, pp. 78–79). All test signs are assigned a code: The code for sign A is RP.06 p.78: No Entry for Vehicles Having Overall Height Exceeding the Specific Height (Height Limit) (text and symbol sign). The code for sign B is RP.10 p.79 No Parking (symbol-only sign), and lastly, the code for sign C is RP.05 p.78 No Entry for Vehicles Having Total Weight Exceeding the Specified Weight (text-only sign). The stimulus for this analysis is shown in Sub-section 3.7.2.3. The instrument was tested using Rasch analysis and test level statistics for all sign is shown in Table 4.29. The table shows descriptive statistics for all signs. The number of item for text + symbol sign is different because it includes items intended for both symbol and text sign. If there is a high proportion of incomplete responses, the respondesnt will be removed from the analysis. Test-level mean, and standard deviation are as follow: Sign A, RP.06 (M = 44.6, SD = 10.98), Sign B RP.10 (M = 37.67, SD = 10.10), and Sign C RP.05 (M = 37.69, SD = 9.55).

Rasch analysis of each test sign is analysed in a different section (Sign A - Sub-section 4.5.2, Sign B – Sub-section 4.5.3, and Sign C- Sub-section 4.5.3). On top of that, PCAR and DIF analysis were also performed. Lastly, Sub-section 4.5.7 provides a summary of Rasch analysis including PCAR, DIF analysis, and decision on item removal.

4.5.1 Test level statistics for all signs

Table 4.29 shows descriptive statistics for all signs. The number of item for text + symbol sign is different because it includes items intended for both symbol and text sign. If there are a high proportion of incomplete responses, the sample is filtered and removed from the analysis. Test-level mean, and standard deviation are as follow: Sign A, RP.06 (M = 44.6, SD = 10.98), Sign B RP.10 (M = 37.67, SD = 10.10), and Sign C RP.05 (M = 37.69, SD = 9.55).

Table 4.29: Test Level Statistics for All Signs

Test level statistics	Text + symbol sign Sign A (RP.06 p.78)	Symbol sign Sign B (RP.10 p.79)	Text sign Sign C(RP.05 p.78)
Number of Items	20	18	18
Examinees	198	203	192
Min	22.0000	15.0000	16.0000
Max	76.0000	64.0000	65.0000
Mean	44.6061	37.6749	37.6875
Median	43.5000	37.0000	36.0000
Standard Deviation	10.9819	10.0952	9.5452
Interquartile Range	16.0000	12.0000	14.0000
Skewness	0.3842	0.2991	0.4869
Kurtosis	-0.1191	-0.1378	-0.2388

4.5.2 Rasch Analysis of sign A (RP.06 p.78)

Table 4.30 shows item analysis summary for the text + symbol sign (sign A, RP.06).

Table 4.30: Item Analysis Summary on Text + Symbol Sign A (RP.06)

Item	Mean	Median	Mode	SD	Var	Range	Min	Max	Diff.	Discri.	Del.*	JMLE Diff.
a1	3.47	4.00	4.00	1.14	1.29	4.00	1.00	5.00	2.49	0.56	0.83	-0.25
a2	3.43	4.00	4.00	1.18	1.38	4.00	1.00	5.00	1.57	0.50	0.83	0.77
a3	3.62	4.00	4.00	1.08	1.17	4.00	1.00	5.00	2.66	0.45	0.84	0.54
a4	2.96	3.00	3.00	1.26	1.59	4.00	1.00	5.00	2.03	0.55	0.83	0.29
a5	3.61	4.00	4.00	0.99	0.98	4.00	1.00	5.00	2.67	0.37	0.84	-0.48
a6	3.07	3.00	3.00	1.17	1.36	4.00	1.00	5.00	1.95	0.51	0.83	0.39
a7	3.28	3.00	4.00	1.07	1.15	4.00	1.00	5.00	2.30	0.22	0.85	0.09
a8	2.72	3.00	2.00	1.12	1.26	4.00	1.00	5.00	2.30	0.50	0.83	0.04
a9	3.53	4.00	4.00	1.05	1.11	4.00	1.00	5.00	2.54	0.56	0.83	-0.31
a10	3.13	3.00	4.00	1.15	1.32	4.00	1.00	5.00	1.81	0.04	0.85	0.49
a11	3.61	4.00	4.00	1.04	1.08	4.00	1.00	5.00	2.64	0.49	0.83	-0.32
a12	2.47	2.00	2.00	1.08	1.16	4.00	1.00	5.00	2.54	0.32	0.84	-0.11
a13	3.87	4.00	4.00	0.86	0.74	4.00	1.00	5.00	2.85	0.30	0.84	-0.85
a14	2.61	3.00	2.00	1.04	1.08	4.00	1.00	5.00	2.38	0.44	0.84	0.03
a15	3.38	4.00	4.00	1.22	1.49	4.00	1.00	5.00	1.39	0.54	0.83	0.08
a16	3.30	3.00	4.00	1.23	1.51	4.00	1.00	5.00	0.97	0.43	0.84	0.83
a17	3.36	3.00	3.00	1.07	1.15	4.00	1.00	5.00	2.32	0.47	0.84	-0.10
a18	2.70	3.00	3.00	0.95	0.91	4.00	1.00	5.00	2.31	0.41	0.84	0.06
a19	3.40	4.00	4.00	1.06	1.11	4.00	1.00	5.00	2.43	0.45	0.84	-0.14
a20	2.57	2.00	2.00	1.09	1.20	4.00	1.00	5.00	2.43	0.38	0.84	0.01

*Refer Appendix H for full table of item deletion reliability statistics

Indicators and statistics in Table 4.30 include non-Rasch item difficulty (Diff), item discriminant (Discri.), item deletion reliability (Del.) as well as Rasch item difficulty based on joint maximum likelihood estimation (JMLE Diff in logits). Items have the same response anchors; therefore, a common threshold is set. Non-Rasch item difficulty (Diff) refers to the agreeableness for the item. For example, item 13 that records difficulty value of 2.85 is the easiest to agree with. Meanwhile, the hardest item to agree with is item 16 with difficulty value of 0.97. In other words, it refers to the tendency to endorse a specific item for all participants of which ranges from 0-4. Item discriminant (Discri.) describes the contribution of each item to the total usability scale in terms of the correlation between item response and overall scores on the instrument. It ranges from 0 to 1. Value 0 would indicate that the item response is unrelated to the overall score. This is a sign that an item does not discriminate between low and high scorers. From the result, item 10 recorded a very low score (0.04). It is an indication that the item does not correlate well with the measure.

Rasch item difficulty (JMLE Diff) value uses Joint Maximum Likelihood Estimation (JMLE) estimation method. Rasch item difficulty centres from 0 (average). A bigger positive value of JMLE Rasch item difficulty indicates the items are more difficult, that is, high usability required for agreement with the item. The most difficult item is item 16 (0.80 logit), and the easiest item is item 5 (-0.48 logit). Table 4.31 shows various reliability analyses for the instruments tested with sign A. Initial Cronbach's alpha value is high (0.84), after deletion of items, there is a slight increase in the final alpha value (0.85). The initial alpha value is already considered to be very reliable.

Table 4.31: Reliability Analysis Text + Symbol Sign (RP.06)

Coefficient	Initial			Final		
	Estimate	**95% Conf. Int.**	**SEM**	**Estimate**	**95% Conf. Int.**	**SEM**
Guttman's L2	0.8543	(0.8233, 0.8819)	4.1917	0.8616	(0.8320, 0.8881)	4.0464
Alpha	0.8439	(0.8107, 0.8735)	4.3371	0.8537	(0.8223, 0.8817)	4.1607
Feldt-Gilmer	0.8485	(0.8163, 0.8772)	4.2535	0.8576	(0.8271, 0.8849)	4.1044
Feldt-Brennan	0.8483	(0.8161, 0.8771)	4.2700	0.8574	(0.8268, 0.8847)	4.1076
Raju's Beta	0.8439	(0.8107, 0.8735)	4.3371	0.8537	(0.8223, 0.8817)	4.1607

Table 4.32 describes scale quality statistics of the instrument. In addition to Cronbach's alpha, Rasch analysis also provides item and person reliability index along with Separation index. The comprehensiveness of reliability analysis is explained in Chapter 3. The instrument elicits excellent reliability, with the item and person reliability recorded 0.96 and 0.87 respectively. Item separation index is 4.9 and person separation index is 2.6. Person separation index should be more than 2 and item separation should be more than 3 (Linacre, 2012). These values prove that persons (or items) who are estimated with high scoring indeed have higher probability do have a higher score.

Table 4.32: Rasch Reliability Statistics Text + Symbol Sign (RP.06)

Statistic	Items	Persons
Observed Variance	0.1743	0.5963
Observed Std. Deviation	0.4176	0.7722
Mean Square Error	0.0069	0.0756
Root MSE	0.0833	0.2750
Adjusted Variance	0.1674	0.5206
Adjusted Std. Deviation	0.4092	0.7216
Separation Index	4.9127	2.6240
Number of Strata	6.8837	3.8320
Reliability	0.9602	0.8732

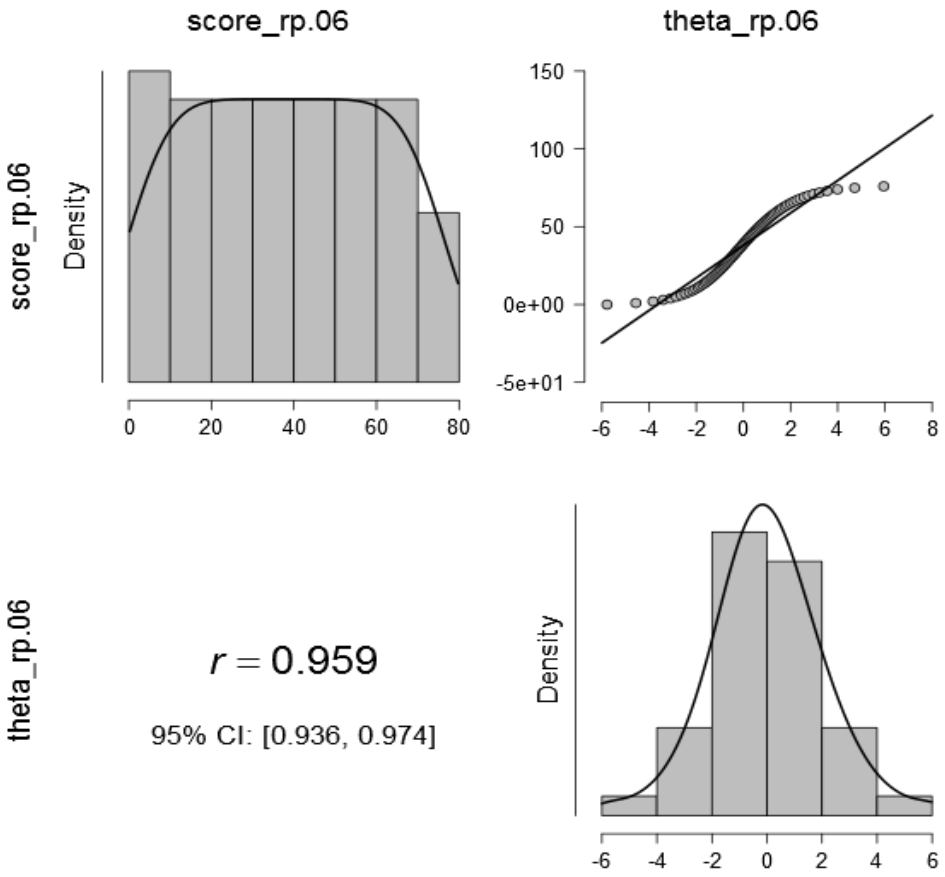


Figure 4.4: Theoretical relationship of Theta vs. Likert scale score (RP.06)

Rasch analysis creates person ability estimate that corresponds to each summed score. Figure 4.4 illustrates the theoretical relationship between summed scores and person ability estimates for sign A, RP.06. The actual relationship is dispersed because the differences in the scoring of different item combination can produce the same summated score. In Rasch model, the raw score is a sufficient statistic to estimate the ability (theta). The lowest and the highest score are considered as extreme scores even though these floor-and-ceiling values are probable. These extreme values are excluded during the JML estimation.

Table 4.33 shows the raw score, ability estimates and the standard error of ability estimate. The scores are between 0 to 76, and ability estimates range from -5.78 to 5.95. Meanwhile, standard error of ability estimates ranges from 0.24 to 1.84.

Table 4.33: Score, Ability Estimate, and Standard Error (RP.06)

RP.06

Score	Theta	Std. Err	Score	Theta	Std. Err	Score	Theta	Std. Err
0.00	-5.78	1.84	26.00	-0.76	0.25	52.00	0.82	0.27
1.00	-4.56	1.02	27.00	-0.70	0.25	53.00	0.89	0.27
2.00	-3.83	0.73	28.00	-0.64	0.24	54.00	0.97	0.27
3.00	-3.40	0.60	29.00	-0.58	0.24	55.00	1.04	0.28
4.00	-3.08	0.52	30.00	-0.52	0.24	56.00	1.12	0.28
5.00	-2.84	0.47	31.00	-0.46	0.24	57.00	1.20	0.29
6.00	-2.63	0.43	32.00	-0.40	0.24	58.00	1.28	0.29
7.00	-2.46	0.41	33.00	-0.34	0.24	59.00	1.37	0.30
8.00	-2.30	0.38	34.00	-0.29	0.24	60.00	1.46	0.30
9.00	-2.16	0.36	35.00	-0.23	0.24	61.00	1.55	0.31
10.00	-2.04	0.35	36.00	-0.17	0.24	62.00	1.65	0.32
11.00	-1.92	0.33	37.00	-0.11	0.24	63.00	1.76	0.33
12.00	-1.82	0.32	38.00	-0.05	0.24	64.00	1.87	0.34
13.00	-1.72	0.31	39.00	0.00	0.24	65.00	1.99	0.35
14.00	-1.62	0.30	40.00	0.06	0.24	66.00	2.11	0.36
15.00	-1.54	0.29	41.00	0.12	0.24	67.00	2.25	0.38
16.00	-1.45	0.29	42.00	0.18	0.24	68.00	2.40	0.40
17.00	-1.37	0.28	43.00	0.24	0.25	69.00	2.57	0.42
18.00	-1.30	0.27	44.00	0.30	0.25	70.00	2.75	0.45
19.00	-1.22	0.27	45.00	0.36	0.25	71.00	2.97	0.48
20.00	-1.15	0.27	46.00	0.42	0.25	72.00	3.23	0.53
21.00	-1.08	0.26	47.00	0.49	0.25	73.00	3.55	0.61
22.00	-1.01	0.26	48.00	0.55	0.25	74.00	3.99	0.73
23.00	-0.95	0.26	49.00	0.62	0.26	75.00	4.72	1.02
24.00	-0.88	0.25	50.00	0.68	0.26	76.00	5.95	1.84
25.00	-0.82	0.25	51.00	0.75	0.26	-	-	-

The relationship between standard errors and ability estimates is illustrated in Figure 4.5.

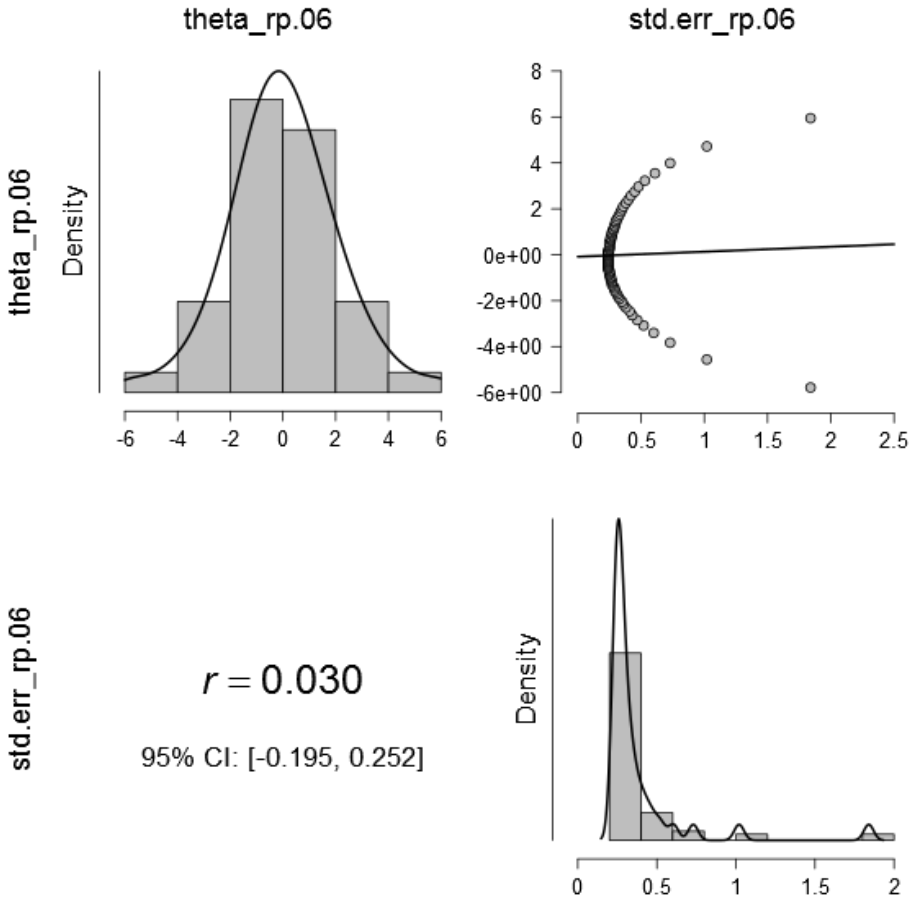


Figure 4.5: Conditional standard error and person ability (RP.06)

From Figure 4.5, it is illustrated that the standard error of ability estimate is not homogenous across person ability estimate. It is regarded as the conditional standard error as it shows the lowest value approaching the average score and ascends toward the lowest and the highest score. Table 4.34 shows weighted mean squares (WMS), and unweighted mean squares (UMS) as fit statistics for instrument tested with sign A. Unfit items are isolated and considered for removal as it decreases the quality of measure. Unfit items also indicate it might belong to another dimension. From the table, one item (item 10) shows unfit value (WMS=1.64, UMS=1.82). After removal of item 10, WMS and UMS statistics for all items fall into acceptable value.

Table 4.34*:* WMS and UMS Fit Statistics for Text + Symbol Sign (RP.06)

Item	Initial		Elimination		Item fit
	WMS	UMS	WMS	UMS	(0.5-1.5)
1	0.80	0.82	0.80	0.83	fit
2	0.98	0.99	1.01	1.01	fit
3	0.91	0.92	0.92	0.92	fit
4	0.89	0.89	0.94	0.93	fit
5	1.03	1.00	1.05	1.02	fit
6	0.93	0.95	0.98	1.00	fit
7	1.21	1.28	1.25	1.31	fit
8	0.93	0.96	0.97	1.01	fit
9	0.80	0.78	0.80	0.77	fit
10	1.64	1.82	-	-	Underfit
11	0.85	0.80	0.84	0.78	fit
12	1.18	1.19	1.28	1.30	fit
13	1.09	1.21	1.15	1.41	fit
14	1.00	1.04	1.06	1.09	fit
15	0.64	0.62	0.65	0.62	fit
16	1.15	1.17	1.21	1.29	fit
17	0.92	0.91	0.94	0.93	fit
18	0.99	1.00	1.04	1.06	fit
19	0.93	0.93	0.96	0.95	fit
20	1.03	1.05	1.09	1.11	fit

Table 4.35 records standardize weighed mean squares (Std. WMS), and standardize unweighted mean squares (Std. UMS) for sign A. The cut-off value for standardized mean squares should be no more than 3 but no less than -2 (Linacre, 2002). The removal of items based on fit statistics is an iterative process until there are no unfit items. The items that are too predictable or too unpredictable are isolated and considered for removal. From the table above, item 10 records very high standardize mean squares (Std. WMS= 5.89, Std. UMS=7.17).

Table 4.35: Item Difficulty and Std. Fit Statistics for Text + Symbol Sign (RP.06)

Item	Initial				Elimination			
	Diff.	Std. Err	Std. WMS	Std. UMS	Diff.	Std. Err	Std. WMS	Std. UMS
1	-0.25	0.07	-2.41	-2.00	-0.23	0.08	-2.37	-1.79
2	0.77	0.07	-0.19	-0.12	0.84	0.08	0.14	0.10
3	-0.54	0.08	-0.95	-0.79	-0.53	0.08	-0.92	-0.77
4	0.29	0.07	-1.31	-1.23	0.33	0.07	-0.61	-0.68
5	-0.48	0.09	0.32	0.01	-0.47	0.09	0.55	0.26
6	0.39	0.07	-0.79	-0.52	0.43	0.08	-0.22	0.04
7	0.09	0.08	2.20	2.74	0.12	0.08	2.49	2.99
8	0.04	0.08	-0.81	-0.43	0.06	0.08	-0.31	0.17
9	-0.31	0.08	-2.14	-2.46	-0.29	0.08	-2.21	-2.53
10	0.49	0.08	5.89	7.17	-	-	-	-
11	-0.32	0.08	-1.50	-2.05	-0.31	0.09	-1.67	-2.22
12	-0.11	0.08	1.77	1.81	-0.10	0.08	2.63	2.69
13	-0.85	0.09	0.86	1.90	-0.85	0.09	1.40	3.48
14	0.03	0.08	0.01	0.43	0.06	0.08	0.62	0.96
15	0.08	0.10	-3.44	-3.43	0.11	0.10	-3.30	-3.30
16	0.83	0.10	1.30	1.36	0.89	0.10	1.73	2.20
17	-0.10	0.08	-0.82	-1.00	-0.07	0.08	-0.59	-0.78
18	0.06	0.09	-0.11	0.08	0.09	0.09	0.44	0.62
19	-0.14	0.08	-0.70	-0.76	-0.12	0.08	-0.43	-0.48
20	0.01	0.08	0.38	0.56	0.03	0.08	0.97	1.13

4.5.3 Rasch Analysis of sign B (RP.10 p.79)

In Table 4.36, item analysis summary on sign B (RP.10 p.79) is presented. Items have the same response anchors; therefore, a common threshold is set. Non-Rasch item difficulty (Diff) shows there are two items that are considered easy items. Item 3 and item 5 have the difficulty value of 2.59. Meanwhile, the hardest item to agree with is item 16 with difficulty value of 1.21. Item discriminant (Discri.) describes the contribution of each item to the total usability scale. In terms of item discriminant, item 10 recorded a very low score (-0.06). This result is consistent with the result for sign A. It is another indication that the item does not correlate well with the measure. Based on JMLE Rasch item difficulty, the most difficult item is item 6 (0.72), and the easiest item is item 5 (-0.47).

Table 4.36: Item Analysis Summary on Symbol-only Sign (RP.10)

Item	Mean	Median	Mode	SD	Var	Range	Min	Max	Diff.	Discri.	Del.*	JMLE Diff.
b1	3.32	3.00	4.00	1.21	1.45	4.00	1.00	5.00	2.30	0.60	0.83	-0.09
b2	3.42	4.00	4.00	1.17	1.36	4.00	1.00	5.00	1.55	0.43	0.84	0.57
b3	3.60	4.00	4.00	1.14	1.29	4.00	1.00	5.00	2.59	0.53	0.83	-0.4
b4	3.10	3.00	3.00	1.16	1.36	4.00	1.00	5.00	1.88	0.42	0.84	0.38
b5	3.60	4.00	4.00	1.01	1.02	4.00	1.00	5.00	2.59	0.48	0.83	-0.47
b6	3.50	4.00	4.00	1.16	1.34	4.00	1.00	5.00	1.48	0.46	0.84	0.72
b7	3.39	4.00	4.00	1.08	1.16	4.00	1.00	5.00	2.37	0.42	0.84	-0.25
b8	2.73	3.00	2.00	1.09	1.18	4.00	1.00	5.00	2.26	0.44	0.84	0.03
b9	3.37	4.00	4.00	1.08	1.17	4.00	1.00	5.00	2.35	0.60	0.83	-0.05
b10	3.24	3.00	3.00	1.11	1.23	4.00	1.00	5.00	1.75	-0.06	0.83	0.42
b11	3.44	4.00	4.00	1.05	1.11	4.00	1.00	5.00	2.42	0.60	0.83	-0.24
b12	2.75	3.00	2.00	1.03	1.06	4.00	1.00	5.00	2.23	0.32	0.84	0.07
b13	3.53	4.00	4.00	1.00	1.00	4.00	1.00	5.00	2.51	0.29	0.84	-0.38
b14	2.71	3.00	2.00a	0.92	0.84	4.00	1.00	5.00	2.28	0.28	0.84	0.02
b15	3.38	4.00	4.00	1.22	1.49	4.00	1.00	5.00	1.34	0.48	0.84	-0.1
b16	2.88	3.00	2.00	1.18	1.39	4.00	1.00	5.00	1.21	0.45	0.84	0.13
b17	3.23	3.00	3.00	1.09	1.18	4.00	1.00	5.00	2.22	0.52	0.83	-0.07
b18	2.64	3.00	3.00	0.89	0.80	4.00	1.00	5.00	2.34	0.27	0.84	-0.28

*Refer Appendix H for full table of item deletion reliability statistics

Table 4.37 shows various reliability indicators for sign B. Initial Cronbach's alpha value is high (0.82). After the deletion of items, there is a slight increase in the final alpha value (0.84). The initial alpha value is already considered to be very reliable.

Table 4.37: Reliability Analysis on Symbol-only Sign (RP.10)

Coefficient	Initial			Final		
	Estimate	95% Conf. Int.	SEM	Estimate	95% Conf. Int.	SEM
Guttman's L2	0.8426	(0.8093, 0.8725)	4.0050	0.8543	(0.8233, 0.8819)	3.8559
Alpha	0.8281	(0.7917, 0.8607)	4.1852	0.8439	(0.8107, 0.8735)	3.9909
Feldt-Gilmer	0.8363	(0.8016, 0.8673)	4.0849	0.8485	(0.8163, 0.8772)	3.9319
Feldt-Brennan	0.8346	(0.7996, 0.8660)	4.1055	0.8483	(0.8161, 0.8771)	3.9344
Raju's Beta	0.8281	(0.7917, 0.8607)	4.1852	0.8439	(0.8107, 0.8735)	3.9909

Table 4.38 describes scale quality statistics of the instrument. The instrument elicits excellent reliability, with the item and person reliability recorded 0.94 and 0.87 respectively. Item separation index is 4.0 and person separation index is 2.6. Person separation index

should be more than 2 and item separation should be more than 3 (Linacre, 2012). These values prove that persons (or items) who are estimated with high scoring indeed have a higher probability to have a higher score.

Table 4.38: Rasch Reliability Statistics for Symbol-only Sign (RP.10)

Statistic	Items	Persons
Observed Variance	0.1163	0.6415
Observed Std. Deviation	0.3410	0.8009
Mean Square Error	0.0068	0.0838
Root MSE	0.0827	0.2895
Adjusted Variance	0.1094	0.5577
Adjusted Std. Deviation	0.3308	0.7468
Separation Index	4.0001	2.5796
Number of Strata	5.6668	3.7729
Reliability	0.9412	0.8694

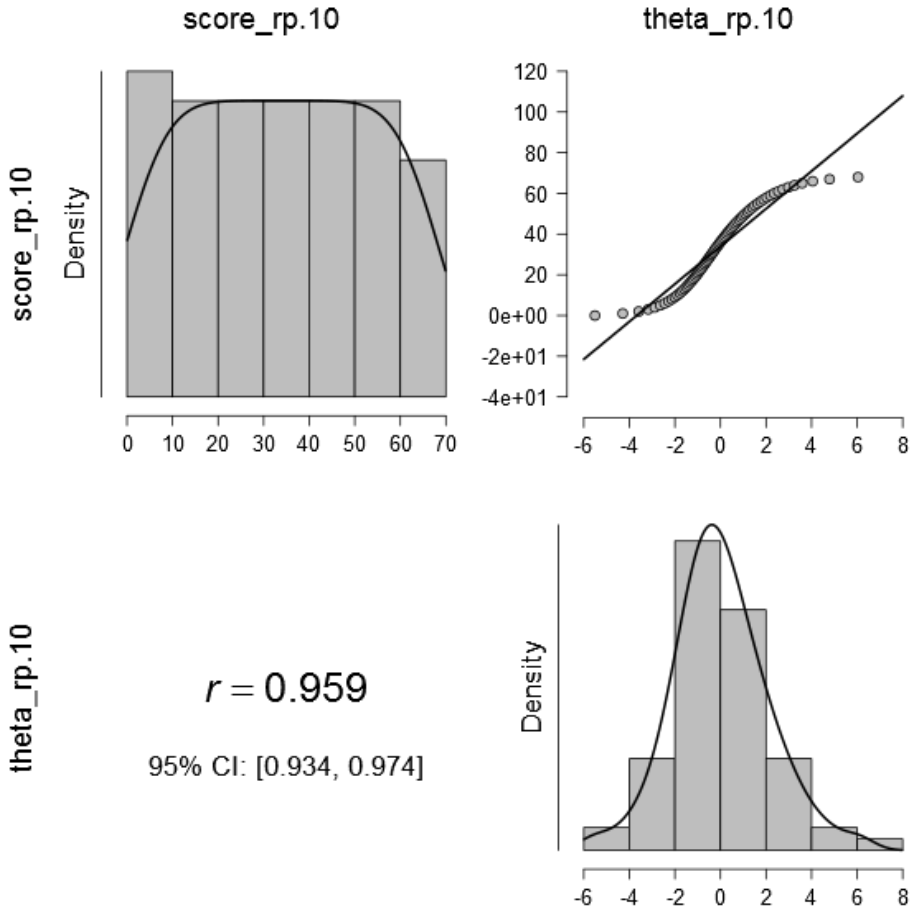


Figure 4.6: Theoretical relationship of Theta vs. Likert scale score (RP.10)

Rasch analysis creates person ability estimates that correspond to each summed score. Figure 4.6 illustrates the theoretical relationship between summed scores and person ability estimates for sign B, RP.10. The graphs have a similar pattern to graphs for sign A. This similarity is expected in the graphs for sign C, Rp.05. Table 4.39 shows the raw score, ability estimates and the standard error of ability estimate for sign B, RP.10. The Likert scales are scored between 0 to 68 which is a smaller range as compared to sign A, RP.06. The ability estimates range from -5.51 to 6.04. Meanwhile, the standard error of ability estimates range from 0.25 to 1.84.

Table 4.39: Scores, Ability Estimate, and Standard Error (RP.10)

RP.10								
Score	**Theta**	**Std. Err**	**Score**	**Theta**	**Std. Err**	**Score**	**Theta**	**Std. Err**
0.00	-5.51	1.83	23.00	-0.84	0.26	46.00	0.73	0.29
1.00	-4.29	1.01	24.00	-0.78	0.25	47.00	0.82	0.29
2.00	-3.59	0.71	25.00	-0.71	0.25	48.00	0.91	0.30
3.00	-3.17	0.59	26.00	-0.65	0.25	49.00	1.00	0.30
4.00	-2.88	0.51	27.00	-0.59	0.25	50.00	1.09	0.31
5.00	-2.64	0.46	28.00	-0.52	0.25	51.00	1.19	0.32
6.00	-2.45	0.42	29.00	-0.46	0.25	52.00	1.29	0.32
7.00	-2.28	0.39	30.00	-0.40	0.25	53.00	1.40	0.33
8.00	-2.14	0.37	31.00	-0.33	0.25	54.00	1.51	0.34
9.00	-2.01	0.35	32.00	-0.27	0.25	55.00	1.63	0.35
10.00	-1.89	0.34	33.00	-0.21	0.25	56.00	1.76	0.36
11.00	-1.78	0.32	34.00	-0.14	0.26	57.00	1.89	0.37
12.00	-1.68	0.31	35.00	-0.07	0.26	58.00	2.03	0.38
13.00	-1.59	0.30	36.00	-0.01	0.26	59.00	2.19	0.40
14.00	-1.50	0.29	37.00	0.06	0.26	60.00	2.35	0.42
15.00	-1.42	0.29	38.00	0.13	0.26	61.00	2.54	0.44
16.00	-1.34	0.28	39.00	0.20	0.27	62.00	2.74	0.47
17.00	-1.26	0.28	40.00	0.27	0.27	63.00	2.98	0.50
18.00	-1.19	0.27	41.00	0.34	0.27	64.00	3.25	0.55
19.00	-1.11	0.27	42.00	0.42	0.27	65.00	3.59	0.62
20.00	-1.04	0.26	43.00	0.49	0.28	66.00	4.05	0.74
21.00	-0.98	0.26	44.00	0.57	0.28	67.00	4.79	1.03
22.00	-0.91	0.26	45.00	0.65	0.29	68.00	6.04	1.84

The relationship between standard errors and ability estimates is illustrated in Figure 4.7. It shows the same pattern as sign A (RP.06).

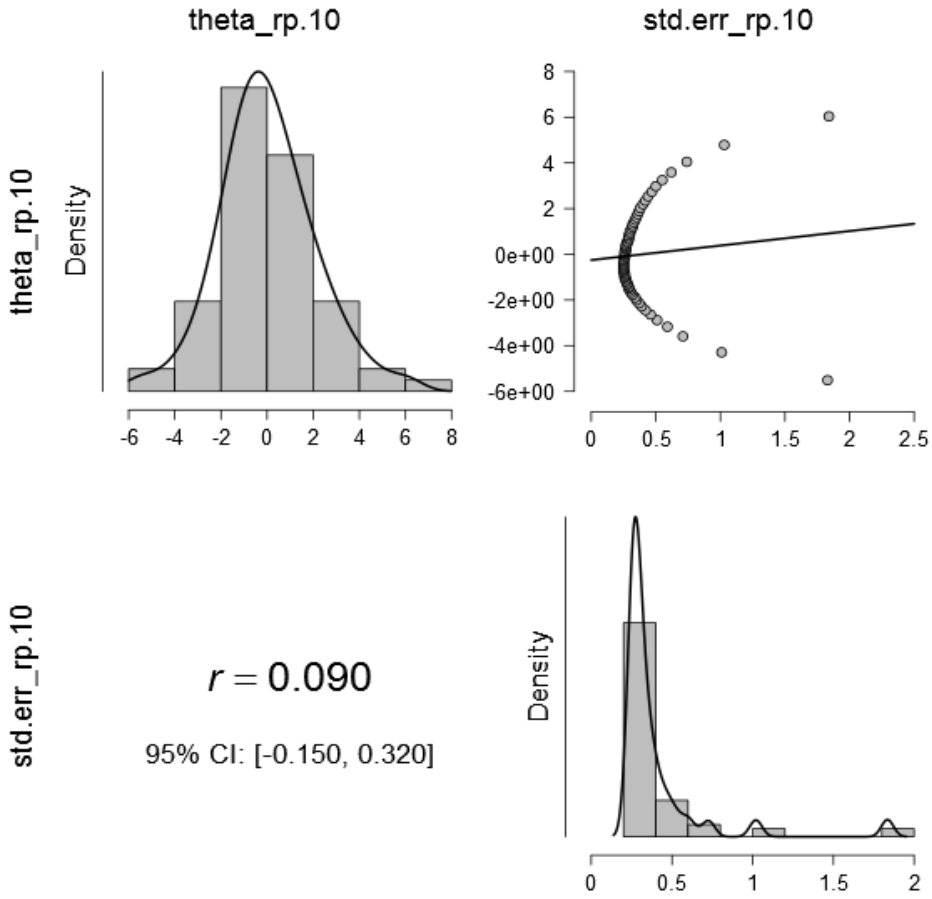


Figure 4.7: Conditional standard error and person ability (RP.10)

Table 4.40 shows weighted mean squares (WMS), and unweighted mean squares (UMS) fit statistics for sign B. From the table, one item (item 10) shows unfit value (WMS=1.77, UMS=1.95). After removal of item 10, WMS and UMS statistics for all items fall into acceptable value.

Table 4.40: WMS and UMS fit Statistics for Symbol-only Sign (RP.10)

Item	Initial		Elimination		Item fit
	WMS	UMS	WMS	UMS	0.5-1.5
1	0.74	0.72	0.76	0.74	fit
2	1.11	1.14	1.17	1.19	fit
3	0.83	0.87	0.83	0.92	fit
4	1.05	1.06	1.13	1.14	fit
5	0.87	0.81	0.88	0.83	fit
6	1.02	1.03	1.13	1.15	fit
7	0.99	0.98	1.03	1.02	fit
8	0.98	0.96	1.06	1.03	fit
9	0.75	0.73	0.75	0.73	fit
10	1.77	1.95	-	-	Underfit
11	0.76	0.75	0.76	0.75	fit
12	1.13	1.11	1.19	1.18	fit
13	1.15	1.18	1.23	1.31	fit
14	1.12	1.15	1.19	1.23	fit
15	0.67	0.66	0.68	0.66	fit
16	0.82	0.79	0.88	0.86	fit
17	0.86	0.89	0.89	0.94	fit
18	1.14	1.16	1.23	1.28	fit

Table 4.41: Item Difficulty and Std. Fit Statistics for Symbol-only Sign (RP.10)

Item	Initial				Elimination			
	Diff.	Std. Err	Std. WMS	Std. UMS	Diff.	Std. Err	Std. WMS	Std. UMS
1	-0.09	0.07	-3.19	-3.37	-0.08	0.07	-2.82	-3.06
2	0.57	0.07	1.17	1.43	0.64	0.08	1.69	1.87
3	-0.40	0.07	-1.89	-1.32	-0.40	0.08	-1.86	-0.79
4	0.38	0.07	0.65	0.72	0.42	0.08	1.46	1.48
5	-0.47	0.08	-1.40	-2.01	-0.47	0.08	-1.22	-1.82
6	0.72	0.08	0.21	0.34	0.79	0.08	1.33	1.50
7	-0.25	0.08	-0.09	-0.21	-0.24	0.08	0.34	0.25
8	0.03	0.08	-0.16	-0.41	0.06	0.08	0.62	0.39
9	-0.05	0.08	-2.77	-2.90	-0.03	0.08	-2.74	-2.91
10	0.42	0.08	6.95	8.29	-	-	-	-
11	-0.24	0.08	-2.70	-2.79	-0.23	0.08	-2.67	-2.81
12	0.07	0.08	1.39	1.17	0.09	0.08	1.95	1.81
13	-0.38	0.08	1.49	1.74	-0.38	0.08	2.18	2.93
14	0.02	0.09	1.20	1.42	0.05	0.09	1.81	2.12
15	-0.10	0.09	-3.05	-2.99	-0.08	0.10	-2.98	-2.92
16	0.13	0.10	-1.65	-1.77	0.17	0.10	-1.00	-1.15
17	-0.07	0.08	-1.52	-1.17	-0.05	0.08	-1.19	-0.62
18	-0.28	0.09	1.38	1.65	-0.26	0.09	2.26	2.65

Table 4.41 records standardize weighed mean squares (Std. WMS) and standardize unweighted mean squares (Std. UMS) for sign B. The complete list of items that are too predictable and too unpredictable is presented in Sub-section 4.5.7. From Table 4.41, item 10 records very high standardize mean squares (Std. WMS= 5.89, Std. UMS=7.17).

4.5.4 Rasch Analysis of sign C (RP.05 p.78)

In Table 4.42, the item analysis summary on sign C (RP.05 p.78) is presented. All items have the same response anchors; therefore, a common threshold is set. Item 13 is the easiest item with non-Rasch item difficulty (Diff) of 2.73. Meanwhile, the hardest item to agree with is item 16 with difficulty value of 1.23. This result is consistent with sign A and sign B. In terms of item discriminant, item 10 has consistently recorded a very low score (-0.20). Negative loaded item on item discriminant should be removed. Based on JMLE Rasch item difficulty, the most difficult item is item 2 (0.65), and the easiest item is item 13 (-0.74).

Table 4.42: Item Analysis Summary on Text-only Sign (RP.05)

Item	Mean	Median	Mode	SD	Var	Range	Min	Max	Diff.	Discri.	Del.*	JMLE Diff.
c1	3.16	3.00	4.00	1.38	1.91	4.00	1.00	5.00	1.97	0.55	0.77	0.21
c2	3.56	4.00	4.00	1.21	1.45	4.00	1.00	5.00	1.45	0.50	0.78	0.65
c3	3.19	3.00	4.00	1.23	1.51	4.00	1.00	5.00	2.26	0.42	0.79	-0.02
c4	3.08	3.00	3.00	1.15	1.33	4.00	1.00	5.00	1.96	0.50	0.78	0.18
c5	3.38	4.00	4.00	1.09	1.19	4.00	1.00	5.00	2.41	0.34	0.79	-0.11
c6	2.94	3.00	3.00	1.09	1.18	4.00	1.00	5.00	2.06	0.37	0.79	0.11
c7	3.17	3.00	3.00	1.00	1.01	4.00	1.00	5.00	2.19	0.40	0.79	0.06
c8	2.84	3.00	2.00	1.17	1.36	4.00	1.00	5.00	2.19	0.40	0.79	0.00
c9	3.13	3.00	4.00	1.15	1.31	4.00	1.00	5.00	2.31	0.45	0.78	-0.23
c10	2.92	3.00	3.00	1.01	1.03	4.00	1.00	5.00	2.03	-0.20	0.82	-0.02
c11	3.29	4.00	4.00	1.15	1.33	4.00	1.00	5.00	2.31	0.51	0.78	-0.09
c12	2.62	3.00	2.00	1.09	1.18	4.00	1.00	5.00	2.35	0.31	0.79	-0.2
c13	3.73	4.00	4.00	0.88	0.77	4.00	1.00	5.00	2.73	0.28	0.79	-0.74
c14	2.57	3.00	3.00	0.95	0.89	4.00	1.00	5.00	2.42	0.40	0.79	-0.26
c15	3.15	3.00	4.00	1.31	1.72	4.00	1.00	5.00	1.24	0.44	0.78	0.2
c16	3.30	3.00	4.00	1.22	1.49	4.00	1.00	5.00	1.23	0.25	0.80	0.28
c17	3.20	3.00	3.00	1.07	1.14	4.00	1.00	5.00	2.23	0.49	0.78	0.03
c18	2.68	3.00	2.00	1.12	1.25	4.00	1.00	5.00	2.32	0.37	0.79	-0.07

*Refer Appendix H for full table of item deletion reliability statistics

Table 4.43 shows various reliability indicators for the instruments tested with sign C. Initial Cronbach's alpha value is high (0.79). After item deletion, there is a slight increase in the final alpha value (0.82). Cronbach alpha values are consistently high for all test signs (>.8).

Table 4.43: Reliability analysis on Text-only Sign (RP.05)

Coefficient	Initial			Final		
	Estimate	95% Conf. Int.	SEM	Estimate	95% Conf. Int.	SEM
Guttman's L2	0.8150	(0.7746, 0.8510)	4.1057	0.8326	(0.7960, 0.8653)	3.8013
Alpha	0.7978	(0.7537, 0.8372)	4.2919	0.8228	(0.7840, 0.8574)	3.9113
Feldt-Gilmer	0.8074	(0.7654, 0.8449)	4.1888	0.8271	(0.7892, 0.8608)	3.8640
Feldt-Brennan	0.8054	(0.7630, 0.8433)	4.2107	0.8270	(0.7891, 0.8607)	3.8651
Raju's Beta	0.7978	(0.7537, 0.8372)	4.2919	0.8228	(0.7840, 0.8574)	3.9113

Table 4.44 describes scale quality statistics of the instrument. The instrument elicits excellent reliability (item reliability = 0.93, person reliability = 0.86, item separation index = 3.6, person separation index = 2.5). These values prove that persons (or items) who are estimated with high scoring indeed have a higher probability to have a higher score. All reliability indicators are consistently high across all sign. This shows the instrument is very reliable for all types of sign.

Table 4.44: Rasch Reliability Statistics for Text-only Sign (RP.05)

Statistic	Items	Persons
Observed Variance	0.0936	0.5998
Observed Std. Deviation	0.3059	0.7744
Mean Square Error	0.0066	0.0830
Root MSE	0.0810	0.2880
Adjusted Variance	0.0870	0.5168
Adjusted Std. Deviation	0.2950	0.7189
Separation Index	3.6425	2.4959
Number of Strata	5.1899	3.6612
Reliability	0.9299	0.8617

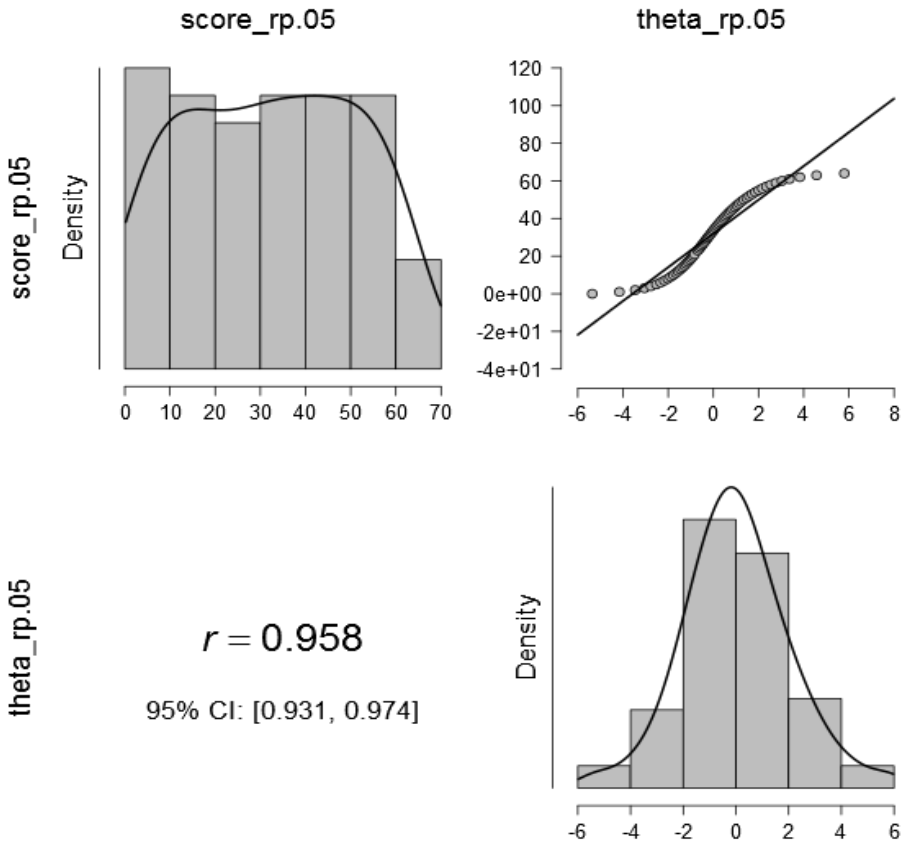


Figure 4.8: Theoretical relationship of Theta vs. Likert scale score (RP.05)

Figure 4.8 illustrates the theoretical relationship between summed scores and person ability estimates for sign C, RP.05. The graphs conform to the theoretical expectation as the distributions exhibit a similar pattern with graphs for sign A and sign B.

Table 4.45 shows the raw score, ability estimates and the standard error of ability estimate for sign C, RP.05. The score ranges from 0 to 64, which is the smallest range among all test signs. The ability estimates range from -5.36 to 5.80. Meanwhile, standard error of ability estimates ranges from 0.25 to 1.84, which is identical to the range of standard error for sign B, RP.10.

Table 4.45: Score, Ability Estimate, and Standard Error (RP.05)

RP.05								
Score	**Theta**	**Std. Err**	**Score**	**Theta**	**Std. Err**	**Score**	**Theta**	**Std. Err**
0.00	-5.36	1.83	23.00	-0.69	0.26	45.00	0.86	0.30
1.00	-4.16	1.00	24.00	-0.62	0.26	46.00	0.95	0.31
2.00	-3.46	0.71	25.00	-0.56	0.26	47.00	1.04	0.31
3.00	-3.05	0.58	26.00	-0.49	0.25	48.00	1.14	0.32
4.00	-2.76	0.51	27.00	-0.43	0.25	49.00	1.24	0.33
5.00	-2.52	0.46	28.00	-0.36	0.25	50.00	1.35	0.33
6.00	-2.33	0.42	29.00	-0.30	0.25	51.00	1.47	0.34
7.00	-2.17	0.39	30.00	-0.24	0.25	52.00	1.59	0.35
8.00	-2.02	0.37	31.00	-0.17	0.25	53.00	1.72	0.37
9.00	-1.89	0.35	32.00	-0.11	0.25	54.00	1.86	0.38
10.00	-1.77	0.34	33.00	-0.04	0.26	55.00	2.01	0.39
11.00	-1.66	0.33	34.00	0.02	0.26	56.00	2.17	0.41
12.00	-1.56	0.32	35.00	0.09	0.26	57.00	2.35	0.43
13.00	-1.46	0.31	36.00	0.16	0.26	58.00	2.54	0.46
14.00	-1.37	0.30	37.00	0.23	0.26	59.00	2.77	0.50
15.00	-1.28	0.29	38.00	0.30	0.27	60.00	3.04	0.54
16.00	-1.20	0.28	39.00	0.37	0.27	61.00	3.37	0.62
17.00	-1.12	0.28	40.00	0.45	0.27	62.00	3.83	0.74
18.00	-1.04	0.27	41.00	0.52	0.28	63.00	4.57	1.02
19.00	-0.97	0.27	42.00	0.60	0.28	64.00	5.80	1.84
20.00	-0.90	0.27	43.00	0.68	0.29			
21.00	-0.83	0.26	44.00	0.77	0.29			

The relationship between standard errors and ability estimates is illustrated in Figure 4.9. All test sign shows similar patterns of distribution.

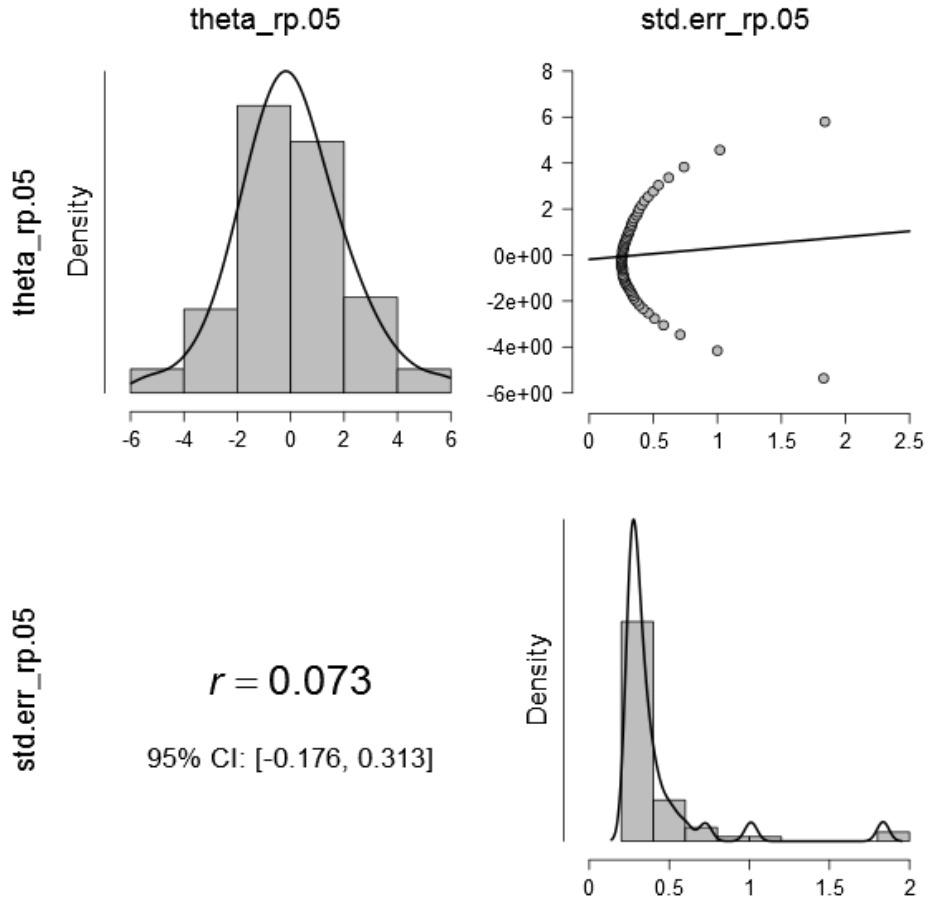


Figure 4.9: Conditional standard error and person ability (RP.05)

Table 4.46 shows weighted mean squares (WMS) and unweighted mean squares (UMS) as fit statistics for instrument tested with sign C. From the table, one item (item 10) shows unfit value (WMS=1.73, UMS=1.85). After removal of item 10, item 16 recorded unfit UMS value (1.70). After removal of two items, all fit statistics are acceptable.

Table 4.46: WMS and UMS fit Statistics for Text-only Sign (RP.05)

Item	Initial		1st elimination		2nd elimination		Item fit
	WMS	UMS	WMS	UMS	WMS	UMS	0.5-1.5
1	0.84	0.81	0.86	0.83	0.88	0.86	fit
2	0.98	1.03	1.03	1.10	1.05	1.16	fit
3	0.93	0.96	0.94	1.00	0.94	1.01	fit
4	0.86	0.87	0.92	0.92	0.94	0.94	fit
5	1.00	1.10	1.03	1.16	1.06	1.30	fit
6	1.03	1.02	1.11	1.10	1.14	1.14	fit
7	0.92	0.93	0.94	0.95	0.96	0.98	fit
8	1.01	1.00	1.07	1.06	1.10	1.08	fit
9	0.92	0.90	0.95	0.93	0.96	0.94	fit
10	1.73	1.85	-	-	-	-	Underfit
11	0.79	0.77	0.79	0.76	0.80	0.77	fit
12	1.11	1.14	1.20	1.23	1.23	1.27	fit
13	1.06	1.04	1.11	1.09	1.12	1.09	fit
14	0.97	0.95	1.05	1.02	1.07	1.04	fit
15	0.72	0.71	0.74	0.72	0.73	0.72	fit
16	1.36	1.51	1.47	1.70	-	-	Underfit
17	0.83	0.93	0.83	0.92	0.83	0.91	fit
18	0.98	1.07	1.04	1.14	1.07	1.18	fit

Table 4.47: Item Difficulty and Std. Fit Statistics for Text-only Sign (RP.05)

Item	Initial				1st Elimination				2nd Elimination			
	Diff.	Std. Error	Std. WMS	Std. UMS	Diff.	Std. Error	Std. WMS	Std. UMS	Diff.	Std. Error	Std. WMS	Std. UMS
1	0.21	0.06	-2.02	-2.16	0.22	0.06	-1.66	-1.80	0.23	0.07	-1.42	-1.47
2	0.65	0.07	-0.15	0.30	0.70	0.08	0.33	1.04	0.74	0.08	0.52	1.58
3	-0.02	0.07	-0.81	-0.35	-0.02	0.07	-0.64	0.06	-0.01	0.07	-0.64	0.09
4	0.18	0.08	-1.50	-1.43	0.19	0.08	-0.87	-0.82	0.22	0.08	-0.57	-0.58
5	-0.11	0.08	0.05	0.99	-0.13	0.08	0.32	1.48	-0.11	0.08	0.64	2.66
6	0.11	0.08	0.31	0.27	0.12	0.08	1.13	1.07	0.14	0.08	1.46	1.48
7	0.06	0.08	-0.83	-0.76	0.07	0.09	-0.61	-0.49	0.10	0.09	-0.34	-0.14
8	0.00	0.07	0.15	0.01	0.00	0.08	0.79	0.62	0.02	0.08	1.09	0.89
9	-0.23	0.08	-0.90	-1.10	-0.23	0.08	-0.51	-0.69	-0.22	0.08	-0.38	-0.58
10	-0.02	0.08	6.18	6.98	-	-	-	-	-	-	-	-
11	-0.09	0.08	-2.38	-2.61	-0.10	0.08	-2.37	-2.59	-0.08	0.08	-2.28	-2.46
12	-0.20	0.08	1.12	1.47	-0.21	0.08	1.98	2.29	-0.20	0.08	2.24	2.61
13	-0.74	0.09	0.61	0.44	-0.78	0.09	1.04	0.88	-0.77	0.10	1.08	0.91
14	-0.26	0.09	-0.26	-0.48	-0.27	0.09	0.47	0.19	-0.25	0.09	0.70	0.39
15	0.20	0.08	-2.72	-2.68	0.19	0.09	-2.51	-2.50	0.21	0.09	-2.60	-2.48
16	0.28	0.09	2.94	3.94	0.30	0.09	3.70	5.04	-	-	-	-
17	0.03	0.08	-1.82	-0.68	0.03	0.08	-1.83	-0.76	0.05	0.08	-1.86	-0.91
18	-0.07	0.08	-0.20	0.72	-0.08	0.08	0.46	1.43	-0.07	0.08	0.74	1.73

Table 4.47 records standardize weighed mean squares (Std. WMS) and standardize unweighted mean squares (Std. UMS) for sign C. The complete list of items that are too predictable and too unpredictable is presented in Table 4.47. From Table 4.47, item 10 records very high standardize mean squares (Std. WMS= 6.18, Std. UMS=6.98). For another removed item, item 16, The Std. WMS value is 3.70 and Std. UMS value is 5.04, which also deviate greatly from the acceptable range.

4.5.5 PCAR analysis and paired t-test of item subsets

Although Item discriminant and fit statistics provide evidence of unidimensionality in the measure, a more sensitive measure is implemented. Principal Components Analysis (pCA) of Standardized Residuals (PCAR) is used to assess the dimensionality of the instrument. The analysis is performed on all instruments (sign A, B, and C) and conclusive evidence is drawn based on the analysis of all three results. In addition to the eigenvalue, PCAR categorized items as positive or negative loading on the first component. The items are then extracted and grouped based on polarity. The paired t-test is performed to assess difference for these person estimates of item subsets from PCAR. The results are summarized in Table 4.48 below.

Table 4.48: PCAR Analysis and Paired T-test of Item Subsets

Sign type	**Eigenvalue <3**	**Paired t-test for person estimates of item subsets from PCAR**						
		Mean 1	SD 1	Mean 2	SD 2	Paired Correlation	Sig.	Cohen's *d*
Text + Symbol	2.90	0.49	1.05	0.45	1.14	0.43	0.622	0.035
Symbol	2.66	0.43	1.35	-0.03	0.88	0.47	<0.001	0.448
Text	2.99	0.23	1.07	0.30	1.34	0.34	0.513	-0.048

As in Table 4.48, Text + symbol sign shows no significant difference in the person estimates for positively loaded item subsets (M=0.49, SD=1.05) and negatively loaded item

subsets (M=0.45, SD=1.14) conditions; t(197)=0.494, p = 0.622 with a trivial effect size of 0.035. Notable difference was detected in the person estimates for positively loaded item subsets in symbol-only sign (M=0.43, SD=1.35) and negatively loaded item subsets (M=-0.03, SD=0.88) conditions; t(202)= 6.246, p <0.001 but with a small effect size of 0.448. Lastly, for the text-only sign, there was no significant difference in the person estimates for positively loaded item subsets (M=0.23, SD=1.07) and negatively loaded item subsets (M=0.30, SD=1.34) conditions; t(191)=-0.656 p = 0.513, with a trivial effect size of -0.048. Although significant difference observed between two item subsets for the symbol-only sign, an insignificant difference between item subsets in text + symbol and text-only sign suggest stronger evidence for unidimensionality rather than the existence of underlying second dimension. This claim is supported as all three types of sign indicate eigenvalue of first Rasch residual ranges from 2.66-2.99, which is less than the suggested cut-off point of 3.

4.5.6 Differential Item Functioning

To investigate if the items have a discriminative effect on a different group of participants, differential item functioning analysis is performed on two individual characteristics: gender and license status. Table 4.49 and Table 4.50 show χ^2 value, p-value and effect size of DIF analysis based on effect size calculated with 95% confident interval. A non-significant p-value is desirable as it indicates the item does not discriminate between two groups.

Table 4.49: DIF Analysis on Gender for All Signs

No	χ^2			p-value			Effect Size (95% C.I.)		
	RP.06	RP.10	RP.05	RP.06	RP.10	RP.05	RP.06	RP.10	RP.05
1	0.02	6.17	0.03	0.90	0.01	0.86	-0.00 (-0.52,0.51)	0.48 (0.03,0.92)	-0.04 (-0.54,0.46)
2	0.00	0.71	0.52	0.96	0.40	0.47	-0.04 (-0.31,0.23)	0.16 (-0.09, 0.40)	0.10 (-0.11,0.31)
3	0.80	0.01	7.77	0.37	0.91	0.01	-0.27 (-0.78,0.23)	-0.09 (-0.59,0.41)	0.65 (0.18,1.11)
4	0.00	4.79	0.05	0.96	0.03	0.82	0.05 (-0.25,0.35)	0.37 (0.04,0.70)	0.05 (-0.23,0.32)
5	2.16	0.23	0.14	0.14	0.63	0.71	-0.26 (-0.65,0.13)	0.04 (-0.36,0.43)	-0.13 (-0.54,0.28)
6	0.33	0.19	1.08	0.56	0.66	0.30	0.17 (-0.12,0.45)	0.03 (-0.23,0.28)	-0.11 (-0.32,0.10)
7	0.05	0.12	0.28	0.82	0.73	0.60	0.02 (-0.40, 0.44)	-0.15 (-0.55,0.25)	0.11 (-0.31,0.53)
8	0.00	0.56	0.33	0.98	0.45	0.56	0.04 (-0.22,0.29)	0.09 (-0.16,0.34)	-0.13 (-0.43,0.18)
9	0.09	0.30	0.65	0.77	0.58	0.42	0.03 (-0.34,0.39)	-0.15 (-0.48,0.17)	0.09 (-0.20,0.39)
10	2.55	0.02	0.14	0.11	0.89	0.71	-0.20 (-0.52,0.12)	0.01 (-0.26,0.28)	-0.02 (-0.30,0.27)
11	0.44	0.61	0.03	0.51	0.44	0.87	-0.21 (-0.61,0.20)	0.15 (-0.23,0.53)	0.00 (-0.37,0.37)
12	0.25	0.22	0.00	0.62	0.64	0.95	-0.08 (-0.39,0.22)	0.07 (-0.21,0.34)	0.03 (-0.24,0.29)
13	0.23	0.63	1.86	0.63	0.43	0.17	-0.17 (-0.63,0.29)	0.12 (-0.35,0.60)	-0.35 (-0.83,0.12)
14	6.35	0.00	0.01	0.01	0.95	0.94	-0.26 (-0.52,-0.01	0.03 (-0.25,0.31)	0.00 (-0.25,0.26)
15	1.00	3.24	2.73	0.32	0.07	0.10	-0.11 (-0.38,0.17)	0.28 (-0.01,0.57)	0.22 (-0.06,0.50)
16	0.56	0.18	0.03	0.45	0.67	0.87	-0.07 (-0.22,0.09)	-0.03 (-0.21,0.15)	-0.01 (-0.17,0.15)
17	0.32	0.80	-	0.57	0.37	-	-0.17 (-0.55,0.22)	0.13 (-0.22,0.47)	-
18	0.77	0.53	-	0.38	0.47	-	-0.11 (-0.36,0.14)	-0.10 (-0.35,0.15)	-
19	0.23	-	1.00	0.63	-	0.32	0.05 (-0.24,0.34)	-	-0.17 (-0.53,0.18)
20	0.34	-	0.46	0.56	-	0.50	-0.08 (-0.34,0.18)	-	0.10 (-0.14,0.35)

Table 4.49 shows results for DIF of gender and Table 4.50 shows DIF of license status. These two tables compile DIF for 3 test signs to ensure DIF detected is not due to chance. Effect size is useful to flag the discriminative item and show the severity of DIF. The patterns of responses based on the category are plotted in the form of graphs.

Table 4.50: DIF Analysis on License Status for All Signs

No	χ^2			p-value			Effect Size (95% C.I.)		
	RP.06	RP.10	RP.05	RP.06	RP.10	RP.05	RP.06	RP.10	RP.05
1	0.98	1.25	0.25	0.32	0.26	0.62	-0.18 (-0.52,0.16)	-0.27(-0.81, 0.28)	-0.08 (-0.52,0.36)
2	0.33	1.58	0.05	0.57	0.21	0.82	-0.06 (-0.26,0.14)	-0.11(-0.36, 0.14)	0.02 (-0.21,0.25)
3	0.14	0.05	0.18	0.71	0.82	0.67	-0.11 (-0.50,0.28)	-0.01(-0.60, 0.58)	0.14 (-0.25,0.53)
4	0.36	0.77	0.12	0.55	0.38	0.73	-0.07 (-0.31,0.16)	0.14(-0.16,0.44)	0.03 (-0.19,0.26)
5	0.02	0.33	0.01	0.90	0.57	0.93	-0.01 (-0.40,0.38)	-0.06(-0.59, 0.47)	0.02 (-0.27,0.31)
6	0.12	2.27	0.01	0.73	0.13	0.91	0.04 (-0.18,0.26)	0.12(-0.17, 0.41)	-0.01 (-0.29,0.26)
7	0.08	0.57	4.55	0.78	0.45	0.03	0.04 (-0.28,0.35)	-0.12(-0.58,0.35)	-0.22 (-0.46,0.01)
8	0.26	0.50	0.37	0.61	0.48	0.54	0.04 (-0.17,0.25)	-0.18(-0.46, 0.09)	0.03 (-0.24,0.30)
9	2.49	0.30	0.74	0.11	0.58	0.39	-0.32 (-0.69,0.04)	0.08 (-0.37, 0.54)	-0.17 (-0.51,0.16)
10	0.73	0.02	2.42	0.39	0.89	0.12	0.12 (-0.16,0.39)	0.09 (-0.26,0.43)	0.28 (-0.02,0.57)
11	0.05	0.00	0.06	0.82	0.96	0.80	-0.07 (-0.47,0.33)	0.03 (-0.43, 0.49)	-0.07 (-0.42,0.28)
12	0.29	2.86	1.06	0.59	0.09	0.30	-0.08 (-0.34,0.18)	-0.20(-0.49, 0.09)	0.15 (-0.14,0.44)
13	0.03	0.14	1.39	0.87	0.70	0.24	0.04 (-0.38,0.45)	0.13 (-0.35, 0.60)	0.18 (-0.25,0.62)
14	1.41	0.43	4.74	0.23	0.51	0.03	0.11 (-0.11,0.32)	-0.07(-0.34,0.19)	0.32 (0.06,0.57)
15	0.01	0.16	0.59	0.92	0.69	0.44	0.04 (-0.27,0.34)	0.09 (-0.28, 0.46)	0.20 (-0.14,0.54)
16	3.96	4.29	5.79	0.05	0.04	0.02	0.16 (-0.04,0.35)	0.12 (-0.04,0.28)	0.21 (0.00,0.43)
17	0.01	0.83	-	0.94	0.36	-	0.02 (-0.36,0.39)	0.14 (-0.23,0.51)	-
18	1.32	0.16	-	0.25	0.69	-	0.12 (-0.08,0.33)	0.11 (-0.18,0.39)	-
19	0.07	-	0.18	0.79	-	0.67	0.09 (-0.25,0.43)	-	0.05 (-0.24,0.34)
20	0.00	-	1.92	0.99	-	0.17	0.03 (-0.23,0.29)	-	0.22 (-0.04,0.48)

Figure 4.10, Figure 4.11, and Figure 4.12 shows test characteristics curves for DIF analysis based on gender for sign A, B, and C respectively with the female group as the focal group and male as the reference group. Meanwhile, Figure 4.13, Figure 4.14, and Figure 4.15 shows DIF analysis based on license status for sign A, B, and C respectively with non-licensed as the focal group and licensed as the reference group. Full graphs of DIF analysis can be found in Appendix M, N, O, P, Q, and R.

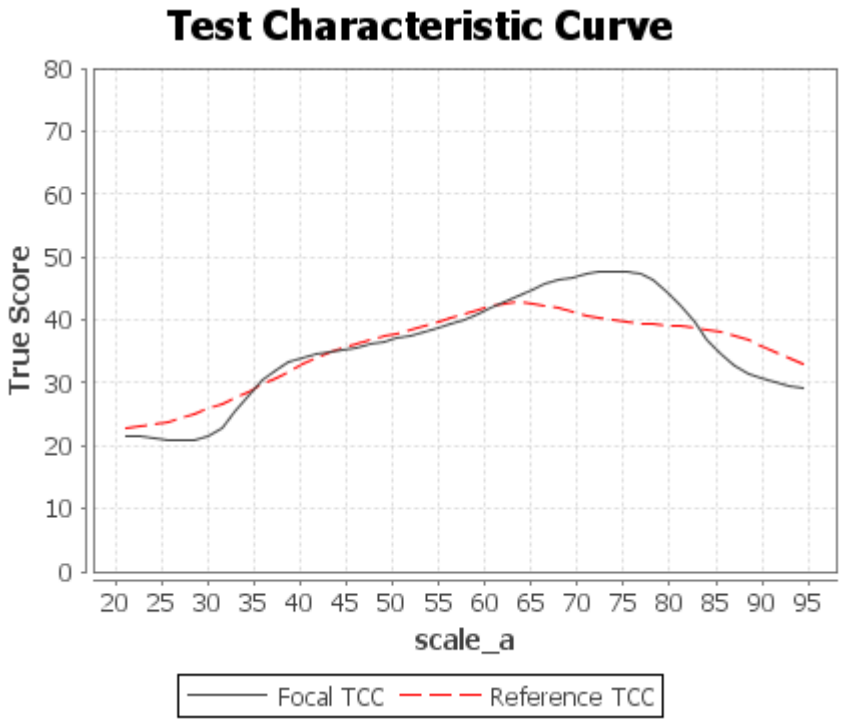


Figure 4.10: TCC of DIF for RP.06 based on gender

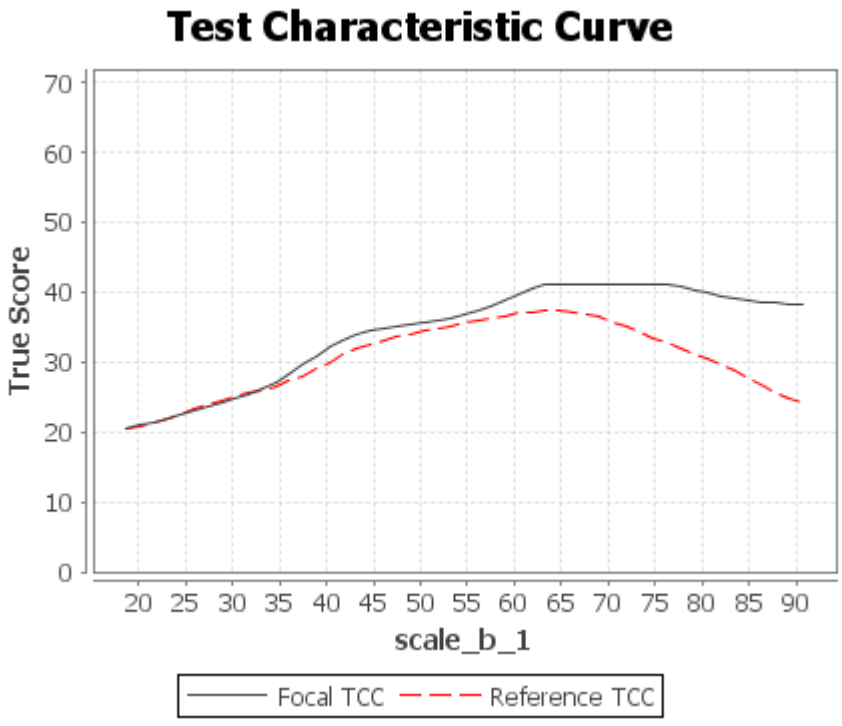


Figure 4.11: TCC of DIF for RP.10 based on gender

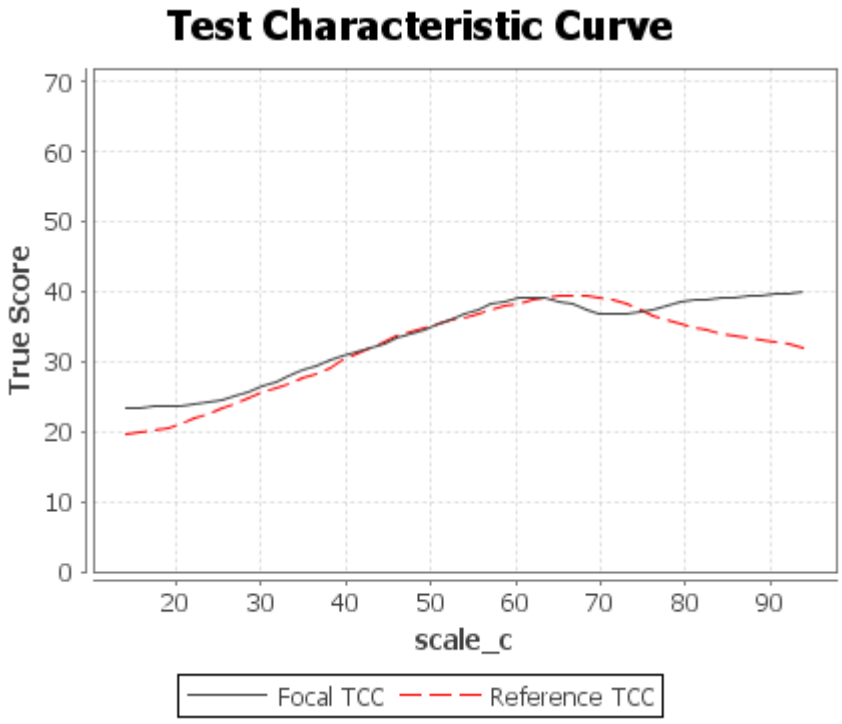


Figure 4.12: TCC of DIF for RP.05 based on gender

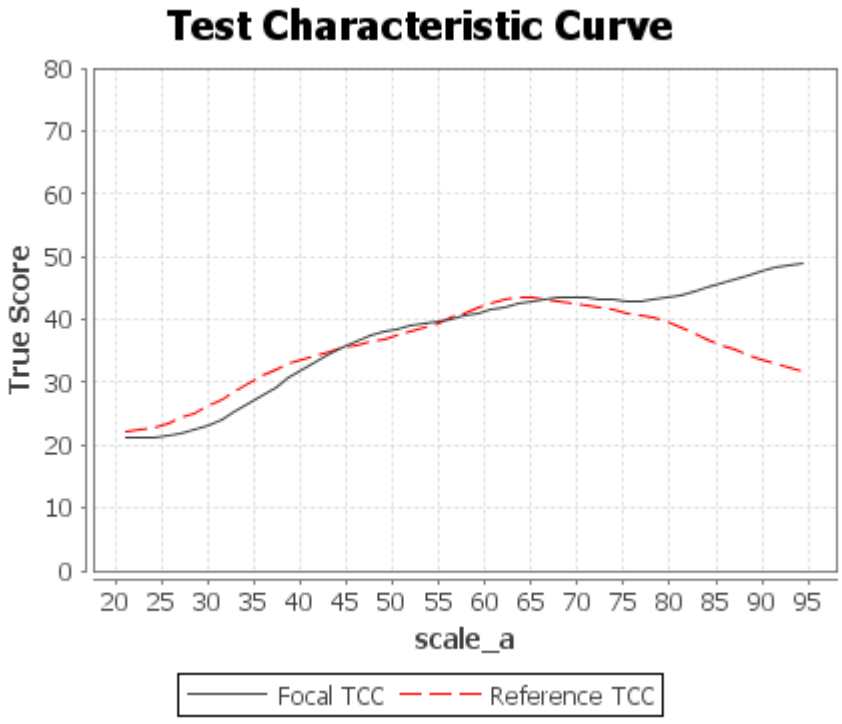


Figure 4.13: TCC of DIF for RP.06 based on license status

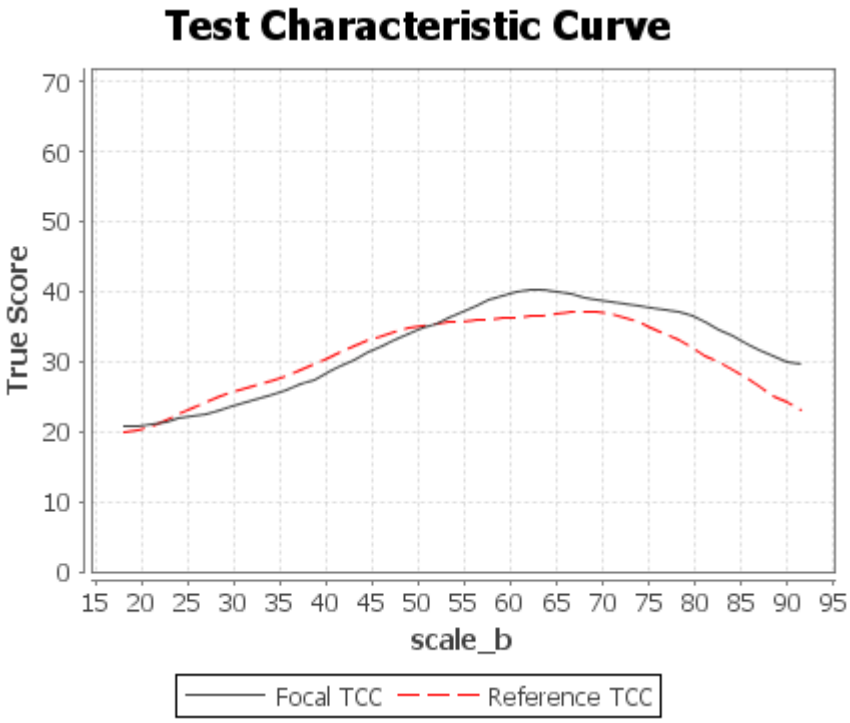


Figure 4.14: TCC of DIF for RP.10 based on license status

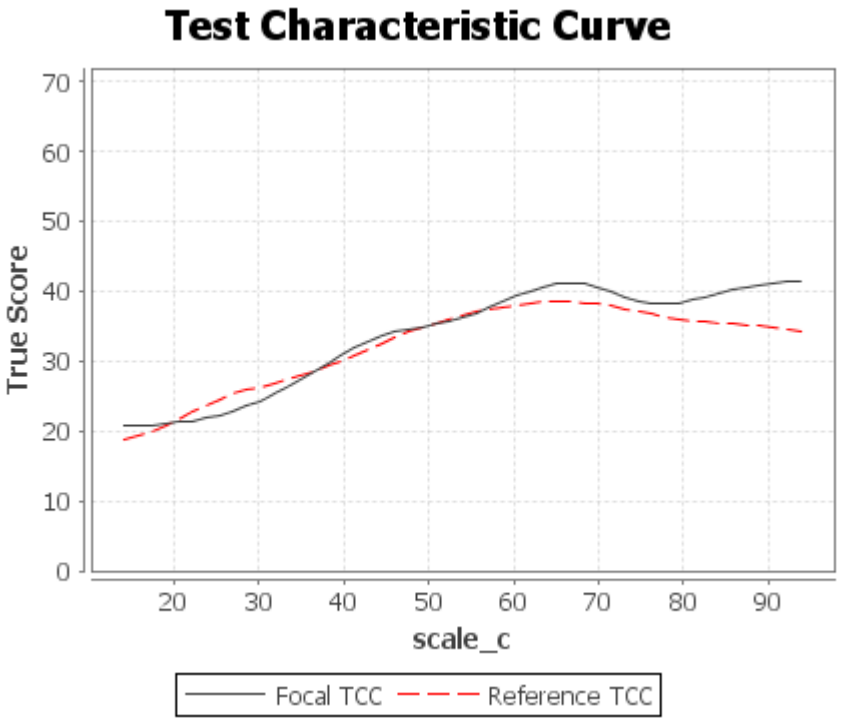


Figure 4.15: TCC of DIF for RP.05 based on license status

Table 4.51: DIF Classification Based on Gender and License Status

Item No.	DIF Classification Level (gender)			DIF Classification Level (license)		
	Text + Symbol	Symbol	Text	Text + Symbol	Symbol	Text
1	AA	CC+	AA	AA	BB-	AA
2	AA	AA	AA	AA	AA	AA
3	BB-	AA	CC+	AA	AA	AA
4	AA	BB+	AA	AA	AA	AA
5	BB-	AA	AA	AA	AA	AA
6	AA	AA	AA	AA	AA	AA
7	AA	AA	AA	AA	AA	BB-
8	AA	AA	AA	AA	AA	AA
9	AA	AA	AA	BB-	AA	AA
10	AA	AA	AA	AA	AA	BB+
11	BB-	AA	AA	AA	AA	AA
12	AA	AA	AA	AA	AA	AA
13	AA	AA	BB-	AA	AA	AA
14	BB-	AA	AA	AA	AA	BB+
15	AA	BB+	BB+	AA	AA	BB+
16	AA	AA	AA	AA	AA	BB+
17	AA	AA	-	AA	AA	-
18	AA	AA	-	AA	AA	-
19	AA	-	AA	AA	-	AA
20	AA	-	AA	AA	-	BB+

To gain a better understanding of DIF for all items, items are classified as shown in Table 4.51. Differential item functioning (DIF) analysis using Mantel-Haenszel Statistics is implemented to examine if different gender responds to items differently. DIF is classified into one of 3 different categories of observed DIF: AA, BB, or CC (negligible, medium, or large). In Table 4.51 DIF based on gender (focal = female, reference = male) and license status (focal = non-licensed, reference = licensed) are investigated. DIF for each sign is compared to achieve stability and consistency before a decision on item removal is made. Based on the results, items were not removed because of item relevance and the extent of DIF shown is not serious as none of the items show consistent existence of DIF. This indicates there is good item functioning and the measure is not discriminative.

DIF explains to what extent the item affects the item endorsement of participants sub-groups (gender, license status). To illustrate how subgroups differ in PSUS score for all test signs, independent t-tests were conducted by using the same groups as in DIF analysis (female=1, male=2; non-licensed=1, licensed=2). Figure 4.16 illustrate descriptive plots for PSUS score based on gender and license status. From the graphs, the graph for license status for RP06 and RP05 show early sign of significant difference.

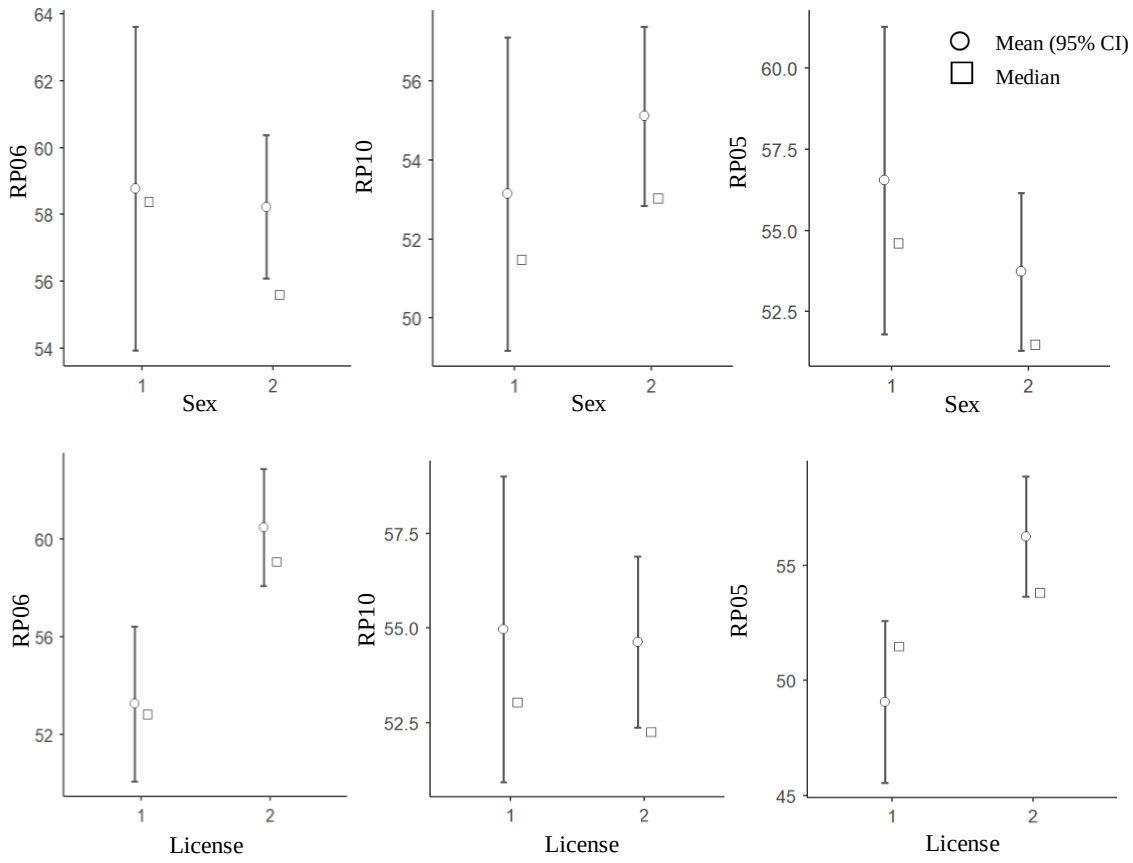


Figure 4.16: Descriptive plots for PSUS score based on gender and license status

This is further probed in Table 4.52 self-reference., independent t-test shows there is a significant difference in PSUS score between non-licensed and licensed road users for two signs: sign RP.06 t(202)= -3.293, p=0.001, Cohen's d = -0.5111 as well as sign RP.05 t(202)= -2.999, p=0.003, Cohen's d = -0.4654. Both signs show moderate effect of

difference, with non-licensed road users tend to score lower on PSUS compared to licensed road users.

Table 4.52: Independent T-Test for PSUS Score (Gender and License Status)

	Sign	Statistics	df	p	Mean difference	SE difference	Cohen's d
Sex	RP.06	0.224	203	0.823	0.549	2.45	0.0379
	RP.10	-0.813	203	0.417	-1.973	2.43	-0.1372
	RP.05	1.055	203	0.293	2.809	2.66	0.1780
License	RP.06	-3.293	202	0.001	-7.223	2.19	-0.5111
	RP.10	0.154	202	0.878	0.344	2.24	0.0239
	RP.05	-2.999	202	0.003	-7.181	2.39	-0.4654

As in Table 4.52, independent t-test shows there is a significant difference in PSUS score between non-licensed and licensed road users for two signs: sign RP.06 t(202)= -3.293, p=0.001, Cohen's d = -0.5111 as well as sign RP.05 t(202)= -2.999, p=0.003, Cohen's d = -0.4654. Both signs show moderate effect of difference, with non-licensed road users tend to score lower on PSUS compared to licensed road users.

4.5.7 Summary and decision on item removal

4.5.7.1 Summary on reliability

Rasch reliability indicators were examined for all test signs. It can be concluded that PSUS has an unalienable relative reproducibility, indicating that the instrument can be repeated with dependable and consistent results. Item and person reliability for PSUS are good: Sign A, RP.06 (item reliability = 0.9602, person reliability = 0.8732, Cronbach's alpha = 0.8439), Sign B RP.10 (item reliability = 0.9412, person reliability = 0.8694, Cronbach's alpha = 0.8281), and Sign C RP.05 (item reliability = 0.9299, person reliability = 0.8617, Cronbach's alpha = 0.8228). High item reliability indicates that the sample size is big enough

to locate items along the variable line. Meanwhile, high person reliability indexes indicate the ability to distinguish sample into enough levels for effective measurement.

Separation indexes of PSUS conform to the acceptable cut-off point. Person separation index should be more than 2 and item separation should be more than 3 (Linacre, 2012). Separation indexes of all tests signs also recorded good value: Sign A, RP.06 (item = 4.9157, person = 2.6240), Sign B RP.10 (item = 4.0001, person = 2.5796), Sign C RP.05 (item = 3.6425, person = 2.4959); indicating persons (or items) who are estimated to have high score indeed have higher probability to get higher score as compared to persons (or items) who were estimated to have low score.

In conclusion, in addition to good person separation index, it can be concluded that the scale could distinguish at least two distinct levels of usability in the sample. Person separation index of at least 1.5 was required to ensure that the scale could differentiate people with at least two different levels (Lerdal et al., 2016; Lerdal, Fagermoen, Bonsaksen, Gay, & Kottorp, 2014).

4.5.7.2 Characteristics curves, item map, and kernel density

Figure 4.17 illustrates Option Characteristic Curves for an item based on response anchors (0-4). The position of curves will vary, depending on item difficulty. This is the Option Characteristic Curves for item 1 for the first test sign. The complete graph can be found in Appendix I, J, and K. Ill functioning item may have a more unpredictable pattern. Note that the response curves are located relative to each other as in response anchor of Likert-types item. The curve for Response "strongly disagree" is on the left (black), followed by "disagree" (red), "neutral" (blue), "agree" (green) and "strongly agree" (purple).

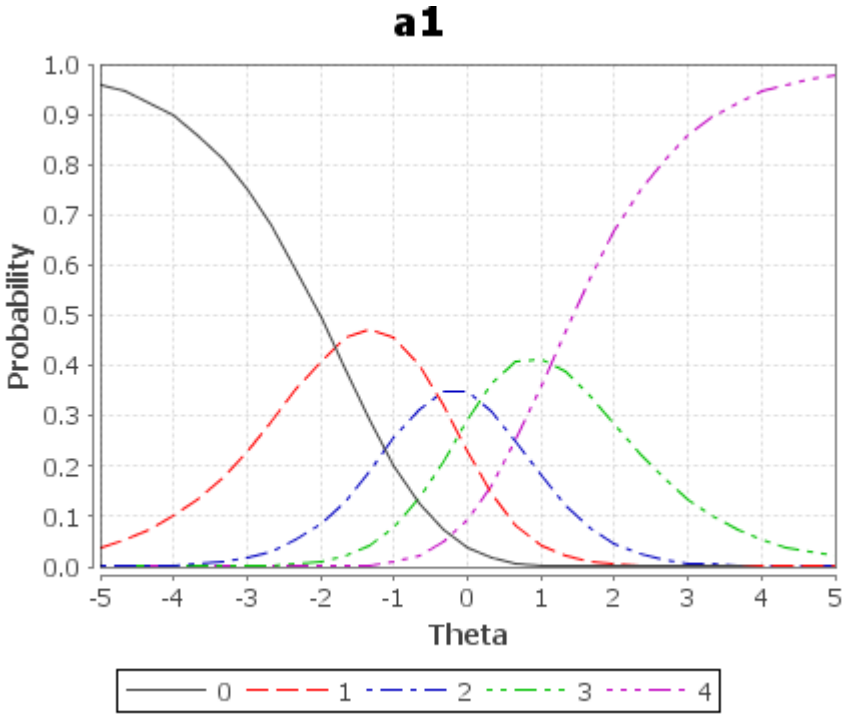


Figure 4.17: Option Characteristic Curves for All Options

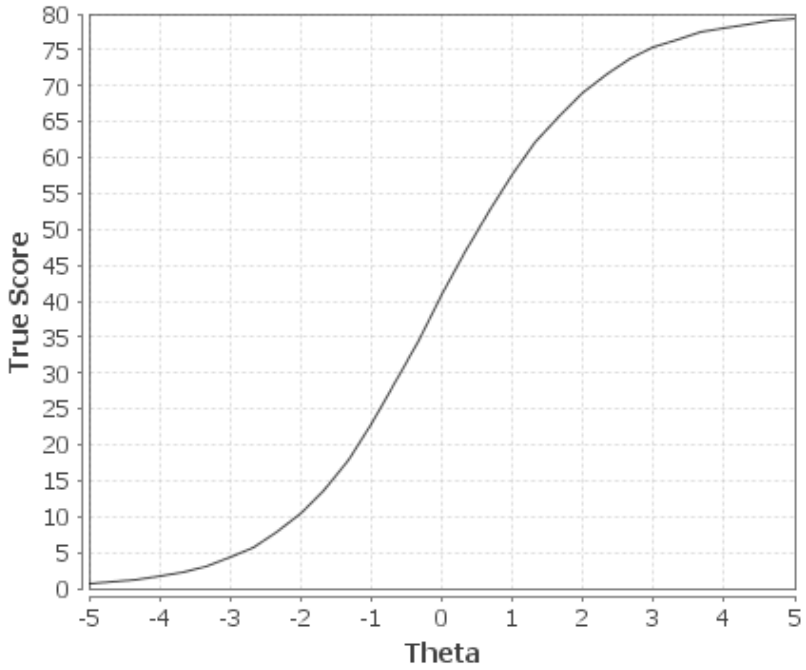


Figure 4.18: Item Characteristics Curve

Item characteristics curve as in Figure 4.18 illustrates item difficulty as well as the association between the latent trait and probability of agreeing to an item (example for sign A). The curves on all signs illustrate the same monotonic form as the probability to agree with an item increases as the item increases. Person plot for all test signs can be found in Appendix L.

Figure 4.19, Figure 4.20, and Figure 4.21 illustrate the range of item difficulty in relation to the people density. Laying out each item based on its difficulty aids interpretation of item and provides an insight about person ability for the measure. Item maps summarize the distribution item difficulty as well as person ability for all items in the instrument. It is insightful to be used to determine whether items are in line with the population. Based on all item maps, it can be concluded that the instrument has good range and is well centred in relation to the distribution of the person measure.

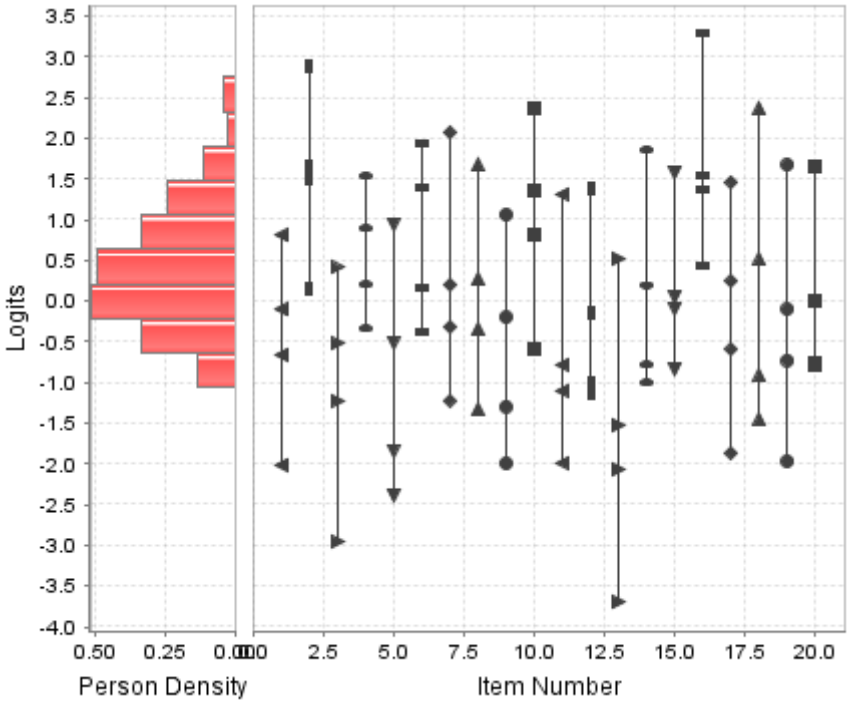


Figure 4.19: Item map for text + symbol sign (RP.06)

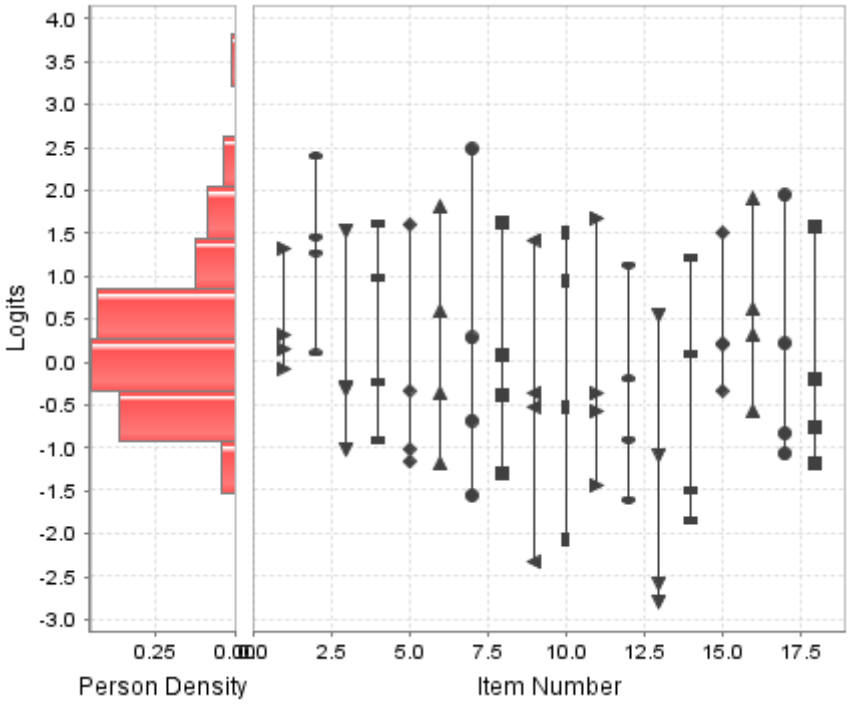


Figure 4.20: Item map for symbol-only sign (RP.06)

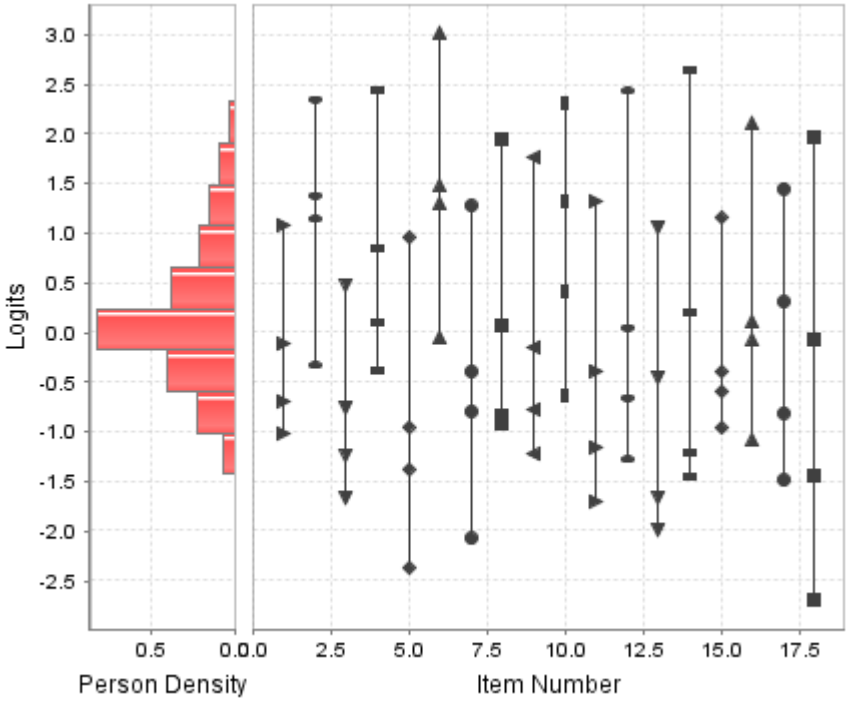


Figure 4.21: Item map for text-only sign (RP.06)

By examining the relationship between item difficulty and ability estimates, it is illustrated that most of the items covered most participants, with some items covering participants whose ability deviates from the norm. This situation is ideal in a measure because this shows that the instrument gives an adequate amount of precision for participants in general.

The functions in Figure 4.19, and Figure 4.20 are jagged and not appear as smooth as the Rasch characteristic curves. The implementing of characteristics curves based on nonparametric regression estimates can solve this limitation. To achieve this, Gaussian Kernel Density Estimate is adapted to illustrate the probability density function for person estimates (theta), as shown in Figure 4.22, Figure 4.23, and Figure 4.24 These non-parametric graphs provide a better visual representation of density function.

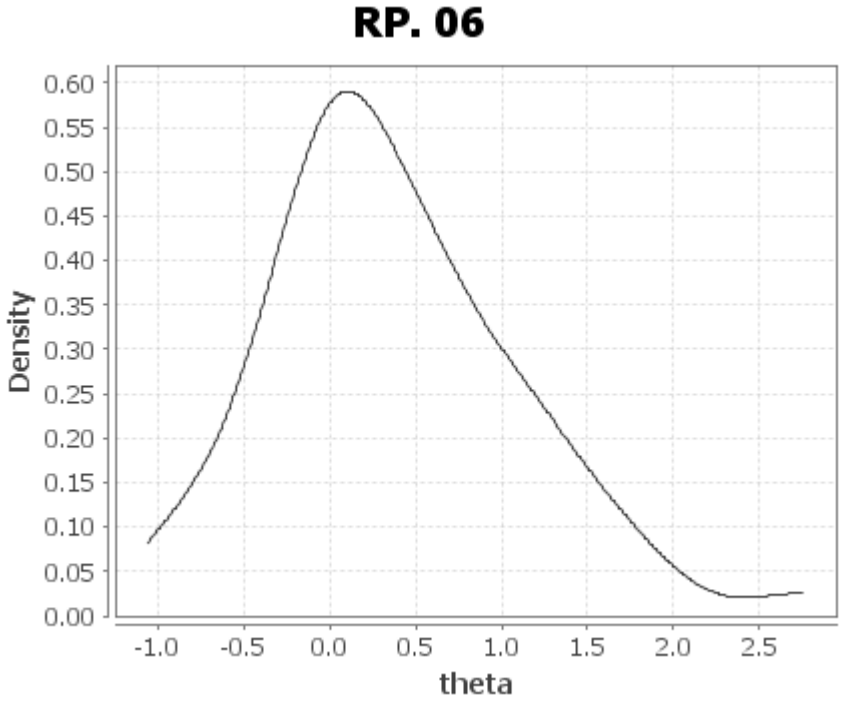


Figure 4.22: Kernel Density Function of person estimates for RP.06

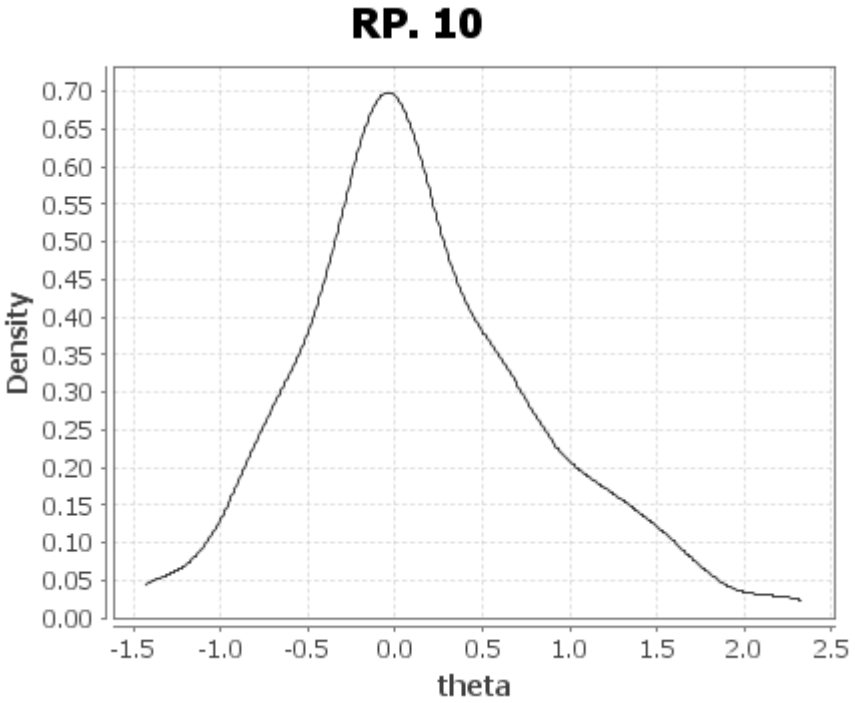


Figure 4.23: Kernel Density Function of person estimates for RP.10

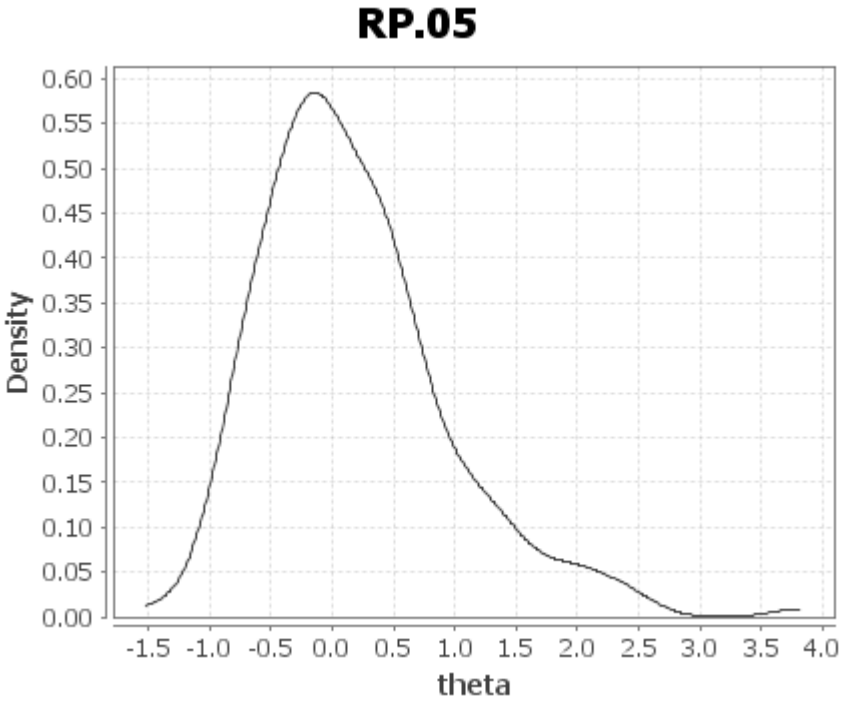


Figure 4.24: Kernel Density Function of person estimates for RP.05

Figure 4.22, Figure 4.23, and Figure 4.24 shows the same information as item map, but provide more details and precision than bar graphs. Based on these parametric graphs,

the density is the highest at theta value of 0 (average). Adequate dispersions of theta are observed.

4.5.7.3 Decision on item removal

Based on observations in the Rasch Analysis, problematic items are removed from the final set of the instrument. Problematic items are detected through series of quality measures inclusive of reliability analysis, item polarity assessment, fit statistics, and DIF analysis. Table 4.53 summarizes problematic items detected in various analyses conducted in this study.

Table 4.53: Screening of Problematic Items

Sign	Quality measures					
	Reliability Analysis	Item polarity	WMS and UMS	Std. WMS & Std. UMS	DIF	Deletion
A	Not detected	10	10	7, 9, 10, 12, 13, 15, 16	Not detected	10, 15, 16
B	Not detected	10	10	1, 9, 10, 11, 13, 15	Not detected	
C	Not detected	10	10, 16	10, 11, 12, 15, 16	Not detected	

As in Table 4.53, problematic items are items that have detrimental effects on the quality of the instrument. The item could be not contributing to the measure, possibility to represent another latent trait, and biased towards a group. Item 10 “I have seen many variations of this sign” is consistently problematic across different quality measures and is therefore removed from the final version of the instrument. Item 16 “It is likely that I will ignore this sign” was deleted because it shows unfit WMS and UMS values as well as unfit Std. WMS and Std. UMS values on two signs. Item 15 “I know what I should do when I see this sign” was also removed because it displays unfit values for Std. WMS and Std. UMS consistently across all types of sign.

4.6 Assessment of Test Signs

Unlike other psychometric assessments, sign usability needs to be evaluated in a different way to utilize the result from open-ended comprehension test and scale transformation of the composite score. This section attempts to assess the usability of three test signs using proposed method in Chapter 3. Cognitive mapping is illustrated to gain insight on participants' response pattern on the comprehension test.

4.6.1 Cognitive mapping

Participant's responses on open-ended comprehension test and its relative responses are recorded in tables. Table 4.54 depicts comprehension for sign A: RP.06 p.78: No Entry for Vehicles Having Overall Height Exceeding the Specific Height.

Table 4.54: Comprehension Test Result for Sign A RP.06

Meaning	Number of Responses	Relative Responses Frequency	Relative Person Frequency
Maximum height limit	115	0.777	0.561
General height indicator	8	0.054	0.039
Height of tunnel	5	0.034	0.024
Distant left	5	0.034	0.024
Vehicle height	2	0.014	0.01
Lorry lane	2	0.014	0.01
Minimum height	1	0.007	0.005
Vehicle length limit	1	0.007	0.005
Slow down	1	0.007	0.005
Bumper	1	0.007	0.005
Vehicle width	1	0.007	0.005
General prohibition	1	0.007	0.005
Maximum height limit (ft)	1	0.007	0.005
Compulsory height	1	0.007	0.005
Road suitability	1	0.007	0.005
Road level	1	0.007	0.005
Road width	1	0.007	0.005
TOTAL	148	1.004	0.723

Table 4.55 shows the result for second test sign, RP.10 p.79 No Parking. Lastly, result for the third test sign, RP.05 p.78 No Entry for Vehicles Having Total Weight Exceeding the Specified Weight is shown in Table 4.56. Based on the tables, cognitive mappings of perceived sign meaning of all three signs can be developed by considering the frequency of the relative response. The largest value of relative responses frequency is 1.004, therefore, the closer the value to one, the more frequent a meaning is assigned to a sign.

Table 4.55: Comprehension Test Result for Sign B RP.10

Meaning	Number of Responses	Relative Responses Frequency	Relative Person Frequency
No entry	63	0.432	0.307
No parking	35	0.240	0.171
No stopping	29	0.199	0.141
No waiting	5	0.034	0.024
Beware of danger	4	0.027	0.02
Stop	2	0.014	0.01
One way	2	0.014	0.01
General prohibition	2	0.014	0.01
Road section closed	2	0.014	0.01
Dead end	1	0.007	0.005
No turning	1	0.007	0.005
TOTAL	146	1.002	0.713

Table 4.56: Comprehension Test Result for Sign C RP.05

Meaning	Number of Responses	Relative Responses Frequency	Relative Person Frequency
Maximum weight limit	95	0.766	0.463
General weight indicator	9	0.073	0.044
Maximum height (feet)	6	0.048	0.029
Minimum weight limit	3	0.024	0.015
Compulsory vehicle load weight	3	0.024	0.015
Motorcycle lane	2	0.016	0.01
No. of construction	2	0.016	0.01
Length of vehicle	2	0.016	0.01
No. of holes on the road	1	0.008	0.005
Bridge height	1	0.008	0.005
TOTAL	124	0.999	0.606

As shown in Table 4.54 and Figure 4.25, RP.06 p.78: No Entry for Vehicles Having Overall Height Exceeding the Specific Height is generally well understood by participants with 0.777 relative responses frequency. This is an indicator that this sign tends to be misjudged.

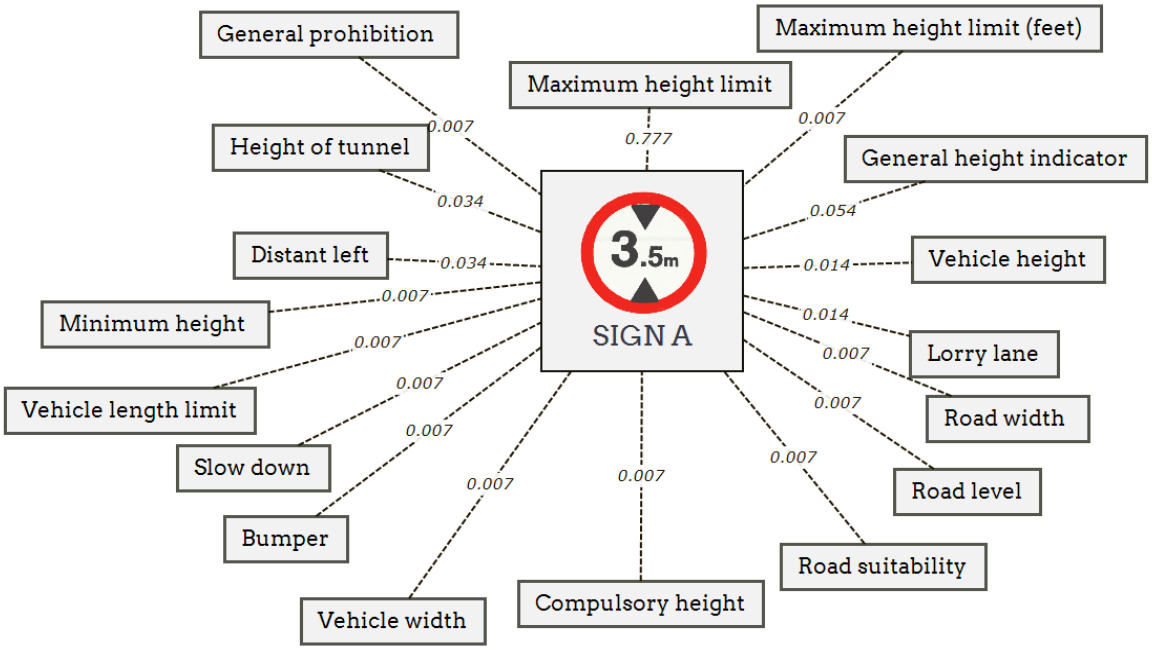


Figure 4.25: Cognitive mapping for sign A RP.06

Responses for sign A, RP.06 p.78: No Entry for Vehicles Having Overall Height Exceeding the Specific Height (Figure 4.25) are: Maximum height limit (0.777), General height indicator (0.054), Height of tunnel (0.034), Distant left (0.034), Vehicle height (0.014), Lorry lane (0.014), Minimum height (0.007), Vehicle width (0.007), Vehicle length limit (0.007), Slow down (0.007), Bumper (0.007), Compulsory height (0.007), Road suitability (0.007), Road level (0.007), Road width (0.007), General prohibition (0.007), Maximum height limit (feet) (0.007).

Responses for sign B, RP.10 p.79 No Parking (as in Figure 4.26) are: No entry (0.432), No parking (0.24), No stopping (0.199), Beware of danger (0.027), No waiting (0.034), Stop (0.014), One way (0.014), General prohibition (0.014), Road section closed (0.014), Dead end (0.007), No turning (0.007).

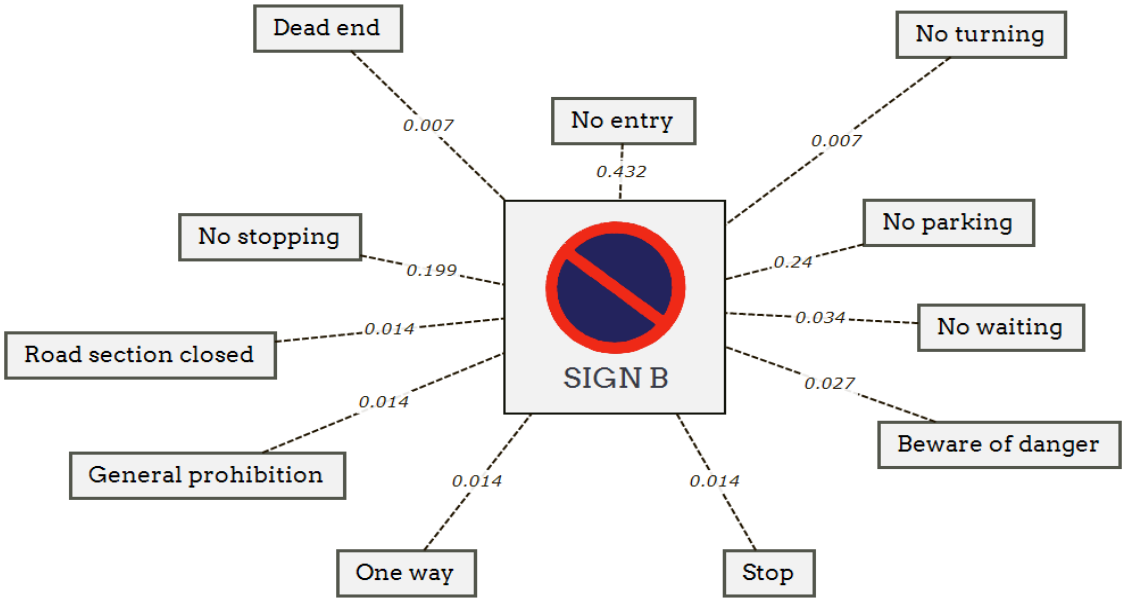


Figure 4.26: Cognitive mapping for sign B RP.10

RP.10 p.79 No Parking (Table 4.55 and Figure 4.26) has a noticeable sign of incorrect comprehension. Top response recorded was “No Entry”, which is an incorrect answer. Ministry of Works has defined this sign as No Parking sign in the manual of traffic sign (Public Works Department Malaysia, 1985b, p. 79), of which only recorded 0.24 of relative responses frequency.

Responses for sign C, RP.05 p.78 NO Entry for Vehicles Having Total Weight Exceeding the Specified Weight (Figure 4.27): Maximum weight limit (0.766), General weight indicator (0.073), Maximum height (feet) (0.048), Compulsory vehicle load weight (0.024), Motorcycle lane (0.016), No. of construction (0.016), Length of vehicle (0.016), Minimum weight limit (0.024), No. of holes on the road (0.008), Bridge height (0.008).

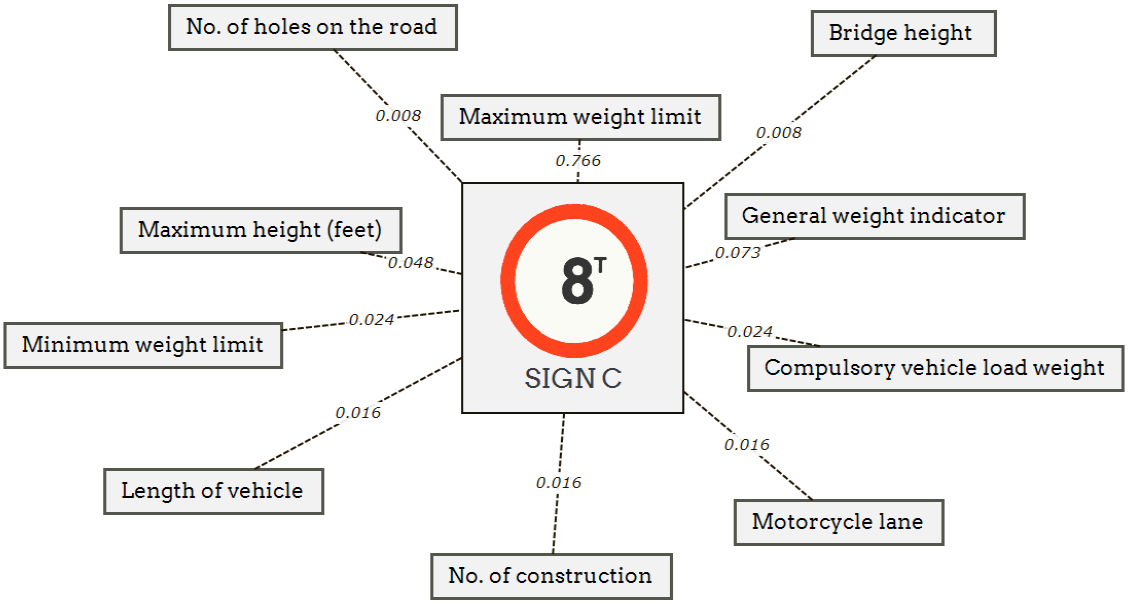


Figure 4.27: Cognitive mapping for sign C RP.05

Text-only sign, RP.05 p.78 NO Entry for Vehicles Having Total Weight Exceeding the Specified Weight (as shown in Table 4.56 and in Figure 4.27) has a similar response pattern with sign A, RP.06 p.78. The correct response recorded 0.766 of relative responses frequency. This result indicates there is no comprehension issue with this sign.

4.6.2 Classification of sign

The results from open-ended comprehension test and PSUS acceptability range are combined and the score from each participant is mapped in a cross-tabulation table and figures. Figure 4.30 shows PSUS of sign A RP.06, Figure 4.29 shows PSUS of sign B RP.10. Lastly, Figure 4.28 shows PSUS of sign C RP.05. These graphs illustrate whether a sign has the tendency to be misinterpreted. These graphs are useful when showing results to non-practitioners.

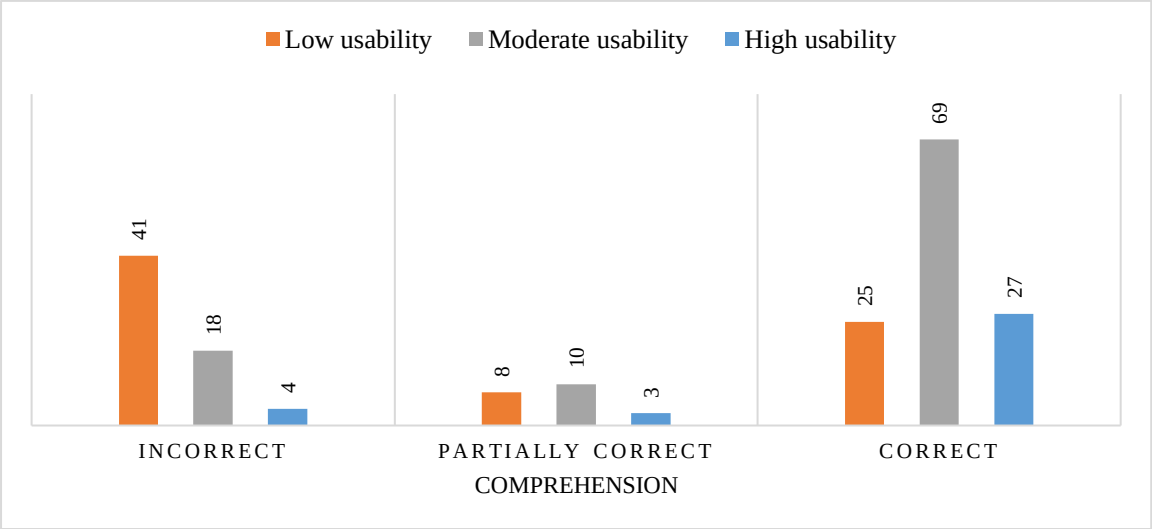


Figure 4.30: PSUS of sign A RP.06

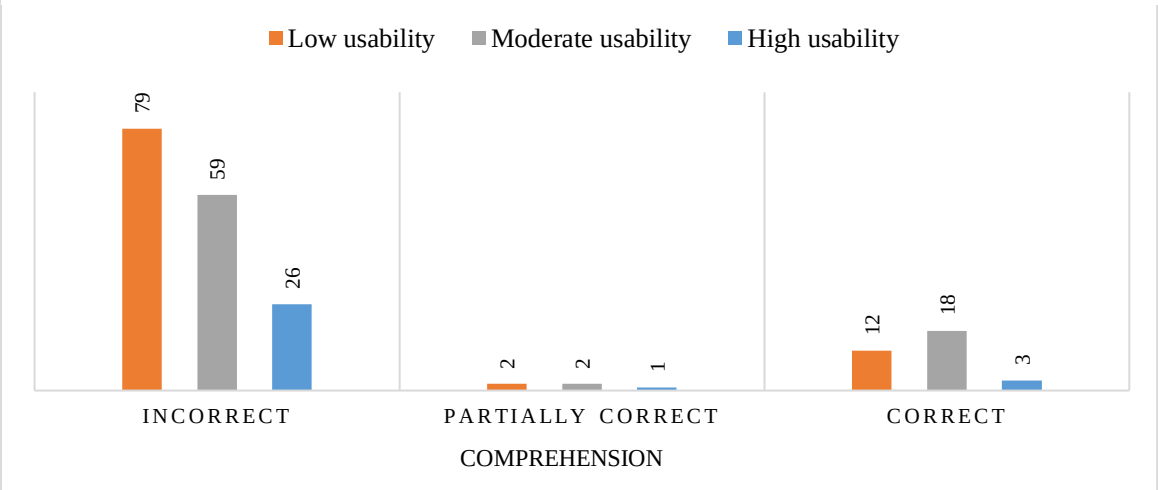


Figure 4.29: PSUS of sign B RP.10

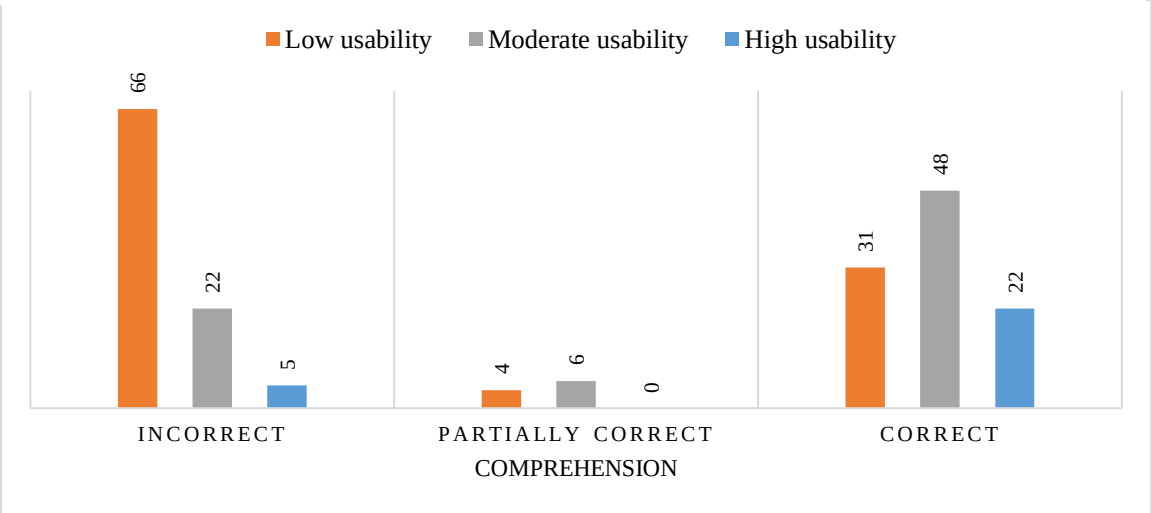


Figure 4.28: PSUS of sign C RP.05

Based on these results, all test signs were classified using sign classification as shown in Table 4.57. From the table, sign A (RP.06) recorded 69 counts of correct comprehension and moderate PSUS range, therefore it was found to have acceptable usability. Meanwhile, sign B (RP.10) and sign C (RP.05) both were classified as extremely low usability.

Table 4.57: Cross-tabulation and Classification of All Signs

Sign	Comprehension	PSUS Acceptability Range			Total	Classification
		Low	Moderate	High		
A (RP.06 p.78)	Correct	25	69	27	121	Acceptable usability
	Partially correct	8	10	3	21	
	Incorrect	41	18	4	63	
	Total	74	97	34	205	
B (RP.10 p.79)	Correct	12	18	3	33	Extremely Low Usability
	Partially correct	2	2	1	5	
	Incorrect	79	59	26	164	
	Total	93	79	30	202	
C (RP.05 p.78)	Correct	31	48	22	101	Extremely Low Usability
	Partially correct	4	6	0	10	
	Incorrect	66	22	5	93	
	Total	101	76	27	204	

Table 4.58: PSUS Classification of Test Sign

		PSUS Acceptability range		
		Low	**Moderate**	**High**
Comprehension	**Correct**	Effective but low satisfaction (3)	Acceptable Usability (2)	Very high usability (1)
	Partial	Somewhat low usability (6)	**Sign A (RP.06)** Below average usability (5)	Intended meaning partially conveyed (4)
	Incorrect	**Sign B (RP.10)** **Sign C (RP.05)** Extremely Low Usability (9)	Low usability (8)	Misinterpreted (7)

Table 4.58 shows the categorization of three test signs in the PSUS classification table. From the table, Sign A is categorized as class 5. Meanwhile, Sign B and C are categorized as class 9, which is the lowest sign usability class.

4.7 Discussion

This section discusses emerging issues related to the findings in this research. This includes the discussion on the difference of sign usability and the usability of other types of interface or system. Issues pertaining to CSC and interpretation of scale are also discussed. Finally, this section addresses the prowess of the PSUS formula derived to be adapted as an alternative for the calculation of composite score of Likert-type items.

4.7.1 Major findings

Table 4.59 shows the final version of PSUS.

Table 4.59: Final version of PSUS

No.	Item for text + symbol signs
1	I am confident that I understand the meaning of this sign.
2	I need help in understanding this sign for the first time.
3	I am familiar with this sign.
4	I easily forget the meaning of this sign.
5	I would learn this sign quickly.
6	This sign can be easily confused with another sign.
7	This sign can easily capture my attention.
8	This sign is complicated.
9	This sign is easy to understand.
10	This sign is easy to read.
11	The colours used in this sign are not suitable.
12	The colour contrast of this sign is good.
13	The shape of this sign is not suitable.
14	*The icon in this sign conveys a clear meaning.
15	*The icon design of this sign is not suitable.
16	**The text in this sign conveys a clear meaning.
17	**The text in this sign is difficult to read.

* These items are removed for text-only signs
** These items are removed for symbol-only signs

From Table 4.59, there are three versions of PSUS to suit three types of signs: Text and symbol sign, text-only sign and symbol-only sign. Text and symbol sign uses all 17 items in PSUS. Text-only and symbol-only signs each has 15 items. The omitted items for text-only are "The icon in this sign conveys a clear meaning" and "The icon design of this sign is not suitable". Meanwhile, the omitted items for symbol-only signs are "The text in this sign conveys a clear meaning" and "The text design of this sign is not suitable". Usability researchers should categorize signs prior to usability testing.

4.7.2 Sign usability vs. usability of other systems

Traditionally, usability has been used as a concept to measure the quality of user-system interaction (Ketola & Röykkee, 2001). Ketola & Röykkee (2001) discussed how the concept of usability is different for other systems by demonstrating different facets of usability in mobile handsets. It was demonstrated that the usability of a product can be very complex and products should be evaluated based on its unique characteristics. Ketola & Röykkee (2001) have also identified three conceptual interfaces that constitute mobile handset usability, which are the user interface, service interface, and external interface. It was suggested that future researchers on mobile handset usability should incorporate all facets of usability.

The analysis of several popular usability questionnaires provides knowledge of certain aspects of a system or product that constitutes to the overall usability. However, it was emphasized in the findings of this study that sign usability should be viewed and interpreted differently as a different type of product has different elements that constitute to its usability. Based on the results of the documentary analysis, 12 CSC were identified. From this, it is proven that sign usability aspects are different than generic system or product.

Therefore, sign usability should only be assessed with instrument specifically designed to assess sign usability. The use of a modified generic instrument is not a correct representation of sign usability and its use should be discouraged.

One of the apparent differences between sign usability and other products is the lack of user interaction. This difference is noted by a researcher who noted that the use of sign requires no user-sign interaction (Lay, 2009). Based on analysis of various usability instruments, usability testing is heavily dependent on the user experience of using the system or product to accomplish intended tasks. This is contradictory to the use of sign whereby a user does not use sign to perform a task because the function of a sign is to provide information to the user. Aside from the lack of user interaction, the use of signs is different from systems because signs do not have many functions, capabilities as well as the ability to recover from user's error. There are no step-by-step operations that need to be carried out to perform a certain task.

In past researches, proof for the use of traffic sign is by detecting eye fixation using computerized eye tracking methods (Inman, 2012). Every sign gives a certain meaning, either to give information or warn drivers about potential danger and regulation. Generic usability questionnaires contain many system-specific items such as error recovery and prevention, on-screen help and user manual, use of commands and operators, aesthetical values, system error as well as system responsiveness. While it was noted that usability of system and sign is different, the two still share some common concepts, namely simplicity, effectiveness, learnability, and memorability.

4.7.3 Cognitive sign characteristics (CSC)

McDougall et al. (1999) identified sign characteristics that constitute sign usability by conducting an intensive literature review. From the study, five characteristics were identified: concreteness, complexity, meaningfulness, familiarity, and semantic distance (McDougall et al., 1999). This is a more prominent attempt to identify sign characteristics and it has been adapted by other sign usability studies (Ng & Chan, 2007a; Ou & Liu, 2012; Yuan et al., 2014). However, it is apparent that after many years another documentary analysis should be conducted to update this finding. A more recent attempt in identifying characteristics includes characteristics of warning labels as part of safety sign. This study identified four effectiveness criteria: judgments of noticeability (saliency), likelihood to read, understandability (comprehension) and the likelihood of complying (authoritativeness) (Laughery et al., 2002).

This study extends the findings of past researchers as 12 CSC have been identified: Authoritativeness, Complexity, Comprehension, Familiarity, Learnability, Legibility, Meaningfulness, Memorability, Priming, Saliency, Standardization, as well as Surface Properties. This study noted the close association between aforesaid characteristics. The first indication is the c-coefficient of CSC. High co-occurrence was detected for following pairs: Complexity-Authoritativeness, Meaningfulness-Comprehension, Complexity-Meaningfulness, and Familiarity-Meaningfulness. This result shows that in the literature, Meaningfulness frequently co-occurs with Comprehension, Complexity, and Familiarity. This observation supports the findings from past researches. For example, attributes pertaining to the meaningfulness of sign such as semantic distance (relatedness of the sign to what it represents) and concreteness (extent to which an icon is abstract or arbitrary) were investigated (McDougall et al., 1999; Mcdougall & Curry, 2001; Ou & Liu, 2012; Zhang &

Chan, 2013). These attributes determine the extent to which a sign can be easily comprehended (Mcdougall & Curry, 2001). This shows strong association between Meaningfulness and Comprehension. In terms of the relationship between Meaningfulness and Complexity, it was noted that a sign must have a appropriate level of complexity in order to be meaningful the user (Charlton & Baas, 2006).

Results from MRA illustrate eight co-occurrences in relation to CSC assigned to an item: Complexity-Priming, Learnability-Priming, Complexity- Learnability, Complexity-Memorability, Comprehension-Meaningfulness, Legibility-Surface Properties, Legibility-Complexity, and Surface Properties-Complexity. These co-occurrences of characteristics are not unexpected, as past researches have noted the relationship between characteristics.

CSC assigned to deleted items in this study are Standardization (item 10), Priming (item 15) and Authoritativeness (item 16). There are other items assigned with Standardization and Priming but Authoritativeness was removed. The deletion of authoritativeness based on Rasch analysis implicates less need for authoritativeness in the measure. This is supported by unstable item discriminant statistics for item 16, ranges from 0.25 to 0.43. Authoritativeness records the lowest responses on MRA but recorded an average number of codes in the coding of literature with 241 codes and relative code occurrence frequency of 0.0489. More evidence needed to disregard the importance of Authoritativeness in the measure. This study noted the need to reassess the significance of Authoritativeness. If Authoritativeness is proved to an indispensable aspect of sign usability, there is a need to develop another item to represent the characteristics.

4.7.4 Interpretation of scale

The instrument was developed with the idea that all characteristics can be combined to represent sign usability. This also acts as a basis for choosing Rasch measurement to develop the instrument instead of factor analysis. It can be tempting to use the single adjective rating of each question alone in place of the composite score to evaluate a specific characteristic of sign usability. However, researchers and practitioners should refrain from doing so because of a few reasons. For example, a study based on analysis of SUS stated that researchers should avoid the temptation to analyse the scores of individual questions and report only the overall usability score instead because of the lack of task success rates or time-on-task measures in the SUS. It was also noted that multiple items are generally superior to a single item (Bangor et al., 2009).

The interpretation of SUS scale is not lacking in controversies. Despite the suggestion by Brooke (1996) and Bangor et al. (2009, 2008) that the structure of SUS is unidimensional, Lewis & Sauro (2009) attempted to categorize the items in the SUS according to attributes of usability by assuming Learnability and Usability as independent factors. A different study tested both models (Unidimensional and two-factor) and noted that unidimensional model is being too narrow to represent the contents of all SUS items. On the other hand, the two-factor model with uncorrelated factors is viewed as being too restrictive for its psychometric assumptions (Borsci, Federici, & Lauriola, 2009). Therefore, it was suggested that is a correlation between usability and learnability in the bidimensional model, which is less restrictive. However, based on a documentary analysis, eight studies listed learnability as one of the attributes of usability and should not be measured separately (Jeng, 2005). This notion is supported by results in this study. Based on Rasch analysis, it was concluded that the instrument is unidimensional.

The juxtaposition of factor structure is illustrated in Figure 4.31. The far left model depicts 2-factor model of SUS (Lewis & Sauro, 2009), followed by the unidimensional model of SUS (Bangor et al., 2008) and lastly the factor structure of PSUS.

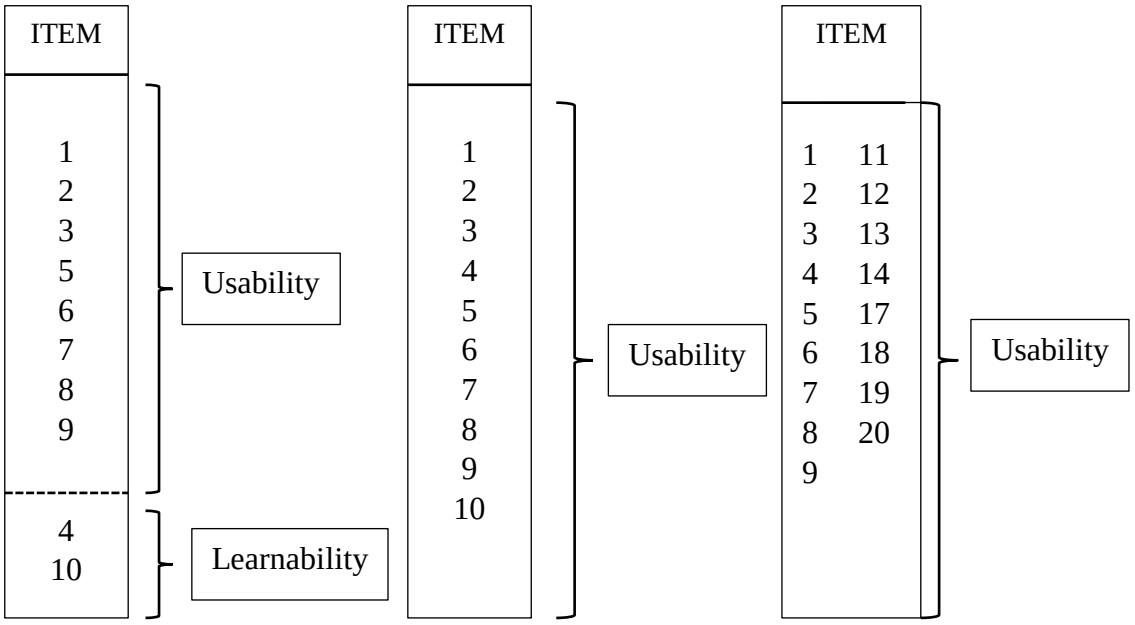


Figure 4.31: Structure of SUS and PSUS

4.7.5 Individual characteristics in sign comprehension

There is an uncertainty as to what extent gender can affect sign comprehension/usability. Several studies indicated that gender can affect sign usability (Al-Madani & Al-Janahi, 2002; Borowsky et al., 2008; Ismail, 2012; Shinar et al., 2003). Even among these studies, the results were not anonymous. Al-Madani & Al-Janahi (2002) and Shinar et al. (2003) found that sign comprehension is significantly higher among male drivers than female drivers. These studies contradict Borowsky et al. (2008) and Ismail (2012) who concluded that female drivers scored better in sign identification and comprehension test. Ismail (2012) attributed the difference to higher level of education received by female drivers.

In this study where all participants received the same level of education, it was found that there is no difference in the perception of sign usability between male and female road users. This result provides supporting evident for Ismail (2012)'s claim. A significant amount of researches stated that male and female drivers did not perform differently in sign comprehension test as well as their perception of sign usability (Arphorn et al., 2003; Brémond & Deugnier, 2006; Costa et al., 2014; Duarte et al., 2014; Makinde & Oluwasegunfunmi, 2014; Makinde & Opeyemi, 2012; Ng et al., 2011a, 2011b; Talab et al., 2013; Yuan et al., 2014). This result reiterates these findings, providing another evident to conclude that gender plays an insignificant role in sign usability.

While gender is the one of the most investigated individual characteristics, the difference in sign comprehension between licensed and non-licensed road users has always been overlooked. Driving experience was often found to influence sign comprehension (Brémond & Deugnier, 2006; Costa et al., 2014; Crundall & Underwood, 2001; Mitchell, 2010; Shaver et al., 2006), but researches often disregard non-licensed road users, such as pedestrian and cyclist who never received their driving education who are not exempted from fatal road accidents. This study detected significant difference in the perception of sign usability between licensed and non-licensed road users in two of the three test signs. PSUS rating is significantly higher among licensed road users for RP.06 (text & symbol) and RP.05 (text-only) sign, but no difference was detected for RP.10 sign (symbol) sign.

As the difference in PSUS rating was only observed in two test signs, this discrepancy in the result reveals a serious error in previous researches on how the effect of individual characteristics was investigated. It is a common practice to use the average rating score or the average comprehension score for all signs as the dependant variable when testing for the

effect of individual characteristics. This practice could be a mistake because differences in sign usability caused by individual characteristics could be unique to the tested signs. This could explain why the findings in previous researches on individual characteristics were greatly diverse and often contradict.

4.8 Chapter Summary

Development of PSUS comprises of several stages. This includes establishing need and objective, conceptualization of concept, instrument design and establishing reliability and quality. To develop the instrument, a total of 284 participants were involved in the study. Establishing the need for a new measure is an important part of instrument development. First item review examines the use of modified SUS.

Second item review was conducted after the addition of new items germane to CSC identified. To facilitate adaptation of the instrument in English-speaking countries, readability tests were also conducted. Results proved that the instrument is easy to read and comprehensible with the approximate grade level of 5 in the US. To ensure the reliability of the readability tests scores from different readability coefficients (correlation of FKGL, GFS, CLI, SMOG, and ARI) were correlated. The scores were positively correlated (r=0.79), indicating consistency across tests. Subsequently, Multiple Responses Analysis (MRA), Cochran's Q, and McNemar tests were conducted to assess if the items relate to at least one of the CSC identified. It was proven that all items are relevant based on CSC, which supports the validity to the measure.

To further establish the quality of the instrument, Rasch analysis was performed. PSUS shows good reliability and separation indexes. Overall of three items were removed. Item 10

“I have seen many variations of this sign” does not contribute to the measure with negative item discriminant as it shows opposite polarity on all test signs. Item 10 also consistently records unfit WMS, UMS, Std. WMS and Std. UMS. Item 15 “I know what I should do when I see this sign” and item 16 “It is likely that I will ignore this sign” were removed because of problematic fit and standardized fit statistics. No problem of differential item functioning was detected for all test signs.

This study introduced several PSUS interpretation methods, including cognitive maps of users’ response on open-ended sign comprehension test and classification of sign based on scale and comprehension test. With this classification, signs can be categorized into nine usability classes. From the use of PSUS, sign A shows acceptable usability, sign B, and C records extremely low usability.

This study acknowledges that there are many aspects of sign usability. Although physical/environmental factors, as well as individual differences also contribute to sign usability, this instrument provides useful and crucial insight of cognitive aspect of sign usability.

CHAPTER 5

SUMMARY AND CONTRIBUTION

5.1 Overview

This chapter summarizes research procedures and findings. The implications of the study are discussed by highlighting limitations of past researches. This section explains how limitations serve as a basis to delimit some aspects of the study. Furthermore, suggestions for future research are also included in this chapter to spearhead the emergence of future studies.

5.2 Summary of the Findings

The main purpose of this study is to develop an instrument to assess road users' perception on the usability of traffic sign based on cognitive sign characteristics. For this purpose, firstly cognitive sign characteristics are identified by analysing researches on sign usability. Test items are generated based on initial item pool consisted of 119 items taken from nine generic usability instruments. The initial item pool is then filtered based on redundancy and relevancy by incorporating cognitive sign characteristics that have been identified.

Item reviews were conducted for several purposes: to assess suitability and limitation, increase comprehensibility as well as to assess the readability of the instrument. Series of item review involved 13 participants consisted of lecturers in the field of ergonomics and graduate students who have experiences in the field of ergonomics/human factors.

To confirm that the items refer to a certain sign characteristic, MRA was conducted. From the results, all items were deemed related to cognitive sign characteristics. The items

were then further analysed with Rasch analysis to assess its reliability and validity. Aspects being investigated include item polarity, item/person separation, Rasch fit statistics, unidimensionality, and differential item functioning. Based on the findings, item 10, 15 and 16 were removed as they deteriorate the quality of the measure. The measure was found to be unidimensional and not biased towards gender and license status.

The instrument consists of two parts: open-ended sign comprehension test and Likert-type items on users' perception on sign usability based on its cognitive sign characteristics. Method to calculate the composite score of the Likert-type items is developed, along with interpretation method of Likert-type items and open-ended comprehension test. Suggestions for interpretation of open-ended comprehension test are mapping sign meaning using relative frequency. The meaning can also be coded as "correct", "partially correct" and "incorrect" and subsequently used in sign categorization. The suggestions to interpret Likert-type items are grade scale and categorization based on score, dichotomous responses method and sign categorization by combining results from open ended comprehension test.

Based on specific objective 1: to develop an instrument to assess sign usability based on its cognitive sign characteristics, summaries according to various processes of instrument development are as follow:

I. Design and development framework for Perceived Sign Usability Scale (PSUS).

   To determine the most suitable design and development framework for an instrument, gaps in literature need to be considered. The design framework for PSUS as in Figure 1.3 is the result of rigorous literature review and the identification of emerging problems. This design framework includes several processes, which are

the review of existing methods, identification of concepts, construction of items and interpretation methods, as well as establishing the quality of the instrument.

II. Review of existing methods to assess sign usability

Sign (traffic, and safety sign) usability researches were mostly conducted in the United States and China, involving non-probabilistic samples of mostly student samples. There were many approaches utilized by researchers to assess the usability of sign, but most methods are computer-aided approaches such as automated road signs detection, visual tracking/eye movement, recall, verbal protocol, naming, driver responses, tachistoscopic presentation, technical analysis of sign luminance and conspicuity, situational awareness, and analysis of accident rates. For cognitive approaches of sign usability, self-administrated rating scale and open-ended comprehension test are widely accepted as two of the most appropriate and effective methods. Other methods include driving simulation and lab experimentation, multiple-choice questions, interview, and documentary analysis. There were efforts to modify System Usability Scale to assess sign usability, but this attempt does not capture the full aspect of cognitive sign characteristics.

III. Identification of cognitive sign's characteristics that constitute road signs usability

This study identified 12 cognitive sign characteristics: Authoritativeness, Complexity, Comprehension, Familiarity, Learnability, Legibility, Meaningfulness, Memorability, Priming, Saliency, Standardization, and Surface Properties.

IV. Development of items for a new measure to assess road users' perception on traffic sign usability.

The instrument developed in this study is named Perceived Sign Usability Scale (PSUS). After establishing reliability and validity, the final version of PSUS consists of 17 items for text + symbol sign, 15 items for symbol only and text only sign. In addition to the Likert-type item, the instrument also incorporates open-ended sign comprehension test to classify signs into nine sign usability categories.

V. Establishing the quality of the instrument

Analysis based on Rasch reliability and Cronbach's alpha shows excellent reliability. Good separation index proves that persons (or items) who are estimated with high scoring indeed have a higher probability to have a higher score. After items removal, the measure is proved to be unidimensional and productive for measurement. Moreover, the instrument does not show differential item functioning based on gender and license status. Item map and item characteristics plot were used to illustrate option curves and item difficulty in relation to person ability.

For the second research objective on the development of a formula to calculate composite score of the new instrument, PSUS was developed to combine all responses on Likert-type items in the instrument into a scale of 0-100. This composite score calculation can be used to substitute the use of mean and sum of which were clouded with controversies for decades. The conventional composite score calculations were also demonstrated to be not intuitive in measuring usability. Using this formula, the recoding of negatively worded items is unnecessary. To ease implementation, constant tables were provided, and calculation method was demonstrated.

This study also suggests methods to interpret the data obtained with the instrument (Objective 3). The interpretation of this instrument is different than other usability instruments whereby the open-ended question and Likert-type items is combined to classify sign into one of nine sign usability categories: Extremely low usability (class 9), Low Usability (class 8), Misinterpreted sign (class 7), Somewhat low usability (class 6), Below average usability (class 5), Intended meaning partially conveyed (class 4), Effective but low satisfaction (class 3), Acceptable usability (class 2), Very high usability (class 1). The open-ended question and Likert-type items can also be individually interpreted. Cognitive mapping technique based on relative responses frequency was used to analyse open-ended responses. For the Likert-type item, the scale can be categorized into Low, Moderate, and High usability.

Lastly, the instrument is used to assess the usability of test signs (Objective 4). PSUS was used to assess the usability of three different sign type: symbol + text sign, symbol only sign, and text only sign. Based on three examples, sign A (RP.06 p.78) was found to have acceptable usability (class 2). Sign B (RP.10 p.79) and sign C (RP.05 p.78) both recorded extremely low usability (class 9).

5.3 Contribution and Implication of Study

5.3.1 A new dimension of sign usability research

The concept of sign usability was not a well-defined concept, as different researchers investigated different aspects of sign usability in various ways. This study noted the diverse direction of sign usability researches. This study also noted three aspects that contribute to sign usability as shown in Figure 5.1. This shift the way sign usability is viewed and how sign usability researches will be conducted in the future. The identification of cognitive sign

characteristics, individual characteristics, and physical/ environmental characteristic acts as a platform for the emergence of possibility for new researches on sign usability. Notable research at identifying sign characteristics was 20 years ago (in 1996) (McDougall et al., 1999). New aspects of sign characteristics have emerged and warrant for attention.

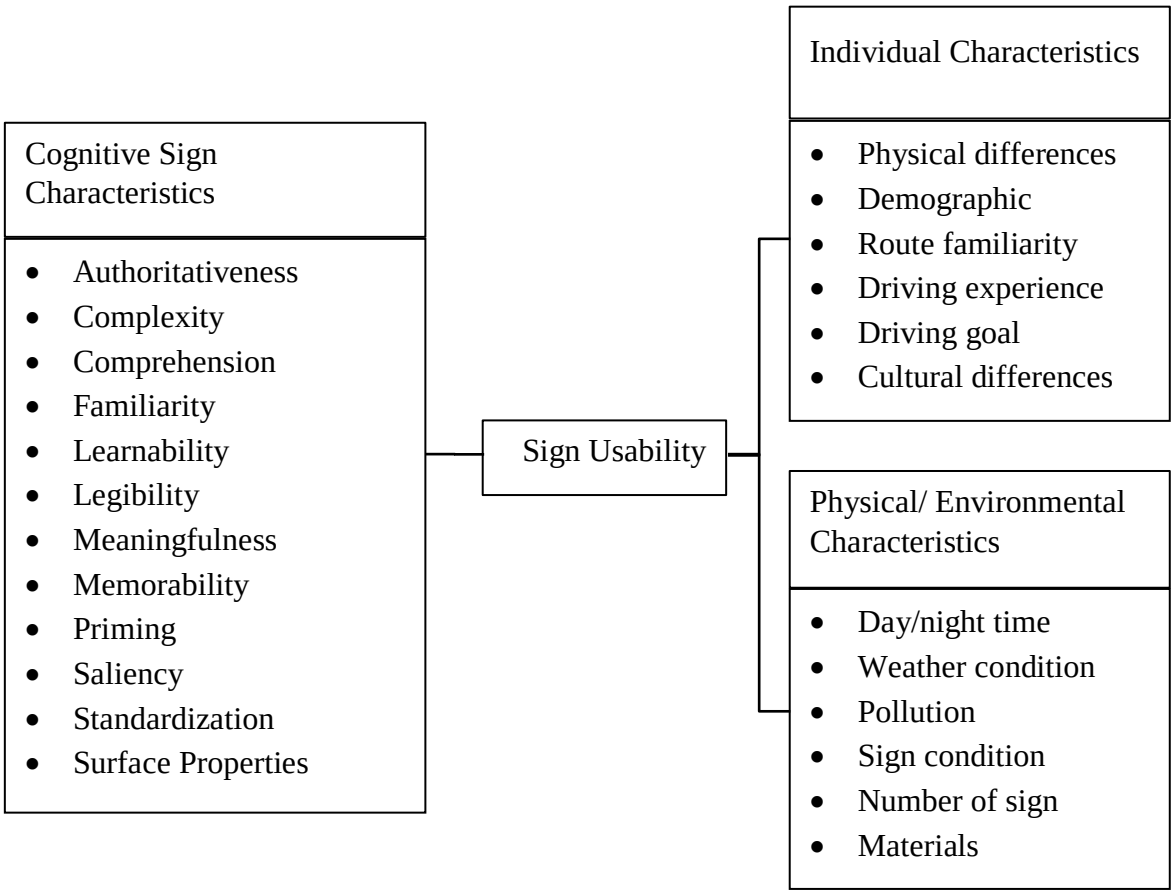


Figure 5.1: Aspects of sign usability

5.3.2 Methods/framework to develop instrument

Design and development framework for PSUS as illustrated in Figure 1.3 can be adapted by other researchers who wish to develop rating scale instrument. This framework serves as an extension of instrument development flow adapted from Brancato et al. (2006) and Radhakrishna (2007). In developing an instrument, some researchers tend to skip the substantive process of establishing need and rush to the construction of items. This

framework avoids this hasty and hazardous approach, as it is crucial to establish the need of the instrument prior to the item construction. This step is crucial because it can save time and resources, as there is no need to develop a new instrument if existing measure can be adapted. This framework is more conclusive as it includes establishing the need for a new measure through identifying gaps in existing measures, the identification of concepts, administration method, composite score calculation method as well as interpretation method.

Aside from design framework, this study explores the use of various statistical methods to develop instrument. For example, in the case whereby an expert on the subject area is not available, this study explores the use of Multiple Responses Analysis (MRA) to ensure the items represent a certain predefined concept. MRA allows participants to assign any number of concepts (0 to all) to an item. A manual of MRA analysis using dichotomy method as used in this study was presented and disseminated as part of workshop series presentation. This manual shows the steps on conducting MRA and description on how to deal with missing data in the analysis. This open-access manual has benefited not only workshop students but as well as other researchers. When analysing results from MRA, typically researchers only report on the relative percentage based on respondents and responses. Thus, the analysis is limited to report only percentages. There was no method found to demonstrate the technique to find differences in responses proportion. This study provides more in-depth analysis of MRA result by incorporating Cochran's Q test and McNemar test in addition to Bonferroni correction to detect these differences. This method is more inclusive and flexible.

5.3.3 PSUS formula to calculate composite score of Liker-type items

The popularity of Likert scales is increasing in usability studies. System Usability Scale (SUS) is one of the popular examples of the implementation of Likert scales in usability studies. SUS was often modified to assess the usability of various types of system and products. SUS converts the responses on Likert-type items into a scale ranging from 0 to 100. Since there was no attempt to develop a formula to calculate the scale of modified SUS, the calculation of modified SUS can be complicated thus may cause confusion and errors. It can also be difficult for researchers whose number of items and scale size is different from the original SUS. This study bridged this gap by introducing a formula to calculate different versions of SUS. Aside from calculating the scale for modified SUS, this formula is robust and can be adapted as a method to compute the composite score of any instruments adapting Likert Scales. Practitioners have several choices to compute the composite score of Likert Scales. Due to controversies and limitations in using mean and sum, the formula serves as a better choice for computing composite score. Scores calculated using the formula have a perfect relationship with mean and sum, indicating it could be used as an alternative. Although there is still much work to do in interpreting the individual score calculated using the formula, it is more intuitive to be comprehended.

This study will contribute as part of the initiative to enhance road safety and enhance the usability of traffic signs. As the number of traffic accidents is on the rise, efforts to enhance safety can benefit the society in general. The output of this study is an instrument to assess users' perception on sign based on cognitive sign characteristics, along with a detailed explanation on how to use the instrument and interpret the result. From the use of this instrument, sign usability can be intuitively investigated and explained. This is particularly useful for researchers or practitioners to explain the result to stakeholders who

have no background in usability testing. Various companies and institutions can adopt this instrument to assess the usability of safety signs in their own context of use. This study can aid the government to identify problematic signs as well as explaining the scenario for sign comprehension among road users. With this input, the government can decide whether to conduct retraining of sign, reinforcement, or redesign of signs.

Moreover, findings can serve as a guideline for researchers to conduct future researches on sign usability. The identification of cognitive sign characteristics serves as a platform for the emergence of possibility for new researches on various aspects of sign usability as well as sign usability.

5.3.4 Methods/technique to assess sign usability

Problems related to existing techniques to assess sign usability has been discussed in Sub-section 1.3.4. The issue of precision was the focal point of the discussion, as the adaptation of inappropriate methods may lead to imprecision of conclusion. This study attempts to provide a solution of these problems by offering a more precise measure. To achieve this, this study combined rating scale of sign usability with comprehension test to gain more understanding of sign comprehension. Doing this also allows sign classification by combining results from open-ended question and usability items.

This study draws cognitive mapping of sign meaning based on responses on sign comprehension test. This acts as an extension for open-opened comprehension tests by previous researches. Cognitive mapping allows the researcher to identify most common interpretation of a sign. In previous researches, the treatment of open-ended comprehension test is by using percentage of correct responses (Lesch, 2008; Razzak & Hasan, 2010; Yuan et al., 2014; Zamanian et al., 2013), multiple-choice question (Ng et al., 2011b), scored as

correct or incorrect answer (Ward et al., 2004), as well as scored as correct, partially correct, and incorrect (Chan & Ng, 2012; Hancock et al., 2004; Jones & Cottrell, 1999; Wogalter et al., 2002).

5.3.5 Practice: Instrument to assess sign usability

In Malaysia, the traffic sign standard was used since 1985. No notable attempt was found to assess the usability of the sign based on road users' perception. This instrument can be the catalyst to spearhead sign usability research and to evaluate existing signs. The research can be conducted in collaboration with local authorities. The outcome will serve as a guideline for policy making, decision to redesign, and retraining traffic and safety signs.

Table 5.1: Summary of Contributions

Aspects	Contribution
Relook the concept of sign usability	• This study identifies various gaps in past researches surrounding the assessment of sign usability. • Findings shift the way sign usability is viewed with the identification of CSC. • Widens the understanding on the matter by explaining different facets of sign usability, explaining how human perceive signs as well as how sign usability differs from other products.
Methods	Development of new: • Instrument design framework • Composite score formula • Sign usability classification methods Using unconventional methods/tools and demonstrate the use of: • C-coefficient, Multiple responses analysis, Cochran's Q, and Multiple McNemar in the instrument development process. • New software (for example: JMetrik & JASP/Jamovi)
Practical	• Provide awareness and a solution to the problem of sign usability of which was well documented but lacking in interventions. • Assess sign usability with the newly developed instrument to identify problematic signs for redesign or retraining. • Allow government agencies to update traffic sign standard of which was adopted since the 80s. • This thesis serves as a manual for assessing sign usability using PSUS.

5.4 Limitations and Delimitations

Among the notabilia of limitation is the identification of CSC. The researcher does not deny the possibility that some relevant documents were being excluded in the analysis. It was also not feasible to include all articles on the domain to be included for analysis. Thus, articles were selected based on several inclusion criteria as described in Sub-section 3.4.1. Several measures were taken to ensure that the researcher is au courant with the matter and has included pivotal documents and researches on sign usability based on CSC. Measures include the use of multiple search engines/databases, asterisk wildcards, and thesaurus. It is also acknowledged that the coding process of CSCS was not error-proof. To increase reliability, each code was verified twice.

Limitations of using self-reported data, such as data acquired using questionnaires were noted in this study. It is a fact that self-reported data are difficult to verify. Moreover, the data obtained with self-reported questionnaires can be affected by several factors. Data can be affected by physical and mental fatigue, boredom and time constraints (Brutus, Aguinis, & Wassmer, 2013). This is more apparent in lengthy questionnaires whereby time constraint also causes participants to answer questions quickly without having a full understanding of the questions. This limitation acts as a basis to delimiting this study to only several signs instead of all standard traffic signs listed by the ministry. Based on feasibility studies, the time needed for the participants to answer all items for a sign (open-ended comprehension test and 18-20 Likert-type items) is about four minutes, not inclusive of time taken for briefing, reading the informed consent, and responding to demographic questions. Based on this limitation, this study delimits the time to administer the instrument for about 20 minutes, which can only accommodate three test signs.

The signs tested in this study are standardized regulatory and warning road signs as listed in the standard road signs manual of Ministry of Works, Malaysia. This study focuses on regulatory and warning signs because they can reduce risk factors associated with potential hazards on the road. Other non-standard, informative, temporary, other non-traffic related signs not included in this research. However, the instrument developed in this study may be adapted for other signs that share a similar characteristic (printed signs with the use of text and icon).

Furthermore, it also was noted that using questionnaire somehow limits respondents into particular response anchors (Simon & Goes, 2013). In the case of Likert-type items, responses are limited based on scale size. To reduce the effect of this limitation, an open-ended comprehension test was implemented along with rating scale to assess sign usability.

The language barrier can be a dangerous limitation in a research as participants may have difficulty to understand and respond to the questions. Sarawak is a multilingual state with about 30-40 different languages spoken (Mis, 2010). To address this limitation, Malay translation was provided, and participants can answer open-ended comprehension test with Malay, Iban or English. Readability test was also conducted to ensure the instrument is generally easy to comprehend. In addition, the researcher is available for any clarification.

Researchers sometimes encounter problems in identifying positive and negative worded statements in the instrument. Although the formula suggested in this study does not require the researchers to recode the item, the formula requires them to identify positive and negative statements. The formula also needs to be tested in various fields other than usability researches and tested using different data set. This article demonstrates the usage of the formula using two statistical tools: IBM SPSS v22 and GNU PSPPIRE v0.8. Computation

using other statistical software, i.e. R, SAS, SOFA, and STATA etc. might need a different approach.

The formula is applicable if value label for the lowest value (0 or 1) indicates an unfavourable response or disagreement; whereas the highest value indicates a favourable response or agreement, corresponds to the nature of original Likert Scale (Likert, 1932) (i.e. 1-Strongly disagree, 2- Disagree, 3-Neutral, 4- Agree, 5- Strongly Agree). For questionnaire with reversed scale (5 for Strongly Disagree), the value calculated must be subtracted from 100.

A scale from 0 to 100 is more intuitive to comprehend. However, researchers should interpret the individual score with caution as it raises many questions about what a single SUS score means in an absolute sense (Bangor et al., 2009). SUS developer noted that one of the limitations in this method is people tend to interpret the scale as a percentage (Brooke, 2013). The value 84 does not mean a product/system is 84% usable. The adjective ratings and grade scale suggested by several studies (Bangor et al., 2009, 2008) may be exclusive to human factors researches. More evaluations needed to apply adjective ratings and grade scale in different fields.

5.5 Suggestions for Future Research

This study opens many possibilities to expand the knowledge on sign usability as delimitations of the studies serve as a guide for aspects that need to be investigated. Suggestion for future research is explained based on three key components: revisiting and expanding theories/concepts, adapting methods in this study as well as measures for intervention.

5.5.1 Revisiting and expanding theories/concept

The aim of developing the instrument is not to create a superlative sign usability instrument but rather attempts to ameliorate known issues of existing methods. In achieving this, research is aware of possibilities that the research can be expanded in several ways:

I. Cognitive sign characteristics: This study is based on cognitive sign characteristics that were identified using a fixed pool of articles. The researcher does not deny the possibility of other cognitive sign characteristics that might also constitute sign usability. Therefore, it is beneficial that future research should revisit the list of sign characteristics. The identification of other sign characteristics can complement the findings of this study.

II. In addition, future research can explore about the interconnection between each sign characteristics. This is particularly useful because c-coefficient result on co-occurrence between concepts indicates possible interconnection between concepts. This result is backed by past researches that also noted the association between various cognitive sign characteristics.

III. It is also critical to investigate the association between physical and cognitive sign characteristics. Past researches have placed considerable effort on investigating individual characteristics on sign usability in general. Future research should continue this effort and incorporate more individual characteristics to be studied. The identification of cognitive sign characteristics allows for more in-depth and specific investigation of the effect of individual characteristics.

5.5.2 Adaptation of instrument and interpretation methods

I. The new Perceived Sign Usability Scale instrument should be tested in new context. This study identified individual differences that were cited to affect sign usability. Findings from previous researches were often contradicting. This indicates the need to revisit individual differences and provide more verification to approve or disprove past findings.

II. Researchers can implement different test conditions while collecting data with the instrument. This includes using different data collection method, diverse and random test samples, as well as the inclusion of more signs to be tested. The test can also be extended to safety signs.

III. Future researches should report test quality indicators for comparison. Quality indicators may include Cronbach's alpha, PCA of Rasch residuals, Item polarity as well as investigation of differential item functioning in another context.

IV. The instrument was developed with Rasch Analysis, which is a modern test theory. Rasch analysis was chosen due to several advantages that have been discussed in Chapter 3. The implementation of classical test theory such as factor analysis, alternate form correlation, item-total correlation, Cronbach's Alpha per scale, or stepwise regression may provide different results. It would be insightful to note the difference.

V. Due to controversies and limitations in using mean and sum as the composite score, the formula was adopted as an alternative for various statistical analyses. Scores calculated using the formula have a perfect relationship with mean and sum, indicating it could be used as an alternative. This formula is robust as an alternative

to calculate the composite score of any instruments adapting Likert Scales. Instead of using sum and mean to calculate the composite score, researchers who adapt Likert-type instruments can use the formula to calculate composite score. This recommendation is not only applicable for human factors researches but also any Likert-type instrument.

VI. This study demonstrates the use of the formula using two statistical tools: IBM SPSS v22 and GNU PSPPIRE v0.8. Computation using other statistical software, i.e. R, SAS, SOFA, and STATA etc. might need a different approach. Researchers who use different tools should report on the procedures to aid other researchers.

VII. The adjective ratings and grade scale suggested by several studies (Bangor et al., 2009, 2008) may be exclusive to human factors researches. Should researcher wish to interpret the composite score calculated using the interpretation methods suggested in this study, they should first conduct a research on rating scale based on their context of use.

VIII. Implement a different version of the instrument in other community/location. Translation of the instrument into local language can be done for this purpose. This is particularly useful for communities that differ in terms of culture and language. It was documented that cultural issues can affect sign comprehension, especially for international drivers (Dissanayake & Lu, 2001; Ng & Chan, 2007b).

5.5.3 Intervention

I. Researchers who use the instrument to evaluate signs should note on signs that are tested to have low usability. Those signs should be improved through methods of design that are well documented in the human factors literature. Providing training

to drivers was cited to be an effective way to enhance sign comprehension (Lesch, 2008; Ng & Chan, 2007c; Ward et al., 2004). Method of training and its effectiveness should be documented to allow replication of training method.

II. Should researchers decide to redesign problematic signs, the instrument should be administered again to assess improvement in usability. This iterative design process and evaluation opens more opportunity for future studies.

5.6 Concluding Remarks

Summary of research contributions and recommendations for future studies are shown in Table 5.2,Table 5.3, and Table 5.4. These tables provide a brief description of contributions and recommendations in relation to statements of problem and objectives of the study.

Table 5.2: Summary- Differences in Usability Aspects

Problem Statement - Difference on usability of sign and other products			
Process	**Findings**	**Contribution**	**Recommendation**
Identifying cognitive sign's characteristics that constitute road signs usability.	12 cognitive sign characteristics were identified: Authoritativeness Complexity Comprehension Familiarity Learnability Legibility Meaningfulness Memorability Priming Saliency Standardization Surface properties	This finding shifts the way traffic sign usability research will be conducted in the future and galvanize future researches investigating CSC. The identification of CSC provides the understanding of how sign usability differs from other systems/products.	If a new cognitive sign characteristic is identified, it can be used to improve the inclusiveness of PSUS. Future studies should investigate the association between physical, cognitive, and environmental aspects of sign usability.

Table 5.3: Summary - Poor Sign Comprehension & Lack of Sign Usability Studies

Problem Statements - Poor traffic sign comprehension, lack of sign usability study and usability instrument, issues related to existing sign usability methods			
Objectives/ processes	**Findings**	**Contribution**	**Recommendation**
Developing a framework for the design and development of PSUS	PSUS development framework incorporates a review of existing methods to establish needs, identification of concepts, composite score calculation, and interpretation methods.	PSUS design and development framework serves as a guideline to develop instruments.	Future researches should report on the implementation of PSUS and its efficacy to improve the framework.
Reviewing existing methods to assess sign usability.	Self-administrated rating scale and open-ended comprehension test are widely accepted as an appropriate and effective format to assess sign usability. However, past researches only utilize several cognitive sign characteristics to assess sign usability.	This study identifies gaps and limitations in existing questionnaires.	Cognitive aspect of traffic sign can be combined with the physical and environmental characteristics to gain more understanding on sign usability in a different situation.
Developing items for a new measure to assess road users' perception on traffic sign usability.	Existing instruments on sign usability are deemed too shallow as they only incorporated several CSC. PSUS items were developed based on modified SUS and item pool from existing generic usability instruments.	PSUS incorporates more CSC. This study also utilizes statistical methods such as Cochran's Q, and Pairwise McNemar to assess item validity.	Translate the instrument into the local language and report the use of the instrument.
Examining psychometric properties of PSUS.	PSUS shows good reliability and validity. The final version of PSUS consists of 17 items for symbol + text signs and 15 items for symbol-only as well sd text-only signs.	Sign usability can be assessed with PSUS.	Incorporate more test signs involving different samples.
To assess the usability of test signs using PSUS	Three test signs were tested: sign A was found to have acceptable usability. Sign B and C both recorded extremely low usability.	Manual to assess sign usability	Adapt suggested methods and report on the use of PSUS with other types of signs

Table 5.4: Summary – Problems in Composite Score Calculations

Problem Statement - Controversial Composite Score Calculation Methods			
Objectives	**Findings**	**Contribution**	**Recommendation**
To develop a formula to calculate the composite score of the new instrument.	PSUS formula was developed. The formula yields a usability score that ranges from 0-100 and corresponds to visual analogue scale (VAS) and semantic differential scale. The formula is unaffected by scale size, the number of items, as well as the number of positive or negative items.	PSUS formula can be used to calculate the composite score of Likert scales. PSUS scale is more intuitive and devoid of the controversies surrounding the use of mean.	Report on the use of the formula with different data sets and tools.
To determine the best methods to interpret the data obtained with the instrument.	This study combines the result from open-ended comprehension test and Likert type items to classify sign into nine usability classes.	This classification method is a new method to assess sign usability that addresses the problem of misinterpretation of sign (high perceived usability but low comprehension)	Report the implementation of interpretation methods.

This thesis has largely focused on the development of PSUS instrument to assess sign usability, inclusive of its interpretation methods. This study took place because of the emerging need for an intervention to enhance road safety and the limitations of sign usability instruments. Traffic sign usability has always been overlooked although it is an indispensable aspect of traffic safety measures. Thus, this study was conducted with the hope that it will serve as the main guide and a catalyst for future sign usability researches. Further analyses utilizing new data and approaches help in fine-tuning and expanding this study. As an initiative effort in developing sign usability instrument that is relevant to cognitive sign characteristics, this is precisely what this study pursues, to constantly evolve itself along with the expansion of knowledge about cognitive sign characteristic

REFERENCES

Abran, A., Khelifi, A., Suryn, W., & Seffah, A. (2003). Consolidating the ISO usability models. *Proceedings of 11th International Software Quality Management Conference*, 23–25.

Agarwal, N., Chaudhari, A., Hansberry, D. R., Tomei, K. L., & Prestigiacomo, C. J. (2013). A comparative analysis of neurosurgical online education materials to assess patient comprehension. *Journal of Clinical Neuroscience*, *20*(10), 1357–1361. https://doi.org/10.1016/j.jocn.2012.10.047

Ahmad, N., Oranye, N. O., & Danilov, A. (2017). Rasch analysis of Stamps's Index of Work Satisfaction in nursing population. *Nursing Open*, *4*(1), 32–40. https://doi.org/10.1002/nop2.61

Akple, M. S., & Biscoff, R. (2012). Assessment of drivers understanding of road signs in Ghana : A case study along Ho – Afloa Togo route. *International Journal of Business, Management and Social Sciences*, *3*(1), 19–25.

Al-Gadhi, S. A., Naqvi, S. A., & Abdul-Jabbar, A. S. (1994). Driver factors affecting traffic sign detection and recall. *Transportation Research Record*, (1461), 36–41.

Al-Madani, H. (2004). Comprehension of Signs: Driver Demographic and Traffic Safety Characteristics. In C. Castro & T. Horberry (Eds.), *The human factors of transport signs* (pp. 169–183). Boca Raton: CRC Press. https://doi.org/doi:10.1201/9780203457412.ch11

Al-Madani, H., & Al-Janahi, A. R. (2002). Assessment of drivers' comprehension of traffic signs based on their traffic, personal and social characteristics. *Transportation Research Part F: Traffic Psychology and Behaviour*, *5*(1), 63–76. https://doi.org/10.1016/S1369-8478(02)00006-2

Allen, I. E., & Seaman, C. A. (2007). Likert scales and data analyses. *Quality Progress*, *40*(7), 64–65. Retrieved from http://mail.asq.org/quality-progress/2007/07/statistics/likert-scales-and-data-analyses.html

Allen, T. M., Lunenfeld, H., & Alexander, G. J. (1971). Driver information needs. *Highway Research Record*, (366), 102–115.

American National Standards Institute. (2007). *Criteria for safety symbols ANSI Z535.3-2007*. Rosslyn, VA: National Electrical Manufacturers Association.

Andrich, D. (1978). Rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561–573.

Anshel, M. H., Weatherby, N. L., Kang, M., & Watson, T. (2009). Rasch calibration of a unidimensional perfectionism inventory for sport. *Psychology of Sport and Exercise*, *10*(1), 210–216. https://doi.org/10.1016/j.psychsport.2008.07.006

Apple, M. T. (2013). Using rasch analysis to create and evaluate a measurement instrument for foreign language classroom speaking anxiety. *Research Journal of the Japan Association for Language Teaching*, *35*(1), 5–28.

Apple, M. T., & Neff, P. (2012). Using Rasch measurement to validate the Big Five factor marker questionnaire for a Japanese university population. *Journal of Applied Measurement*, *13*(3), 276–296.

Ariffin, S. R., Asari, S. M., Mohamed, S., Shahar, S. N., Ishak, N. M., Din, R., Ahmad, A. G., Yamat, H., Majid, R., & Yasin, S. F. M. (2010). Validity of UKM1 intelligence test using Rasch analysis. *Procedia - Social and Behavioral Sciences*, *7*(2), 205–209. https://doi.org/10.1016/j.sbspro.2010.10.029

Arphorn, S., Augsornpeug, N., Srisorrachatr, S., & Pruktharathikul, V. (2003). Comprehension of safety signs for construction workers: comparison of existing and newly designed signs. *Journal of Human Ergology*, *32*(2), 87–94.

Ayele, D. G., Zewotir, T., & Mwambi, H. (2014). Using rasch modeling to re-evaluate rapid malaria diagnosis test analyses. *International Journal of Environmental Research and Public Health*, *11*(7), 6681–6691. https://doi.org/10.3390/ijerph110706681

Aziz, A. A., Masodi, M. S., & Zaharim, A. (2013). *Asas model mengukuran Rasch: Pembentukan skala & struktur pengukuran*. Bangi: Penerbit Universiti kebangsaaan Malaysia.

Badia, X., Prieto, L., & Linacre, J. M. (2002). Differential item functioning and differential test functioning (DIF - DTF). *Rasch Measurement Transactions*, *16*(3), 889.

Baghaei, P. (2008). The Rasch model as a construct validation tool. *Rasch Measurement Transactions*, *22*(1), 1145–1146. Retrieved from http://www.rasch.org/rmt/rmt221a.htm

Bailin, A., & Grafstein, A. (2001). The linguistic assumptions underlying readability formulae: A critique. *Language and Communication*, *21*(3), 285–301. https://doi.org/10.1016/S0271-5309(01)00005-2

Bangor, A., Kortum, P., & Miller, J. (2009). Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of Usability Studies*, *4*(3), 114–123.

Bangor, A., Kortum, P. T. P., & Miller, J. J. T. (2008). An empirical evaluation of the System Usability Scale. *International Journal of Human-Computer Interaction*, *24*(6), 574–594. https://doi.org/10.1080/10447310802205776

Bazire, M., Tijus, C., Brézillon, P., & Cambon, B. (2005). Legal meanings and driver's interpretation of road signs. In *Proceedings of the 13th International Conference on Road Safety on Four continents* (pp. 5–7). Warsaw, Poland.

Bazire, M., Tijus, C., & Lavalette, B. C. (2006). Modeling drivers' decision-making processes from road signs. In *Joint International Conference on Computing and Decision Making in Civil and Building Engineering* (pp. 2584–2593). Montréal, Canada.

Belvedere, S. L., & De Morton, N. A. (2010). Application of Rasch analysis in health care is increasing and is applied for variable reasons in mobility instruments. *Journal of Clinical Epidemiology*, *63*(12), 1287–1297. https://doi.org/10.1016/j.jclinepi.2010.02.012

Ben-Bassat, T., & Shinar, D. (2006). Ergonomic guidelines for traffic sign design increase sign comprehension. *The Journal of the Human Factors and Ergonomics Society*, *7*(4–5), 209–227. https://doi.org/10.1016/j.trf.2004.09.001

Blachman, N., & Jerry, P. (2007). The * Operator. Retrieved January 25, 2015, from www.GoogleGuide.com

Blackman, I., Bradman, L., & Graetz, L. (2015). Assessment of the psychometric properties of the Tampa scale of kinesiophobia adapted for people with dystonia using Rasch analysis. In *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings* (pp. 173–188). Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-37592-7

Bolderston, A. (2008). Writing an effective literature review. *Journal of Medical Imaging and Radiation Sciences*, *39*(2), 86–92. https://doi.org/10.1016/j.jmir.2008.04.009

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch Model: Fundamental measurement in the human sciences* (3rd ed.). New York: Routledge.

Boone, H. N., & Boone, D. A. (2012). Analyzing Likert data. *Journal of Extension*, *50*(2), 1-5. Retrieved from http://www.joe.org/joe/2012april/tt2.php

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch Analysis in the human sciences*. Dordrecht: Springer Science & Business Media. https://doi.org/10.10007/978-94-007-6857-4_1

Borowsky, A., Shinar, D., & Parmet, Y. (2008). Sign location, sign recognition, and driver expectancies. *Transportation Research Part F: Traffic Psychology and Behaviour*, *11*(6), 459–465. https://doi.org/10.1016/j.trf.2008.06.003

Borsci, S., Federici, S., & Lauriola, M. (2009). On the dimensionality of the System Usability Scale: A test of alternative measurement models. *Cognitive Processing*, 193–197. https://doi.org/10.1007/s10339-009-0268-9

Bowen, G. a. (2009). Document analysis as a qualitative research method. *Qualitative Research Journal*, *9*(2), 27–40. https://doi.org/10.3316/QRJ0902027

Brancato, G., Macchia, M., Murgia, M., Signore, M., Simeoni, G., Blanke, K., Körner, T., Nimmergut, A., Lima, P., Paulino, R. Hoffmeyer-Zlotnik, J. (2006). Handbook of recommended practices for questionnaire development and testing in the European statistical system. *European Commission Grant Agreement*, *1*, 142.

Brémond, R., & Deugnier, M. (2006). Saliency of road signs in urban areas. In *Proceedings of the 11th International Conference on Vision in Vehicles* (pp. 1–8). Dublin.

Brézillon, P., Lavalette, B. C. de, Tijus, C., Leproux, C., Lacaste, A., & Bazire, M. (2004). External and internal representations of road pictographic signs. In L. Magnani & R. Dossena (Eds.), *Computing, philosophy and cognition* (pp. 1–12). Paiv: College Publications.

Brooke, J. (1996). SUS-A quick and dirty usability scale. In & A. L. M. P. W. Jordan, B. Thomas, B. A. Weerdmeester (Ed.), *Usability evaluation in industry*. London: Taylor and Francis.

Brooke, J. (2013). SUS: a retrospective. *Journal of Usability Studies*, *8*(2), 29–40.

Brown, J. (2011). Likert items and scales of measurement. *Statistics*, *15*(3), 20–24.

Brutus, S., Aguinis, H., & Wassmer, U. (2013). Self-reported limitations and future directions in scholarly reports : Analysis and recommendations. *Journal of Management*, *39*(1), 48–75. https://doi.org/10.1177/0149206312455245

Campbell, B., & Cowley, J. (2007). Response time and decision accuracy for "No Turn on Red" signs. In *Proceedings of the Human Factors and Ergonomics Society 51st Annual Meeting* (pp. 1439–1443). Baltimore, Maryland.

Carifio, J., & Perla, R. (2007). Ten common misunderstandings, misconceptions, persistent myths and urban legends about Likert scales and Likert response formats and their antidotes. *Journal of Social Sciences*, *3*(3), 106–116.

Castro, C., Moreno-Ríos, S., & Tornay, F. J. (2012). Cognitive representations of obligation and prohibition signs when they provide the same amount of semantic information. *Psicologica: International Journal of Methodology and Experimental Psychology*, *33*(2), 275–291.

Chan, A. H. S., & Ng, A. W. Y. (2012). The guessing of mine safety signs meaning: Effects of user factors and cognitive sign. *International Journal of Occupational Safety and Ergonomics*, *18*(2), 195–208.

Chan, K. L., & Chan, A. H. S. (2011). Understanding industrial safety signs: implications for occupational safety management. *Industrial Management & Data Systems*, *111*(9), 1481–1510. https://doi.org/10.1108/02635571111182809

Chang, D., Dooley, L., & Tuovinen, J. E. (2002). Gestalt Theory in Visual Screen Design — A New Look at an old subject. *Proceedings of the Seventh World Conference on Computers in Education Conference on Computers in Education: Australian Topics*, *8*, 5–12.

Charlton, S., & Baas, P. (2006). *Assessment of hazard warning signs used on New Zealand roads*. Wellington: Land Transport New Zealand. Retrieved from http://trid.trb.org/view.aspx?id=803865

Charlton, S. G. (2006). Conspicuity, memorability, comprehension, and priming in road hazard warning signs. *Accident Analysis and Prevention*, *38*(3), 496–506. https://doi.org/10.1016/j.aap.2005.11.007

Chen, W. H., Lenderking, W., Jin, Y., Wyrwich, K. W., Gelhorn, H., & Revicki, D. A. (2014). Is Rasch model analysis applicable in small sample size pilot studies for assessing item characteristics? An example using PROMIS pain behavior item bank data. *Quality of Life Research*, *23*(2), 485–493. https://doi.org/10.1007/s11136-013-0487-5

Child, D., & Colmer, R. (2011). Measure text readability. Retrieved November 4, 2015, from https://readability-score.com/text/

Clarkson, J., Coleman, R., Hosking, I., & Waller, S. (2013). *Inclusive design toolkit*. Cambrdge: Engineering Design Centre, University of Cambridge.

Clason, D., & Dormody, T. (1994). Analyzing data measured by individual Likert-type items. *Journal of Agricultural Education*, *35*(4), 31–35.

Clauser, B., & Linacre, J. M. (1999). Relating Cronbach and Rasch reliabilities. *Rasch Measurement Transactions*, *13*(2), 696.

Contreras, R. B. (2011). Examining the context in qualitative analysis: The Role of the co-occurrence tool in ATLAS.ti. *Atlas.Ti Newsletter*, *2*(8), 5–6.

Costa, M., Simone, A., Vignali, V., Lantieri, C., Bucchi, A., & Dondi, G. (2014). Looking behavior for vertical road signs. *Transportation Research Part F: Traffic Psychology and Behaviour*, *23*, 147–155. https://doi.org/10.1016/j.trf.2014.01.003

Couper, M. P. (2006). Evaluating the effectiveness of visual analog scales: A Web experiment. *Social Science Computer Review*, *24*(2), 227–245. https://doi.org/10.1177/0894439305281503

Crundall, D., & Underwood, G. (2001). The priming function of road signs. *Transportation Research*

Part F: Traffic Psychology and Behaviour, *4*(3), 187–200. https://doi.org/10.1016/S1369-8478(01)00023-7

Cunningham, T. J., Berkman, L. F., Gortmaker, S. L., Kiefe, C. I., Jacobs, D. R., Seeman, T. E., & Kawachi, I. (2011). Assessment of differential item functioning in the experiences of discrimination index: The Coronary Artery Risk Development in Young Adults (CARDIA) Study. *American Journal of Epidemiology*, *174*(11), 1266–1274. https://doi.org/10.1093/aje/kwr253

Curtin, M., Browne, J., Staines, A., & Perry, I. J. (2015). The early development instrument: An Evaluation of its five domains using Rasch analysis. *BioMed Central Pediatrics*, *16*(1), 10. https://doi.org/10.1186/s12887-016-0543-8

Dencker, A., Sunnerhagen, K. S., Taft, C., & Lundgren-Nilsson, A. (2015). Multidimensional fatigue inventory and post-polio syndrome - a Rasch analysis. *Health and Quality of Life Outcomes*, *13*, 20. https://doi.org/10.1186/s12955-015-0213-9

Dissanayake, S., & Lu, J. J. (2001). Traffic control device comprehension: Differences between domestic and international Drivers in USA. *IATSS Research*, *25*(2), 80–87.

Downing, A. J., Baguley, C. J., & Hills, B. L. (1991). Road safety in developing countries : an overview. In *Proceedings of Nineteenth Transport, Highways and Planning Summer Annual Meeting, PTPR* (pp. 1–49). London: PTRC Education and Research Services Ltd.

Duarte, E., Rebelo, F., Teles, J., & Wogalter, M. S. (2014). Safety sign comprehension by students, adult workers and disabled persons with cerebral palsy. *Safety Science*, *61*, 66–77. https://doi.org/10.1016/j.ssci.2013.10.010

DuBay, W. (2004). The principles of readability: A brief introduction to readability research. *Impact Information*, *21*, 489–508. https://doi.org/10.1.1.91.4042

Dumas, J. S., & Redish, J. C. (1999). *A practical guide to usability testing*. Exeter: Intellect Books.

Duncan, P. W., Bode, R. K., Lai, S. M., & Perera, S. (2003). Rasch analysis of a new stroke-specific outcome scale: The stroke impact scale. *Archives of Physical Medicine and Rehabilitation*, *84*(7), 950–963. https://doi.org/10.1016/S0003-9993(03)00035-2

Eakman, A. M. (2012). Measurement characteristics of the engagement in meaningful activities survey in an age-diverse sample. *American Journal of Occupational Therapy*, *66*(2), e20–e29. https://doi.org/10.1016/j.surg.2006.10.010.Use

Egleston, B. L., Miller, S. M., & Meropol, N. J. (2011). The impact of misclassification due to survey response fatigue on estimation and identifiability of treatment effects. *Statistics in Medicine*, *30*(30), 3560–3572. https://doi.org/10.1002/sim.4377

Eklund, M., Erlandsson, L.-K., Persson, D., & Hagell, P. (2009). Rasch analysis of an instrument for measuring occupational value: Implications for theory and practice. *Scandinavian Journal of Occupational Therapy*, *16*(2), 118–28. https://doi.org/10.1080/11038120802596253

Elliot, A. J., & Maier, M. A. (2014). Color Psychology: Effects of Perceiving Color on Psychological Functioning in Humans. *Annual Review of Psychology*, *65*(1), 95–120. https://doi.org/10.1146/annurev-psych-010213-115035

Estabrooks, C., & Wallin, L. (2004). Where do we stand on the measurement of research utilization. In *4th Annual knowledge utilization Colloquia (KU04)*. Belfast, North Ireland.

Eusofe, Z., & Evdorides, H. (2017). Assessment of road safety management at institutional level in Malaysia: A case study. *Journal of International Association of Traffic and Safety Sciences*, *41*(4), 172–181. https://doi.org/10.1016/j.iatssr.2017.03.002

Evans, A., & Rooney, B. (2014). *Methods in psychological research* (3rd ed.). Los Angeles: SAGE Publications.

Fereday, J., & Muir-Cochrane, E. (2006). Demonstrating rigor using thematic analysis: A Hybrid approach of inductive and deductive coding and theme development. *International Journal of Qualitative Methods*, *5*(1), 80–92. https://doi.org/10.1063/1.2011295

Finstad, K. (2006). The system usability scale and non-native english speakers. *Journal of Usability Studies*, *1*(4), 185–188.

Finstad, K. (2010a). Response interpolation and scale sensitivity: Evidence against 5-point scales. *Journal of Usability Studies*, *5*(3), 104–110.

Finstad, K. (2010b). The usability metric for user experience. *Interacting with Computers*, *22*(5), 323–327.

Firestone, C., & Scholl, B. J. (2016). Cognition does not affect perception: Evaluating the evidence for "top-down" effects. *Behavioral and Brain Sciences*, *39*, 1–77. https://doi.org/10.1017/S0140525X15000965

Fleyeh, H. (2008). *Traffic and road sign recognition*. *Doctoral dissertation*. Napier University. Retrieved from http://dl.acm.org/citation.cfm?id=2509317

Friedman, D. B., & Hoffman-Goetz, L. (2006). A systematic review of readability and comprehension instruments used for print and web-based cancer information. *Health Education & Behavior*, *33*(3), 352–373. https://doi.org/10.1177/1090198105277329

Funkhouser, D., Chrysler, S., Nelson, A., & Park, E. S. (2008). Traffic sign legibility for different sign background colors : Results of an open road study at freeway speeds procedure research participants research activities. In *Proceedings Of The Human Factors And Ergonomics Society 52nd Annual Meeting* (pp. 1855–1859).

Galli, S., Chiesi, F., & Primi, C. (2008). The construction of a scale to measure mathematical ability in psychology students: An application of the Rasch model. *Testing Psicometria Metodologia*, *15*(1), 3–18.

Gothwal, V. K., Wright, T., Lamoureux, E. L., & Pesudovs, K. (2010). Psychometric properties of visual functioning index using Rasch analysis. *Acta Ophthalmologica*, *88*(7), 797–803. https://doi.org/10.1111/j.1755-3768.2009.01562.x

Graham, L. (2008). Gestalt Theory in Interactive Media Design. *Journal of Humanities & Social Sciences*, *2*(1), 1–12. https://doi.org/10.1145/1754393.1754394

Granger, C. V. (2008). Rasch analysis is important to understand and use for measurement. *Rasch Measurement Transactions*, *21*(3), 1122–1123.

Greenhalgh, T., & Peacock, R. (2005). Effectiveness and efficiency of search methods in systematic reviews of complex evidence: Audit of primary sources. *The British Journal of Medicine (Clinical Research Ed.)*, *331*(7524), 1064–1065. https://doi.org/10.1136/bmj.38636.593461.68

Hagell, P. (2014). Testing rating scale unidimensionality using the principal component analysis (PCA)/ t -Test Protocol with the Rasch model: the primacy of theory over statistics. *Open Journal of Statistics*, *4*, 456–465. https://doi.org/10.4236/ojs.2014.46044

Hagquist, C., & Andrich, D. (2004). Is the sense of coherence instrument applicable on adolescents? A latent trait analysis using Rasch modelling. *Personality and Individual Differences*, *36*(4), 955–968. https://doi.org/10.1016/S0191-8869(03)00164-8

Ham-Baloyi, W. ten, & Jordan, P. (2016). Systematic review as a research method in post-graduate nursing education. *Health SA Gesondheid*, *21*(0), 120–128. https://doi.org/10.1016/j.hsag .2015.08.002

Hancock, G. R., & Mueller, R. O. (2010). *The reviewer's guide to quantitative methods in the social sciences*. New York: Routledge.

Hancock, H. E., Rogers, W. a, Schroeder, D., & Fisk, A. D. (2004). Safety symbol comprehension: effects of symbol type, familiarity, and age. *Human Factors*, *46*(2), 183–195. https://doi.org /10.1518/hfes.46.2.183.37344

Hassenzahl, M. (2006). Hedonic, emotional, and experiential perspectives on product quality. *Encyclopedia of Human Computer Interaction*, 266–268. https://doi.org/10.4018/978-1-59140-562-7.ch042

Hendriks, J., Fyfe, S., Styles, I., Rachel Skinner, S., & Merriman, G. (2012). Scale construction utilising the Rasch unidimensional measurement model: A measurement of adolescent attitudes towards abortion. *Australasian Medical Journal*, *5*(5), 251–261. https://doi.org/10.4066/AMJ.2012.952

Heydari, P., & Riazi, A. M. (2012). Readability of texts : Human evaluation versus computer index. *Mediterranean Journal of Social Sciences*, *3*(1), 177–190. https://doi.org/10.5901/mjss.2012.03.01.177

Hibbard, J. H., Stockard, J., Mahoney, E. R., & Tusler, M. (2004). Development of the Patient Activation Measure (PAM): Conceptualizing and measuring activation in patients and consumers. *Health Services Research*, *39*(4 Pt 1), 1005–1026. https://doi.org/10.1111/j.1475-6773.2004.00269.x

Ho, G., Scialfa, C. T., Caird, J. K., & Graw, T. (2001). Visual search for traffic signs: the effects of clutter, luminance, and aging. *Human Factors*, *43*(2), 194–207. https://doi.org/10.1518/001872001775900922

Hobart, J., & Cano, S. (2009). Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods. *Health Technology Assessment (Winchester, England)*, *13*(12). https://doi.org/10.3310/hta13120

Horberry, T., Castro, C., Martos, F., Mertova, P., Cándida, C., Martos, F., & Mertova, P. (2004). An Introduction to transport signs and an overview of this book. In C. Cándida & T. Horberry (Eds.), *The human factors of transport signs.* (pp. 1–13). Boca Raton: CRC press. https://doi.org/10.1201/9780203457412.ch1

Hornbæk, K., & Law, E. L.-C. (2007). Meta-analysis of correlations among usability measures. *Proceedings of the Special Interest Group for Computer-Human Interaction Conference on Human Factors in Computing Systems*, 617–626. https://doi.org/10.1145/1240624.1240722

Houldin, C. (2011). *Traffic signs policy review: Research project into the Awareness of the meaning of traffic signs*. Manchester: Department for Transport, UK. Retrieved from https://www.gov.uk/government/publications/signing-the-way-traffic-signs-policy-review

Houldin, C., Stuart, T., & Murphy, P. (2012). Road user awareness and understanding of the meaning of traffic signs. In *European Transport Conference 2012* (pp. 1–22). Glasgow.

Inman, V. W. (2012). Conspicuity of traffic signs assessed by eye tracking and immediate recall. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *56*(1), 2251–2255. https://doi.org/10.1177/1071181312561474

International Organization for Standardization. (2010). *ISO 9241-210:2010(E)*. Geneva, Switzerland: International Organization for Standardization.

Ismail, A. (2012). Comprehension of posted highway traffic signs in Iraq. *Tikrit Journal of Engineering Science*, *19*(1), 62–70.

Jalali, S., & Wohlin, C. (2012). Systematic literature studies: database searches vs. backward snowballing. In *Proceedings of the 2012 6th ACM_IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)* (pp. 29–38). Lund. https://doi.org/10.1145/2372251.2372257

Jamieson, S. (2004). Likert scales: how to (ab)use them. *Medical Education*, *38*(12), 1217–8. https://doi.org/10.1111/j.1365-2929.2004.02012.x

Jamson, S., Tate, F., & Jamson, A. . (2005). Evaluating the effects of bilingual traffic signs on driver performance and safety. *Ergonomics*, *48*(15), 1734–1748. https://doi.org/10.1080/00140130500142191

Jann, B. (2005). Tabulation of multiple responses. *Stata Journal*, *5*(1), 92–122.

Jeng, J. (2005). Usability assessment of academic digital libraries: Effectiveness, efficiency, satisfaction, and learnability. *Libri*, *55*(2–3), 96–121. https://doi.org/10.1515/LIBR.2005.96

Jian, R., Ruili, G., Zhonghua, W., & Yongyan, W. (2011). Study on traffic sign classification for assessing sign comprehension. *Proceedings 2011 International Conference on Transportation, Mechanical, and Electrical Engineering (TMEE)*, 1831–1835. https://doi.org/10.1109/TMEE .2011.6199570

Jones, S., & Cottrell, B. (1999). Assessment of advanced warning signs for flagging operations. *Ghana Medical Journal*, *45*(3), 1–13.

Karami, H. (2012). An Introduction to Differential Item Functioning. *The International Journal of Eductional and Psychological Assessment*, *11*(2), 59–76.

Karwowski, W., Weaver, J. L., Gielo-perczak, K., Washburn, D. A., & Hancock, P. A. (2003). Considering the importance of individual differences in Human Factors research: No Longer simply confounding noise. In *Proceedings of the Human Factors and Ergonomics Society 47th Annual Meeting* (pp. 1082–1086).

Katz, B., Dagnall, E., & Bertola, M. A. (2012). *Comprehension and legibility of selected symbol signs: Phase 3*. Reston, Virginia. Retrieved from http://www.pooledfund.org/Document /Download/3557

Kendel, F., Wirtz, M., Dunkel, A., Lehmkuhl, E., Hetzer, R., & Regitz-Zagrosek, V. (2010). Screening for depression: Rasch analysis of the dimensional structure of the PHQ-9 and the HADS-D. *Journal of Affective Disorders*, *122*(3), 241–246. https://doi.org/10.1016/j.jad .2009.07.004

Ketola, P., & Röykkee, M. (2001). The three facets of usability in mobile handsets. In *Computer-Human Interaction Workshop 2011: Mobile Communications: Understanding Users, Adoption & Design*. Seattle.

Khan, A., Yavorsky, C., Liechti, S., Opler, M., Rothman, B., DiClemente, G., Lucic, L., Jovic, S., Inada, T. & Yang, L. (2013). A rasch model to test the cross-cultural validity in the positive and negative syndrome scale (PANSS) across six geo-cultural groups. *BioMed Central Psychology*, *1*(1), 5. https://doi.org/10.1186/2050-7283-1-5

Kirakowski, J., & Corbett, M. (1993). SUMI: the Software Usability Measurement Inventory. *British Journal of Educational Technology*, *24*(3), 210–212. https://doi.org/10.1111/j.1467-8535.1993.tb00076.x

Koh, W., Yen, S., & Lan, O. S. (2012). Validating Student-parent Actions Model (Spa) of consistency using the Rasch Measurement Model. *Asia Pasific Journal of Educators and Education*, *27*, 119–138.

Kook, S. H., & Varni, J. W. (2008). Validation of the Korean version of the pediatric quality of life inventory™ 4.0 (PedsQL™) generic core scales in school children and adolescents using the rasch model. *Health and Quality of Life Outcomes*, *6*(1), 41. https://doi.org/10.1186/1477-7525-6-41

Koyuncu, M., & Amado, S. (2008). Effects of stimulus type, duration and location on priming of road signs: Implications for driving. *Transportation Research Part F: Traffic Psychology and Behaviour*, *11*(2), 108–125. https://doi.org/10.1016/j.trf.2007.08.005

Kurniawan, S., & Zaphiris, P. (2001). Investigating the age effects on subjective assessments of traffic signs. In *Proceedings of the Institute of Industrial Engineers Annual Conference* (pp. 20–23). Dallas.

Kurtaiş, Y., Őztuna, D., Küçükdeveci, A. A., Kutlay, Ş., Hafiz, M., & Tennant, A. (2011). Reliability, construct validity and measurement potential of the ICF comprehensive core set for osteoarthritis. *BioMed Central Musculoskeletal Disorders*, *12*(1), 255. https://doi.org/10.1186/1471-2474-12-255

Lajunen, T., & Özkan, T. (2011). Self-report instruments and methods. In B. E. Porter (Ed.), *Handbook of Traffic Psychology* (pp. 43–59). London: Elsevier Inc. https://doi.org/10.1016/B978-0-12-381984-0.10004-9

Lamoureux, E. L., Pallant, J. F., Pesudovs, K., Hassell, J. B., & Keeffe, J. E. (2006). The impact of vision impairment questionnaire: An Evaluation of its measurement properties using Rasch analysis. *Investigative Ophthalmology & Visual Science*, *47*(11), 4732–4741. https://doi.org/10.1167/iovs.06-0220

Lansdown, T. (2004). Considerations in evaluation and design of roadway signage from the perspective of driver attentional allocation. In C. Castro & T. Horberry (Eds.), *The human factors of transport signs* (pp. 71–80). Boca Raton: CRC Press.

Laughery, K. R., Paige, D. L., Laughery, B. R., Wogalter, M. S., Kalsher, M. J., & Leonard, S. D. (2002). Guidelines for warnings design: do they matter? In *Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting* (pp. 1708–1712).

Laughlin, H. M. (1969). SMOG Grading- a New readability formula. *Journal of Reading*, *12*(8), 639–646.

Law, T. H., Radin Umar, R. S., & Wong, S. V. (2005). The Malaysian government's road accident death reduction target for year 201. *International Association of Traffic and Safety Sciences Research*, *29*(1), 42–49. https://doi.org/10.1016/S0386-1112(14)60117-5

Lay, M. G. (2009). Design of traffic signs. In *Handbook of road technology* (4th ed.). London: Spon Press.

Lee, C., Chuang, Y., & Young, M. (2013). Legibility of traffic warning signs by drivers. *Bulletin of Japanese Society for the Science of Design*, *60*(2), 21–28.

Lerdal, A., Fagermoen, M., Bonsaksen, T., Gay, C. L., & Kottorp, A. (2014). Rasch analysis of the sense of coherence scale in a sample of people with morbid obesity – a cross-sectional study. *BioMed Central Psychology*, *2*(1), 1. https://doi.org/10.1186/2050-7283-2-1

Lerdal, A., Kottorp, A., Gay, C., Aouizerat, B. E., Lee, K. A., & Miaskowski, C. (2016). A Rasch analysis of assessments of morning and evening fatigue in oncology patients using the Lee Fatigue Scale. *Journal of Pain and Symptom Management*, *51*(6), 1002–1012. https://doi.org/10.1016/j.jpainsymman.2015.12.331

Lesch, M. F. (2003). Comprehension and memory for warning symbols: Age-related differences and impact of training. *Journal of Safety Research*, *34*(5), 495–505. https://doi.org/10.1016/j.jsr.2003.05.003

Lesch, M. F. (2004). Age-related effects on warning symbol comprehension. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *48*(2), 233–237. https://doi.org/10.1177/154193120404800202

Lesch, M. F. (2008). A comparison of two training methods for improving warning symbol comprehension. *Applied Ergonomics*, *39*(2), 135–143. https://doi.org/10.1016/j.apergo.2007.07.002

Lesch, M. F., Horrey, W. J., Wogalter, M. S., & Powell, W. R. (2011). Age-related differences in warning symbol comprehension and training effectiveness: effects of familiarity, complexity, and comprehensibility. *Ergonomics*, *54*(10), 879–890. https://doi.org/10.1080/00140139.2011.606924

Lesch, M. F., Powell, W. R., Horrey, W. J., & Wogalter, M. S. (2013). The use of contextual cues to improve warning symbol comprehension: making the connection for older adults. *Ergonomics*,

56(8), 1264–79. https://doi.org/10.1080/00140139.2013.802019

Lewis-Beck, M., Bryman, A. E., & Liao, T. F. (2003). *The SAGE encyclopedia of social science research methods*. Thousand Oaks: SAGE Publications.

Lewis, J. R. (1992). Psychometric evaluation of the post-study system usability questionnaire: The PSSUQ. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *36*(16), 1259–1260.

Lewis, J. R. (1995). IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction, 7*(1), 57–78.

Lewis, J. R., & Sauro, J. (2009). The factor structure of the system usability scale. In *International Conference on Human Centered Design* (pp. 94–103). Springer Berlin Heidelberg.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *140*, 22–55.

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions*, *7*(4), 328.

Linacre, J. M. (1996). True-Score Reliability or Rasch Validity? *Rasch Measurement Transactions*, *9*(4), 455.

Linacre, J. M. (1997). KR-20 / Cronbach alpha or Rasch person reliability: Which tells the truth? *Rasch Measurement Transactions*, *11*(3), 580–581.

Linacre, J. M. (1998). Detecting multidimensionality: Which residual data-typeworks Best? *Journal of Outcome Measurement*, *2*(3), 266–283.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean. *Rasch Measurement Transactions*, *16*(2), 878.

Linacre, J. M. (2006). Reliability and separation of measures. In *A user's guide to WINSTEPS MINISTEP Rasch-model computer programs*. Chicago IL: Winstep.com. Retrieved from http://www.winsteps.com/winman/reliability.htm 3/7

Linacre, J. M. (2011). Rasch measures and unidimensionality. *Rasch Measurement Transactions, 24*(4), 1310.

Linacre, J. M. (2012). *A user's guide to Winsteps Ministeps Rasch-model computer programs [version 3.74. 0].* Chicago IL: Winstep.com.

Liu, Y.-C. (2005). A simulated study on the effects of information volume on traffic signs, viewing strategies and sign familiarity upon driver's visual search performance. *International Journal of Industrial Ergonomics, 35*(12), 1147–1158. https://doi.org/10.1016/j.ergon.2005.06.009

Lo, C., Liang, W. M., Hang, L. W., Wu, T. C., Chang, Y. J., & Chang, C. H. (2015). A psychometric assessment of the St. George's respiratory questionnaire in patients with COPD using Rasch model analysis. *Health and Quality of Life Outcomes, 13,* 131. https://doi.org/10.1186/s12955-015-0320-7

Mailloux, S. L., Johnson, M. E., Fisher, D. G., & Pettibone, T. J. (1995). How reliable is computerized assessment of readability? *Computers in Nursing*, *13*(5), 221–225.

Makinde, O., & Oluwasegunfunmi, V. (2014). Comprehension of traffic devices amongst urban drivers- A study of Ado-Ekiti, Ekiti State, Nigeria. *European Journal of Engineering and Technology, 2*(1), 9–19.

Makinde, O., & Opeyemi, D. A. (2012). Understanding of traffic signs by drivers: a Case of Akure City, Ondo State, Nigeria. *Asian Research Publishing Network Journal of Science and Technology, 2*(7), 608–612.

Malaysian Department of Road Transport. (2016). *Kurikulum pendidikan pemandu* (2nd ed.). Melaka: Akademi Pengangkutan Jalan Malaysia.

Malaysian Institute of Road Safety Research. (2016). Official website of Malaysian Institute of Road Safety Research (MIROS). Retrieved January 9, 2017, from https://www.miros.gov .my/1/page.php?id=10

McCreary, L. L., Conrad, K. M., Conrad, K. J., Scott, C. K., Funk, R. R., & Dennis, M. L. (2013). Using the Rasch Measurement Model in psychometric analysis of the family effectiveness measure. *Nursing Research*, *62*(3), 149–159. https://doi.org/DOI10.1097/NNR.0b013 e31828eafe6

McDougald, B. R., & Wogalter, M. S. (2011). Increased comprehension of warning pictorials with color highlighting. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *55*(1), 1769–1772. https://doi.org/10.1177/1071181311551367

McDougall, S. J., Curry, M. B., & de Bruijn, O. (1999). Measuring symbol and icon characteristics: norms for concreteness, complexity, meaningfulness, familiarity, and semantic distance for 239 symbols. *Behavior Research Methods, Instruments, & Computers : A Journal of the Psychonomic Society, Inc*, *31*(3), 487–519.

Mcdougall, S. J. P., & Curry, M. B. (2001). The effects of visual information on users ' mental models : an Evaluation of pathfinder analysis as a measure of icon usability. *International Journal of Cognitive Ergonomics*, *5*(1), 59–84. https://doi.org/10.1207/S15327566IJCE0501

McDougall, S. J., Tyrer, V., & Folkard, S. (2006). Searching for signs, symbols, and icons: effects of time of day, visual complexity, and grouping. *Journal of Experimental Psychology*, *12*(2), 118–128. https://doi.org/10.1037/1076-898X.12.2.118

Meyer, J. P. (2014). *Applied Measurement with JMetrik*. New York: Routledge.

Miller, D. B., Webster, S. E., & MacIntosh, R. (2002). What's there and what's not: Measuring daily hassles in urban African-American adolescents. *Research on Social Work Practice*, *12*(3), 375–388. https://doi.org/10.1177/1049731502012003003

Mis, M. A. (2010). Lingua Franca in Sarawak : the Application of language choice theory. *GEMA Online Journal of Language Studies*, *10*(2), 97–116.

Miska, E., Leur, P. de, & Sayed, T. (2002). Road safety performance associated with improved traffic signal design and increased signal conspicuity. In *Institute of Transportation Engineers 2002 Annual Meeting and Exhibit* (pp. 1–11). Philadelphia.

Mitchell, M. (2010). An analysis of road signage and advertising from a pragmatic visual communication perspective : Case study of the M1 Motorway between the Gold Coast and Brisbane. *Journal of the Australasian College of Road Safety*, *21*(2), 55–64.

Mogey, N. (1999). So you want to use a Likert scale. *Learning Technology Dissemination Initiative*, 21. Retrieved from http://www.icbl.hw.ac.uk/ltdi/cookbook/info_likert_scale/printable.pdf

Moroni, J. L. D. S., & Aymone, J. L. F. (2012). Ergonomics and design: traffic sign and street name sign. *Work (Reading, Mass.)*, *41 Suppl 1*, 1541–6. https://doi.org/10.3233/WOR-2012-0350-1541

Mukhopadhyay, P., Kaur, J., Kaur, L., Arvind, A., Kajabaje, M., Mann, J., … Chakravarty, S. (2013). Ergonomic design analysis of some road signs in India. *Information Design Journal*, *20*(3), 220–227. https://doi.org/10.1075/idj.20.3.03muk

Mulcahy, J., & Vaughan, B. (2015). Exploring the construct validity of the Patient Perception Measure - Osteopathy (PPM-O) using classical test theory and Rasch analysis. *Chiropractic & Manual Therapies*, *23*(1), 6. https://doi.org/10.1186/s12998-015-0055-x

Mulrow, C. D. (1994). Rationale for systematic reviews. *BMJ (Clinical Research Ed.)*, *309*(6954), 597–9.

Mustafa, M. N. (2005). Overview of current road safety situation in Malaysia. *Highway Planning Unit, Road Safety Section, Ministry of Works*, 5–9.

Netemeyer, R. G., Bearden, W. O., & Subhash, S. (2003). *Scaling procedures: Issues and applications*. Thousand Oaks: Sage Publications.

Ng, A. W. Y., & Chan, A. H. S. (2007a). Cognitive design features on traffic signs. *Engineering Letters*, *14*(1), 97–102.

Ng, A. W. Y., & Chan, A. H. S. (2007b). Culture issues in traffic sign usability. In *International Conference on Usability and Internationalization* (pp. 379–387). Berlin, Heidelberg: Springer.

Ng, A. W. Y., & Chan, A. H. S. (2007c). Symbol training in past 40 years. *International Association of Engineers International Journal of Applied Mathematics*, *36*(2), 37–41.

Ng, A. W. Y., & Chan, A. H. S. (2008). Visual and cognitive features on icon effectiveness. In *Proceedings of the International Multiconference of Engineers and Computer Scientists* (Vol. 2, pp. 19–21). Hong Kong.

Ng, A. W. Y., & Chan, A. H. S. (2011). Investigation of the effectiveness of traffic sign training in terms of training methods and sign characteristics. *Traffic Injury Prevention*, *12*(3), 283–295. https://doi.org/10.1080/15389588.2011.556171

Ng, A. W. Y., & Chan, A. H. S. (2013). Recognition performance for damaged safety signs with different levels of color deterioration. In *Proceedings of the International MultiConference of Engineers and Computer Scientists* (Vol. 2). Hong Kong.

Ng, A. W. Y., Lo, H., & Chan, A. H. S. (2011a). Measuring the Usability of Safety Signs: A use of system usability scale (SUS). In *Proceedings of the International MultiConference of Engineers and Computer Scientists* (Vol. 2, pp. 19–21).

Ng, A. W. Y., Lo, H., & Chan, A. H. S. (2011b). Usability assessment of safety signs with the system

usability scale (SUS): the influence of demographic factors. *In Iaeng Transactions on Engineering Technologies of the International Multiconference of Engineers and Computer Scientists*, *7*, 271–283.

Nielson, J. (2012). Usability 101: Introduction to Usability. Retrieved March 28, 2017, from https://www.nngroup.com/articles/usability-101-introduction-to-usability/

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education : Theory and Practice*, *15*(5), 625–32. https://doi.org/10.1007/s10459-010-9222-y

Obeidat, M. S., Rys, M. J., Rys, A., & Du, J. (2016). Evaluation of overhead guide sign sheeting materials to increase visibility and safety for drivers. *Applied Ergonomics*, *56*, 136–143. https://doi.org/10.1016/j.apergo.2016.03.016

Ohrbach, R. (2008). The Jaw Functional Limitation Scale: development, reliability, and validity of 8- item and 20-item versions. *Journal of Orofacial Pain*, *22*(3), 219–230.

Ou, Y.-K., & Liu, Y.-C. (2012). Effects of sign design features and training on comprehension of traffic signs in Taiwanese and Vietnamese user groups. *International Journal of Industrial Ergonomics*, *42*(1), 1–7. https://doi.org/10.1016/j.ergon.2011.08.009

Owen, R., & Brooks, L. L. (2008). *Answering the ultimate question : how Net Promoter can transform your business*. San Francisco: Jossey-Bass.

Page, D. (2008). Systematic literature searching and the bibliographic database haystack. *Electronic Journal of Business Research Methods*, *6*(2), 171–180. https://doi.org/10.1177/01655515145219 36

Pallant, J. (2010). *SPSS survival manual: A step by step guide to data analysis using SPSS*. McGraw-Hill International.

Pallant, J. F., Miller, R. L., & Tennant, A. (2006). Evaluation of the Edinburgh Post Natal Depression Scale using Rasch analysis. *BioMed Central Psychiatry*, *6*, 28. https://doi.org/10.1186/1471-244X-6-28

Peschke, C., Olk, B., & Hilgetag, C. C. (2013). Should I stay or should I go--cognitive conflict in multi-attribute signals probed with East and West German "Ampelmännchen" traffic signs. *PloS One*, *8*(5), e64712. https://doi.org/10.1371/journal.pone.0064712

Peterson, D. J., & Berryhill, M. E. (2013). The Gestalt principle of similarity benefits visual working memory. *Psychonomic Bulletin and Review*, *20*(6), 1282–1289. https://doi.org/10.3758/s13423-013-0460-x

Petrillo, J., Cano, S. J., McLeod, L. D., & Coon, C. D. (2015). Using classical test theory, item response theory, and rasch measurement theory to evaluate patient-reported outcome measures: A comparison of worked examples. *Value in Health*, *18*(1), 25–34. https://doi.org/10.1016/j.jval.2014.10.005

Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences: a Practical guide*. *Blackwell Publishing Ltd*. Malden: Blackwell Publishing Ltd. https://doi.org/10.1027/1016-9040.11.3.244

Pinna, B. (2010). New Gestalt principles of perceptual organization: an extension from grouping to shape and meaning. *Gestalt Theory*, *23*(1), 11–78. Retrieved from http://gestalttheory.net/download/Pinna - GestaltTheory32(1).pdf

Plašil, M., & Vlach, P. (2006). Analysis of multiple-response data. In *9th International Scientific Conference on Applications of Mathematics and Statistics in Economics*. Trutnov.

Podany, K. I., Wogalter, M. S., & Mayhorn, C. B. (2004). Perceived Effectiveness of "No Turn on Red" Traffic Signs. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *48*(16), 2004–2007. https://doi.org/10.1177/154193120404801645

Polgar, S., & Thomas, S. A. (2008). *Introduction to research in the health sciences* (5th ed.). Philadelphia: Elsevier.

Porter, S. R., Whitcomb, M. E., & Weitzer, W. H. (2004). Multiple surveys of students and survey fatigue. *New Directions for Institutional Research*, *2004*(121), 63–73. Retrieved from https://oia.unm.edu/surveys/survey-fatigue.pdf

Public Works Department Malaysia. (1985a). *Manual on traffic control devices : Road marking and dileanation*. *Arahan Teknik (Jalan) 2D/85*. Kuala Lumpur: Roads Branch, Public Works Department Malaysia.

Public Works Department Malaysia. (1985b). *Manual on traffic control devices: standard traffic signs. Arahan Teknik (Jalan) 2A/85* (AT(J)2A/85). Kuala Lumpur: Cawangan Jalan JKR.

Public Works Department Malaysia. (1985c). *Manual on traffic control devices: Traffic sign application*. *Arahan Teknik (Jalan) 2B/85* (AT(J)2B/85). Kuala Lumpur: Cawangan Jalan JKR.

Purduski, J. M., & Rys, M. J. (1999). Evaluations of a new Advance Flagger traffic sign. *International Journal of Industrial Ergonomics*. https://doi.org/10.1016/S0169-8141(98)00092-4

Radhakrishna, R. B. (2007). Tips for developing and testing questionnaires/instruments. *Journal of Extension*, *45*(1), 1–4.

Razzak, A., & Hasan, T. (2010). Motorist understanding of traffic signs : a study in Dhaka city. *Journal of Civil Engineering*, *38*(1), 17–29.

Reips, U., & Funke, F. (2008). Interval-level measurement with visual analogue scales in Internet-based research: VAS Generator. *Behavior Research Methods*, *40*(3), 699–704. https://doi.org/10.3758/BRM.40.3.699

Reisinger, M. (2015). Optimal luminance of internally illuminated wayfinding signs. *Lighting*

Research and Technology, 1–13. https://doi.org/10.1177/1477153515619467

Revicki, D. A., Chen, W. H., & Gelhorn, H. (2013). Rasch Model analysis and small sample size. In D. Neasham & A. Cox (Eds.), *Evidence Matters: Innovative approaches to database study design and analysis* (pp. 19–22). Blue Bell, Pennsylvania: United BioSource Corporation.

Ridley, D. (2008). *The literature review: a Step-by-step guide for students*. Los Angeles: SAGE.

Road Safety Department of Malaysia. (2013). Official website Road Safety Department of Malaysia: Background. Retrieved January 9, 2017, from http://www.jkjr.gov.my/en/mengenai _kami/latarbelakang.html

Road Safety Department of Malaysia. (2014). Road safety plan of Malaysia 2014-2020. Putrajaya: JKJR. Retrieved from http://www.mot.gov.my/SiteCollectionDocuments/Darat/Road_Safety_ Plan_2014-2020_booklet-EN.pdf

Road Safety Department of Malaysia. (2018). *Buku Statistik Keselamatan Jalan Raya*. Kuala Lumpur. Retrieved from http://www.kpwkm.gov.my/kpwkm/uploads/files/Penerbitan/Buku statistik/BUKU STATISTIK 2015 KPWKM.pdf

Robertson, J. (2012). Likert-type scales, statistical methods, and effect sizes. *Communications of the ACM*, *55*(5), 6. https://doi.org/10.1145/2160718.2160721

Rohayu, S., Sharifah, A., Jamilah, M. M., & Wong, S. V. (2012). *Predicting Malaysian road fatalities for year 2020* (MMR06/2012). *Malaysian Institute of Road Safety Research*. Kuala Lumpur: Malaysian Institute of Road Safety Research.

Ryan, F., Coughlan, M., & Cronin, P. (2007). Step-by-step guide to critiquing research. Part 2: Qualitative research. *The British Journal of Nursing*, *16*(12), 738–744.

Ryu, Y. S., & Smith-jackson, T. L. (2005). Development of usability questionnaire items for mobile products and content validity. In *Proceedings of Human Computer Interaction International*

(pp. 22–27). Las Vegas.

Sagawa, K. (2002). Visual functions of older people and visibility of traffic signs. *Gerontechnology*, *1*(4), 296–299.

Sanders, M., & McCormick, E. (1993). *Human Factors in Engineering and Design* (7th ed.). New York: McGraw-Hill.

Santos, J. R. (2000). Getting the most out of multiple response questions. *Journal of Extension*, *38*(3), 3–8.

Sauro, J. (2004). The standardized universal percentile rank questionnaire (SUPR--Q). In J. Sauro & J. R. Lewis (Eds.), *Quantifying the user experience: Practical statistics for user research* (2nd ed., pp. 227–229). Amsterdam: Elsevier.

Sauro, J. (2011). *A practical guide to the system usability scale: background, benchmarks & best practices.* Denver, CO: Measuring Usability LLC.

Sauro, J., & Dumas, J. S. (2009). Comparison of three one-question, post-task usability questionnaires. In *Proceedings of the Special Interest Group on Computer–Human Interaction Conference on Human Factors in Computing Systems* (pp. 1599–1608). Boston: Association for Computing Machinery ACM.

Sauro, J., & Lewis, J. R. (2009). Correlations among prototypical usability metrics: evidence for the construct of usability. *Proceedings of the 27th International Conference on Human Factors in Computing Systems - CHI 09*, 1609. https://doi.org/10.1145/1518701.1518947

Schnell, T., Aktan, F., & Li, C. (2004). Traffic sign luminance requirements of nightime drivers for symbolic signs. *Journal of the Transportation Research Board*, (1862), 24–35. https://doi.org/http://dx.doi.org/10.3141/1862-04

Scialfa, C., McPhee, L., & Ho, G. (2000). The effects of a simulated cellular phone conversation on

search for traffic signs in an elderly sample. *Proceedings of the Symposium on Eye Tracking Research & Applications - ETRA '00*, 45–50. https://doi.org/10.1145/355017.355024

Scialfa, C., Spadafora, P., Klein, M., Lesnik, A., Dial, L., & Heinrich, A. (2008). Iconic sign comprehension in older adults: the role of cognitive impairment and text enhancement. *Canadian Journal on Aging*, *27*(3), 253–265. https://doi.org/10.3138/cja.27.3.253

Shaver, E. F., Young, S. L., Frantz, J. P., Rhoades, T. P., Hall, S. M., & Shah, R. J. (2006). Comparison of ANSI and ISO standard formats on people ' s response to product warnings. In *Proceedings of the Human Factors and Ergonomics Society 50th Annual Meeting* (Vol. 100, pp. 2197–2201).

Shinar, D., Dewar, R., Summala, H., & Zakowska, L. (2003). Traffic sign symbol comprehension: a cross-cultural study. *Ergonomics*, *46*(15), 1549–1565. https://doi.org/10.1080/00140130320001216 15

Shinar, D., & Vogelzang, M. (2013). Comprehension of traffic signs with symbolic versus text displays. *Transportation Research Part F: Traffic Psychology and Behaviour*, *18*, 72–82. https://doi.org/10.1016/j.trf.2012.12.012

Shneier, M. (2006). Road sign detection and recognition. In *IEEE Computer Society International Conference on Computer Vision and Pattern Recognition*. San Diego.

Shohov, S. P. (2004). *Advances in psychology research vol. 27*. New York: Nova Science Publisher Inc.

Sick, J. (2010). Rasch measurement in language education Part 5: Assumptions and requirements of Rasch measurement. *Japan Association for Language TeachingTesting & Evaluation Special Interest Group Newsletter*, *14*(2), 23–29.

Siegel, S. (2016). The Contents of Perception. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 201). Stanford: Metaphysics Research Lab, Stanford University.

Siegmann, P. (2005). Automatic evaluation of traffic sign visibility using SVM recognition methods. In *Proceedings of the 5th World Scientific and Engineering Aademy and Society International Conference on Signal Processing, Computational Geometry & Artificial Vision* (Vol. 5, pp. 170–175). Malta.

Simon, M. (2011). Assumptions, limitations and delimitations. In *Dissertation and scholarly research: Recipes for success* (p. 344). Seattle, WA: Dissertation Success, LLC.

Simon, M., & Goes, J. (2013). Scope, limitations, and delimitations. In *Dissertation and Scholarly Research: Recipes for Success*. Seattle, WA: Dissertation Success LLC.

Smith, A. B., Rush, R., Fallowfield, L. J., Velikova, G., & Sharpe, M. (2008). Rasch fit statistics and sample size considerations for polytomous data. *BioMed Central Medical Research Methodology*, *8*, 33. https://doi.org/10.1186/1471-2288-8-33

Smith, R. M. (1996). Polytomous mean-square fit statistics. *Rasch Measurement Transactions*, *10*(3), 516–517.

Smith, E. V. (2001). Evidence for the reliability of measures and validity of measure interpretation: a Rasch measurement perspective. *Journal of Applied Measurement*, *2*(3), 281–311.

Smith, E. V. (2002). Understanding Rasch measurement: Detecting and evaluating the impact of multidimenstionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement*, *3*(2), 205–231.

Soh, S.-E., Barker, A., Morello, R., Dalton, M., & Brand, C. (2016). Measuring safety climate in acute hospitals: Rasch analysis of the safety attitudes questionnaire. *BioMed Central Health Services Research*, *16*(1), 497. https://doi.org/10.1186/s12913-016-1744-4

Spector, P. E. (1992). *Summated rating scale construction: an Introduction*. Newbury Park, CA: SAGE Publications.

Spönemann, M. (2015). *Graph layout support for model-driven engineering* (2015/2). Kiel: Books on Demand, Norderstedt.

Stone, M. H., & Wright, B. D. (1999). *Measurement essentials*. *Measurement*. Wilmington, Delaware: Wide Range.

Sullivan, G. M., & Artino, J. (2013). Analyzing and interpreting data from Likert-type scales. *Journal of Graduate Medical Education*, *5*(4), 541–542.

Sun, L., Lu, J., Yao, L., & Rong, J. (2011). Simulation analysis on driving behavior during traffic sign recognition. *International Journal of Computational Intelligence Systems*, *4*(3), 353–360. https://doi.org/10.2991/ijcis.2011.4.3.9

Svider, P. F., Agarwal, N., Choudhry, O. J., Hajart, A. F., Baredes, S., Liu, J. K., & Eloy, J. A. (2013). Readability assessment of online patient education materials from academic otolaryngology-head and neck surgery departments. *American Journal of Otolaryngology - Head and Neck Medicine and Surgery*, *34*(1), 31–35. https://doi.org/10.1016/j.amjoto.2012.08.001

Talab, A. D., Meshkani, M., & Mofidi, A. (2013). Evaluation of the perception of workplace safety signs and effective factors. *International Journal of Occupational Hygiene*, *5*(3), 117–122.

Tang, W. K., Wong, E., Chiu, H. F. K., Lum, C. M., & Ungvari, G. S. (2005). The Geriatric Depression Scale should be shortened: Results of Rasch analysis. *International Journal of Geriatric Psychiatry*, *20*(8), 783–789. https://doi.org/10.1002/gps.1360

Tang, W. K., Wong, E., Chiu, H. F. K., & Ungvari, G. S. (2007). Rasch analysis of the scoring scheme of the HADS Depression subscale in Chinese stroke patients. *Psychiatry Research*, *150*(1), 97–103. https://doi.org/10.1016/j.psychres.2006.01.015

Teo, T. (2011). Towards greater precision in latent construct measurement: What's the Rasch? *British Journal of Educational Technology*, *42*(6), 122–124. https://doi.org/10.1111/j.1467-8535.2011.01213.x

Thomas, J., & Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. *BioMed Central Medical Research Methodology*, *8*(45), 1–10. https://doi.org/10.1186/1471-2288-8-45

Tijus, C., Barcenilla, J., Lavalette, B. C. De, & Meunier, J.-G. (2007). The design , understanding and usage of pictograms. *Studies in Writing*, *21*(17), 1–21. https://doi.org/10.1163/9789004253254_003

Todorovic, D. (2008). Gestalt principles. *Scholarpedia*, *3*(12), 5345. https://doi.org/10.4249/scholarpedia.5345

Uman, L. S. (2011). Information management for the busy practitioner: Systematic reviews and meta-analyses. *Journal of the American Academy of Child and Adolescent Psychiatry*, *20*(1), 57–59. https://doi.org/10.1016/j.revmed.2014.05.011

Vagias, W. (2006). Likert-type scale response anchors. *Clemson International Institute for Tourism and Research Development*, *1*, 3–4.

Vaughan, B., Mulcahy, J., & McLaughlin, P. (2014). The DREEM, part 2: psychometric properties in an osteopathic student population. *BioMed Central Medical Education*, *14*(1), 100-110. https://doi.org/10.1186/1472-6920-14-100

Walsh, T. M., Volsko, T. A., & Faarc, M. R. (2008). Readability Assessment of Internet-Based Consumer Health Information. *Respiratory Care*, *53*(10), 1310–1315.

Ward, S. J., Wogalter, M. S., & Mercer, A. W. (2004). Comprehension and training of international road signs. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 48, pp. 2104–2108). Los Angeles: Sage Publications.

Warmbrod, J. R. (2014). Reporting and interpreting scores derived from likert-type scales. *Journal of Agricultural Education*, *55*(5), 30–47. https://doi.org/10.5032/jae.2014.05030

Waterson, P., Pilcher, C., Evans, S., & Moore, J. (2012). Developing safety signs for children on board trains. *Applied Ergonomics*, *43*(1), 254–265. https://doi.org/10.1016/j.apergo.2011.05.012

Wiberg, M. (2007). *Measuring and detecting differential item functioning in criterion-referenced licensing test: a Theoretic comparison of methods*. Umeå: Umeå University Press. Retrieved from http://www.edusci.umu.se/digitalAssets/59/59534_em-no-60.pdf

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Erlbaum.

Wogalter, M. S. (2004). Factors that influence the effectiveness of warning signs and labels. *MORAES, Anamaria de; AMADO, Giuseppe. ColetâNea de Palestras de Convidados Internacionais e Nacionais: Ergodesign e USIHC. Rio de Janeiro: FAPERJ*, 25–36.

Wogalter, M. S., Murray, L. A., Glover, B. L., Shaver, E. F., & Carolina, N. (2002). Comprehension of different types of prohibitive safety symbols with glance exposure. In *Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting* (pp. 1753–1757). Baltimore, Maryland.

Wohlin, C. (2014). Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering* (pp. 1–10). New York: ACM Press. https://doi.org/10.1145/2601248.2601268

Wolfe, E. W., & Smith, E. V. (2007a). Instrument development tools and activities for measure validation using Rasch models: Part I - instrument development tools. *Journal of Applied Measurement*, *8*(1), 97–123.

Wolfe, E. W., & Smith, E. V. (2007b). Instrument development tools and activities for measure validation using Rasch models: Part II - validation activities. *Journal of Applied Measurement*, *8*(2), 204–234.

Wolff, J. S., & Wogalter, M. S. (1998). Comprehension of pictorial symbols: Effects of context and test method. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *40*(2), 173–186. Retrieved from http://hfs.sagepub.com/content/40/2/173.short

World Health Organization. (2009). *Global status report on road safety: time for action*. Gevena.

Wright, B. D. (1996). Reliability and separation. *Rasch Measurement Transactions*, *9*(4), 472.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*(3), 370.

Wuensch, K. L. (2005). What is a Likert Scale? and how do you pronounce "Likert?" Retrieved November 14, 2014, from http://core.ecu.edu/psyc/wuenschk/StatHelp/Likert.htm

Xie, H., Filippidis, L., Galea, E. R., Gwynne, S., Blackshields, D., & Lawrence, P. (2007). Experimental study and theoretical analysis of signage legibility distances as a function of observation angle. *Pedestrian and Evacuation Dynamics*, *17*(41), 131–143. https://doi.org/10.1177/1042391507064025

Yuan, L., Ma, Y., Lei, Z., & Xu, P. (2014). Driver's comprehension and improvement of warning signs. *Advances in Mechanical Engineering*, *2014*, 1–9. https://doi.org/10.1155/2014/582606

Zakir, U., Leonce, A., & Edirisinghe, E. (2010). Road sign segmentation based on colour spaces: A Comparative Study. In *Proceedings of the 11th International Conference Computer Graphics*

and Imaging (pp. 72–79). Innsbruck.

Zakowska, L. (2001). Perception and recognition of traffic signs in relation to drivers characteristics and safety - A case study in Poland. In *Proceedings of 14th International Co-operation on Theories and Concepts in Traffic Safety Workshop*. Caserta.

Zamanian, Z., Afshin, A., Davoudiantalab, A. H., & Hashemi, H. (2013). Comprehension of workplace safety signs : A case study in Shiraz industrial park. *Journal of Occupational Health and Epidemiology*, *2*(1), 37–43.

Zhang, T., & Chan, A. (2013). Traffic sign comprehension: a Review of influential factors and future directions for research. In *Proceedings of the International MultiConference of Engineers and Computer Scientists* (Vol. 2, pp. 1–9). Hong Kong.

Zoanetti, N., Griffin, P., Beaves, M., & Wallace, E. M. (2009). Rasch scaling procedures for informing development of a valid Fetal Surveillance Education Program multiple-choice assessment. *BioMed Central Medical Education*, *9*(1), 20. https://doi.org/10.1186/1472-6920-9-20

APPENDICES

Appendix A

Item Pool From 8 Generic Usability Instrument

Item no.	Usability Instrument
	1. After Scenario Questionnaire (ASQ) (Lewis, 1995)
1	Overall, I am satisfied with the ease of completing this task.
2	Overall, I am satisfied with the amount of time it took to complete this task.
3	Overall, I am satisfied with the support information (on-line help, messages, documentation) when completing this task.
	2. Post Study System Usability Questionnaire (PSSUQ) (Lewis, 1992, 1995)
4	Overall, I am satisfied with how easy it is to use this system
5	It was simple to use this system
6	I could effectively complete the tasks and scenarios using this system.
7	I was able to complete the tasks and scenarios quickly using this system.
8	I was able to efficiently complete the tasks and scenarios using this system.
9	I felt comfortable using this system.
10	It was easy to learn to use this system.
11	I believe I could become productive quickly using this system.
12	The system gave error messages that clearly told me how to fix problems
13	Whenever I made a mistake using the system, I could recover easily and quickly.
14	The information (such as on-line help, on-screen messages and other documentation) provided with this system was clear.
15	It was easy to find the information I needed.
16	The information provided for the system was easy to understand.
17	The information was effective in helping me complete the tasks and scenarios
18	The organization of information on the system screens was clear
19	The interface of this system was pleasant.
20	I liked using the interface of this system.
21	This system has all the functions and capabilities I expect it to have.
22	Overall, I am satisfied with this system.
	3. Computer System Usability Questionnaire (CSUQ) (Lewis, 1995)
23	Overall, I am satisfied with how easy it is to use this system
24	It was simple to use this system

25 I can effectively complete my work using this system
26 I am able to complete my work quickly using this system
27 I am able to efficiently complete my work using this system
28 I feel comfortable using this system
29 It was easy to learn to use this system
30 I believe I became productive quickly using this system
31 The system gives error messages that clearly tell me how to fix problems
32 Whenever I make a mistake using the system, I recover easily and quickly
33 The information (such as online help, on-screen messages, and other documentation) provided with this system is clear
34 It is easy to find the information I needed
35 The information provided for the system is easy to understand
36 The information is effective in helping me complete the tasks and scenarios
37 The organization of information on the system screens is clear
38 The interface of this system is pleasant
39 I like using the interface of this system
40 This system has all the functions and capabilities I expect it to have
41 Overall, I am satisfied with this system

---

4. Software usability measurement inventory (SUMI) (Kirakowski & Corbett, 1993)

42 This software responds too slowly to inputs.
43 I would recommend this software to my colleagues
44 The instructions and prompts are helpful
45 The software has at some time stopped unexpectedly
46 Learning to operate this software initially is full of problems
47 I sometimes don't know what to do next with this software
48 I enjoy my sessions with this software
49 I find that the help information given by this software is not very useful
50 If this software stops, it is not easy to restart it
51 It takes too long to learn the software commands
52 I sometimes wonder if I'm using the right command.
53 Working with this software is satisfying
54 The way that system information is presented is clear and understandable
55 I feel safer if I use only a few familiar commands or operations
56 The software documentation is very informative
57 This software seems to disrupt the way I normally like to arrange my work
58 Working with this software is mentally stimulating
59 There is never enough information on the screen when it’s needed.
60 I feel in command of this software when I am using it
61 I prefer to stick to the facilities that I know best.
62 I think this software is inconsistent
63 I would not like to use this software every day
64 I can understand and act on the information provided by this software
65 This software is awkward when I want to do something which is not standard

66 There is too much to read before you can use the software
67 Tasks can be performed in a straightforward manner using this software
68 Using this software is frustrating
69 The software has helped me overcome any problems I have had in using it.
70 The speed of this software is fast enough
71 I keep having to go back to look at the guides
72 It is obvious that user needs have been fully taken into consideration
73 There have been times in using this software when I have felt quite tense
74 The organisation of the menus or information lists seems quite logical
75 The software allows the user to be economic of keystrokes
76 Learning how to use new functions is difficult
77 There are too many steps required to get something to work
78 I think this software has made me have a headache on occasion
79 Error prevention messages are not adequate
80 It is easy to make the software do exactly what you want
81 I will never learn to use all that is offered in this software
82 The software hasn't always done what I was expecting
83 The software has a very attractive presentation
84 Either the amount or quality of the help information varies across the system
85 It is relatively easy to move from one part of a task to another
86 It is easy to forget how to do things with this software.
87 This software occasionally behaves in a way which can't be understood
88 This software is really very awkward
89 It is easy to see at a glance what the options are at each stage
90 Getting data files in and out of the system is not easy
91 I have to look for assistance most times when I use this software.

---

5. System Usability Scale (SUS) (Brooke, 1996, 2013)

92 I think that I would like to use this system frequently
93 I found the system unnecessarily complex
94 I thought the system was easy to use
95 I think that I would need the support of a technical person to be able to use this system
96 I found the various functions in this system were well integrated
97 I thought there was too much inconsistency in this system
98 I would imagine that most people would learn to use this system very quickly
99 I found the system very cumbersome to use
100 I felt very confident using the system
101 I needed to learn a lot of things before I could get going with this system

---

6. Single Ease Question (SEQ) (Sauro & Dumas, 2009)

102 Overall, how difficult or easy was the task to complete?

7. Standardized Universal Percentile Rank Questionnaire (SUPR-Q) (Sauro, 2004)

103 This website is easy to use
104 It is easy to navigate within the website.
105 I enjoy using the website.
106 I feel comfortable purchasing from this website.
107 I am able to find what I need quickly on this website.
108 I can count on the information I get on this website
109 I found the website to be attractive
110 I feel confident conducting business with this website
111 The website has a clean and simple presentation
112 The information on this website is valuable
113 The website keeps the promises it makes to me
114 I will likely return to this website in the future.
115 How likely are you to recommend this website to a friend or colleague?

---

8. The Usability Metric for User Experience (UMUX) (Finstad, 2010)

116 This system's capabilities meet my requirements.
117 Using this system is a frustrating experience.
118 This system is easy to use.
119 I have to spend too much time correcting things with this system.

Appendix B

Non-redundant Items

Item no.	Item
21	This system has all the functions and capabilities I expect it to have
37	The organization of information on the system screens is clear
40	This system has all the functions and capabilities I expect it to have
47	I sometimes don't know what to do next with this software
50	If this software stops, it is not easy to restart it
52	I sometimes wonder if I'm using the right command.
55	I feel safer if I use only a few familiar commands or operations
56	The software documentation is very informative
58	Working with this software is mentally stimulating
60	I feel in command of this software when I am using it
61	I prefer to stick to the facilities that I know best
62	I think this software is inconsistent
66	There is too much to read before you can use the software
67	Tasks can be performed in a straightforward manner using this software
68	Using this software is frustrating
69	The software has helped me overcome any problems I have had in using it
71	I keep having to go back to look at the guides
72	It is obvious that user needs have been fully taken into consideration
73	There have been times in using this software when I have felt quite tense
75	The software allows the user to be economic of keystrokes
77	There are too many steps required to get something to work
78	I think this software has made me have a headache on occasion
82	The software hasn’t always done what I was expecting
83	The software has a very attractive presentation
86	It is easy to forget how to do things with this software
87	This software occasionally behaves in a way which can’t be understood
91	I have to look for assistance most times when I use this software
96	I found the various functions in this system were well integrated
99	I found the system very cumbersome to use
100	I felt very confident using the system
109	I found the website to be attractive
113	The website keeps the promises it makes to me
116	This system’s capabilities meet my requirements

Appendix C

Layout of PSUS



What is the meaning of this sign? Please explain your answer.

		Strongly Disagree		Neutral		Strongly Agree
1	I am confident that I understand the meaning of this sign.	1	2	3	4	5
2	I need help in understanding this sign for the first time.	1	2	3	4	5
3	I am familiar with this sign.	1	2	3	4	5
4	I easily forget the meaning of this sign.	1	2	3	4	5
5	I would learn this sign quickly.	1	2	3	4	5
6	This sign can be easily confused with another sign.	1	2	3	4	5
7	This sign can easily capture my attention.	1	2	3	4	5
8	This sign is complicated.	1	2	3	4	5
9	This sign is easy to understand.	1	2	3	4	5
10	I have seen many variations of this sign.	1	2	3	4	5
11	This sign is easy to read.	1	2	3	4	5
12	The colours used in this sign are not suitable.	1	2	3	4	5
13	The colour contrast of this sign is good.	1	2	3	4	5
14	The shape of this sign is not suitable.	1	2	3	4	5
15	I know what I should do when I see this sign.	1	2	3	4	5
16	It is likely that I will ignore this sign.	1	2	3	4	5
17	The icon in this sign conveys a clear meaning.	1	2	3	4	5
18	The icon design of this sign is not suitable.	1	2	3	4	5
19	The text in this sign conveys a clear meaning.	1	2	3	4	5
20	The text in this sign is difficult to read.	1	2	3	4	5



What is the meaning of this sign? Please explain your answer.

		Strongly Disagree		Neutral		Strongly Agree
1	I am confident that I understand the meaning of this sign.	1	2	3	4	5
2	I need help in understanding this sign for the first time.	1	2	3	4	5
3	I am familiar with this sign.	1	2	3	4	5
4	I easily forget the meaning of this sign.	1	2	3	4	5
5	I would learn this sign quickly.	1	2	3	4	5
6	This sign can be easily confused with another sign.	1	2	3	4	5
7	This sign can easily capture my attention.	1	2	3	4	5
8	This sign is complicated.	1	2	3	4	5
9	This sign is easy to understand.	1	2	3	4	5
10	I have seen many variations of this sign.	1	2	3	4	5
11	This sign is easy to read.	1	2	3	4	5
12	The colours used in this sign are not suitable	1	2	3	4	5
13	The colour contrast of this sign is good.	1	2	3	4	5
14	The shape of this sign is not suitable.	1	2	3	4	5
15	I know what I should do when I see this sign.	1	2	3	4	5
16	It is likely that I will ignore this sign.	1	2	3	4	5
17	The icon in this sign conveys a clear meaning.	1	2	3	4	5
18	The icon design of this sign is not suitable.	1	2	3	4	5



What is the meaning of this sign? Please explain your answer.

		Strongly Disagree		Neutral		Strongly Agree
1	I am confident that I understand the meaning of this sign.	1	2	3	4	5
2	I need help in understanding this sign for the first time.	1	2	3	4	5
3	I am familiar with this sign.	1	2	3	4	5
4	I easily forget the meaning of this sign.	1	2	3	4	5
5	I would learn this sign quickly.	1	2	3	4	5
6	This sign can be easily confused with another sign.	1	2	3	4	5
7	This sign can easily capture my attention.	1	2	3	4	5
8	This sign is complicated.	1	2	3	4	5
9	This sign is easy to understand.	1	2	3	4	5
10	I have seen many variations of this sign.	1	2	3	4	5
11	This sign is easy to read.	1	2	3	4	5
12	The colours used in this sign are not suitable.	1	2	3	4	5
13	The colour contrast of this sign is good.	1	2	3	4	5
14	The shape of this sign is not suitable.	1	2	3	4	5
15	I know what I should do when I see this sign.	1	2	3	4	5
16	It is likely that I will ignore this sign.	1	2	3	4	5
17	The text in this sign conveys a clear meaning.	1	2	3	4	5
18	The text in this sign is difficult to read.	1	2	3	4	5

Appendix D

Malay Translation of PSUS

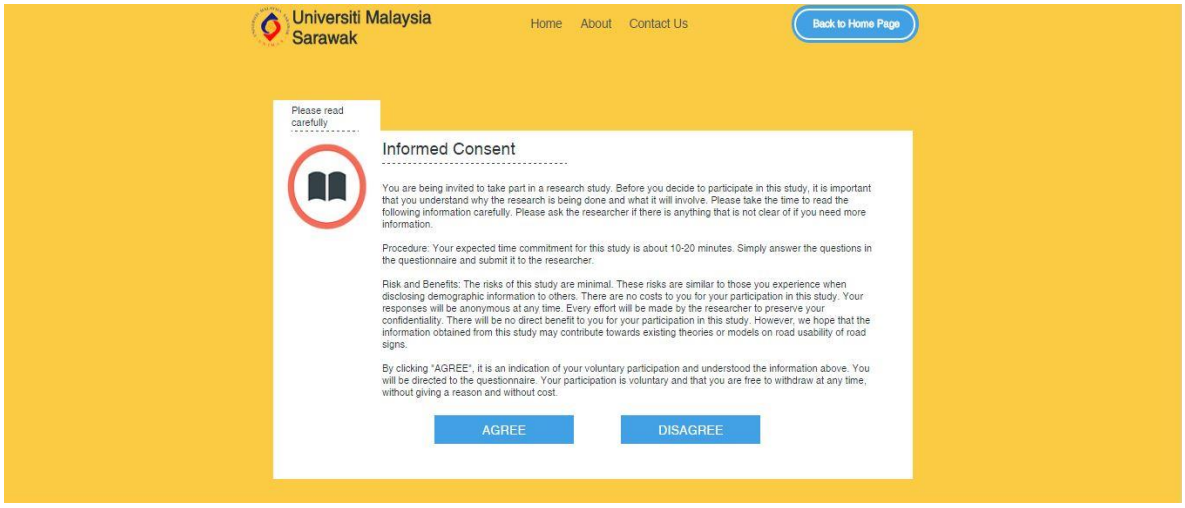
Item no.	Malay Translation
1	Saya yakin bahawa saya faham akan maksud papan tanda ini
2	Saya memerlukan bantuan untuk memahami maksud papan tanda ini buat kali pertama
3	Saya biasa dengan papan tanda ini
4	Saya mudah terlupa maksud papan tanda ini
5	Saya dapat memahami papan tanda ini dengan cepat
6	Papan tanda ini senang dikelirukan dengan papan tanda yang lain
7	Papan tanda ini mudah menarik perhatian saya
8	Papan tanda ini adalah rumit
9	Papan tanda ini mudah difahami
10	Saya pernah melihat banyak variasi papan tanda ini
11	Papan tanda ini mudah dibaca
12	Warna yang digunakan untuk papan tanda ini adalah tidak bersesuaian
13	Perbezaan warna bagi papan tanda ini adalah jelas
14	Bentuk papan tanda ini adalah tidak sesuai
15	Saya tahu apa yang perlu dilakukan apabila terlihat papan tanda ini.
16	Kemungkinan besar saya akan mengabaikan tanda ini
17	Ikon dalam papan tanda ini memberi makna yang jelas
18	Reka bentuk ikon papan tanda ini adalah tidak sesuai
19	Teks dalam papan tanda ini memberi makna yang jelas
20	Teks dalam papan tanda ini sukar untuk dibaca

Appendix E

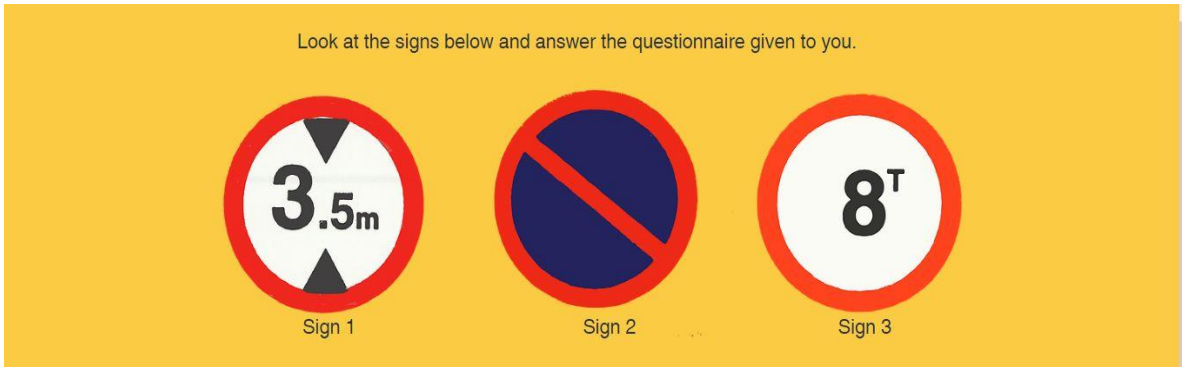
Website for Sign Display (http://donaldstephen89.wixsite.com/roadsigns)



Front page



Informed Consent Page



Appendix F

PSUS Constant Based on Scale Size and Number of Items

Item	Scale Size								
	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	100.0000	50.0000	33.3333	25.0000	20.0000	16.6667	14.2857	12.5000	11.1111
2	50.0000	25.0000	16.6667	12.5000	10.0000	8.3333	7.1429	6.2500	5.5556
3	33.3333	16.6667	11.1111	8.3333	6.6667	5.5556	4.7619	4.1667	3.7037
4	25.0000	12.5000	8.3333	6.2500	5.0000	4.1667	3.5714	3.1250	2.7778
5	20.0000	10.0000	6.6667	5.0000	4.0000	3.3333	2.8571	2.5000	2.2222
6	16.6667	8.3333	5.5556	4.1667	3.3333	2.7778	2.3810	2.0833	1.8519
7	14.2857	7.1429	4.7619	3.5714	2.8571	2.3810	2.0408	1.7857	1.5873
8	12.5000	6.2500	4.1667	3.1250	2.5000	2.0833	1.7857	1.5625	1.3889
9	11.1111	5.5556	3.7037	2.7778	2.2222	1.8519	1.5873	1.3889	1.2346
10	10.0000	5.0000	3.3333	2.5000	2.0000	1.6667	1.4286	1.2500	1.1111
11	9.0909	4.5455	3.0303	2.2727	1.8182	1.5152	1.2987	1.1364	1.0101
12	8.3333	4.1667	2.7778	2.0833	1.6667	1.3889	1.1905	1.0417	0.9259
13	7.6923	3.8462	2.5641	1.9231	1.5385	1.2821	1.0989	0.9615	0.8547
14	7.1429	3.5714	2.3810	1.7857	1.4286	1.1905	1.0204	0.8929	0.7937
15	6.6667	3.3333	2.2222	1.6667	1.3333	1.1111	0.9524	0.8333	0.7407
16	6.2500	3.1250	2.0833	1.5625	1.2500	1.0417	0.8929	0.7813	0.6944
17	5.8824	2.9412	1.9608	1.4706	1.1765	0.9804	0.8403	0.7353	0.6536
18	5.5556	2.7778	1.8519	1.3889	1.1111	0.9259	0.7937	0.6944	0.6173
19	5.2632	2.6316	1.7544	1.3158	1.0526	0.8772	0.7519	0.6579	0.5848
20	5.0000	2.5000	1.6667	1.2500	1.0000	0.8333	0.7143	0.6250	0.5556
21	4.7619	2.3810	1.5873	1.1905	0.9524	0.7937	0.6803	0.5952	0.5291
22	4.5455	2.2727	1.5152	1.1364	0.9091	0.7576	0.6494	0.5682	0.5051
23	4.3478	2.1739	1.4493	1.0870	0.8696	0.7246	0.6211	0.5435	0.4831
24	4.1667	2.0833	1.3889	1.0417	0.8333	0.6944	0.5952	0.5208	0.4630
25	4.0000	2.0000	1.3333	1.0000	0.8000	0.6667	0.5714	0.5000	0.4444
26	3.8462	1.9231	1.2821	0.9615	0.7692	0.6410	0.5495	0.4808	0.4274
27	3.7037	1.8519	1.2346	0.9259	0.7407	0.6173	0.5291	0.4630	0.4115
28	3.5714	1.7857	1.1905	0.8929	0.7143	0.5952	0.5102	0.4464	0.3968
29	3.4483	1.7241	1.1494	0.8621	0.6897	0.5747	0.4926	0.4310	0.3831
30	3.3333	1.6667	1.1111	0.8333	0.6667	0.5556	0.4762	0.4167	0.3704
31	3.2258	1.6129	1.0753	0.8065	0.6452	0.5376	0.4608	0.4032	0.3584
32	3.1250	1.5625	1.0417	0.7813	0.6250	0.5208	0.4464	0.3906	0.3472

33	3.0303	1.5152	1.0101	0.7576	0.6061	0.5051	0.4329	0.3788	0.3367
34	2.9412	1.4706	0.9804	0.7353	0.5882	0.4902	0.4202	0.3676	0.3268
35	2.8571	1.4286	0.9524	0.7143	0.5714	0.4762	0.4082	0.3571	0.3175
36	2.7778	1.3889	0.9259	0.6944	0.5556	0.4630	0.3968	0.3472	0.3086
37	2.7027	1.3514	0.9009	0.6757	0.5405	0.4505	0.3861	0.3378	0.3003
38	2.6316	1.3158	0.8772	0.6579	0.5263	0.4386	0.3759	0.3289	0.2924
39	2.5641	1.2821	0.8547	0.6410	0.5128	0.4274	0.3663	0.3205	0.2849
40	2.5000	1.2500	0.8333	0.6250	0.5000	0.4167	0.3571	0.3125	0.2778
41	2.4390	1.2195	0.8130	0.6098	0.4878	0.4065	0.3484	0.3049	0.2710
42	2.3810	1.1905	0.7937	0.5952	0.4762	0.3968	0.3401	0.2976	0.2646
43	2.3256	1.1628	0.7752	0.5814	0.4651	0.3876	0.3322	0.2907	0.2584
44	2.2727	1.1364	0.7576	0.5682	0.4545	0.3788	0.3247	0.2841	0.2525
45	2.2222	1.1111	0.7407	0.5556	0.4444	0.3704	0.3175	0.2778	0.2469
46	2.1739	1.0870	0.7246	0.5435	0.4348	0.3623	0.3106	0.2717	0.2415
47	2.1277	1.0638	0.7092	0.5319	0.4255	0.3546	0.3040	0.2660	0.2364
48	2.0833	1.0417	0.6944	0.5208	0.4167	0.3472	0.2976	0.2604	0.2315
49	2.0408	1.0204	0.6803	0.5102	0.4082	0.3401	0.2915	0.2551	0.2268
50	2.0000	1.0000	0.6667	0.5000	0.4000	0.3333	0.2857	0.2500	0.2222

Item	Scale Size									
	11	12	13	14	15	16	17	18	19	20
1	10.0000	9.0909	8.3333	7.6923	7.1429	6.6667	6.2500	5.8824	5.5556	5.2632
2	5.0000	4.5455	4.1667	3.8462	3.5714	3.3333	3.1250	2.9412	2.7778	2.6316
3	3.3333	3.0303	2.7778	2.5641	2.3810	2.2222	2.0833	1.9608	1.8519	1.7544
4	2.5000	2.2727	2.0833	1.9231	1.7857	1.6667	1.5625	1.4706	1.3889	1.3158
5	2.0000	1.8182	1.6667	1.5385	1.4286	1.3333	1.2500	1.1765	1.1111	1.0526
6	1.6667	1.5152	1.3889	1.2821	1.1905	1.1111	1.0417	0.9804	0.9259	0.8772
7	1.4286	1.2987	1.1905	1.0989	1.0204	0.9524	0.8929	0.8403	0.7937	0.7519
8	1.2500	1.1364	1.0417	0.9615	0.8929	0.8333	0.7813	0.7353	0.6944	0.6579
9	1.1111	1.0101	0.9259	0.8547	0.7937	0.7407	0.6944	0.6536	0.6173	0.5848
10	1.0000	0.9091	0.8333	0.7692	0.7143	0.6667	0.6250	0.5882	0.5556	0.5263
11	0.9091	0.8264	0.7576	0.6993	0.6494	0.6061	0.5682	0.5348	0.5051	0.4785
12	0.8333	0.7576	0.6944	0.6410	0.5952	0.5556	0.5208	0.4902	0.4630	0.4386
13	0.7692	0.6993	0.6410	0.5917	0.5495	0.5128	0.4808	0.4525	0.4274	0.4049
14	0.7143	0.6494	0.5952	0.5495	0.5102	0.4762	0.4464	0.4202	0.3968	0.3759
15	0.6667	0.6061	0.5556	0.5128	0.4762	0.4444	0.4167	0.3922	0.3704	0.3509
16	0.6250	0.5682	0.5208	0.4808	0.4464	0.4167	0.3906	0.3676	0.3472	0.3289
17	0.5882	0.5348	0.4902	0.4525	0.4202	0.3922	0.3676	0.3460	0.3268	0.3096
18	0.5556	0.5051	0.4630	0.4274	0.3968	0.3704	0.3472	0.3268	0.3086	0.2924
19	0.5263	0.4785	0.4386	0.4049	0.3759	0.3509	0.3289	0.3096	0.2924	0.2770
20	0.5000	0.4545	0.4167	0.3846	0.3571	0.3333	0.3125	0.2941	0.2778	1.1728
21	0.4762	0.4329	0.3968	0.3663	0.3401	0.3175	0.2976	0.2801	0.2646	0.2506
22	0.4545	0.4132	0.3788	0.3497	0.3247	0.3030	0.2841	0.2674	0.2525	0.2392
23	0.4348	0.3953	0.3623	0.3344	0.3106	0.2899	0.2717	0.2558	0.2415	0.2288
24	0.4167	0.3788	0.3472	0.3205	0.2976	0.2778	0.2604	0.2451	0.2315	0.2193
25	0.4000	0.3636	0.3333	0.3077	0.2857	0.2667	0.2500	0.2353	0.2222	0.2105
26	0.3846	0.3497	0.3205	0.2959	0.2747	0.2564	0.2404	0.2262	0.2137	0.2024
27	0.3704	0.3367	0.3086	0.2849	0.2646	0.2469	0.2315	0.2179	0.2058	0.1949
28	0.3571	0.3247	0.2976	0.2747	0.2551	0.2381	0.2232	0.2101	0.1984	0.1880
29	0.3448	0.3135	0.2874	0.2653	0.2463	0.2299	0.2155	0.2028	0.1916	0.1815
30	0.3333	0.3030	0.2778	0.2564	0.2381	0.2222	0.2083	0.1961	0.1852	0.1754
31	0.3226	0.2933	0.2688	0.2481	0.2304	0.2151	0.2016	0.1898	0.1792	0.1698
32	0.3125	0.2841	0.2604	0.2404	0.2232	0.2083	0.1953	0.1838	0.1736	0.1645
33	0.3030	0.2755	0.2525	0.2331	0.2165	0.2020	0.1894	0.1783	0.1684	0.1595
34	0.2941	0.2674	0.2451	0.2262	0.2101	0.1961	0.1838	0.1730	0.1634	0.1548

35	0.2857	0.2597	0.2381	0.2198	0.2041	0.1905	0.1786	0.1681	0.1587	0.1504
36	0.2778	0.2525	4.1667	0.2137	0.1984	0.1852	0.1736	0.1634	0.1543	0.1462
37	0.2703	0.2457	4.0541	0.2079	0.1931	0.1802	0.1689	0.1590	0.1502	0.1422
38	0.2632	0.2392	3.9474	0.2024	0.1880	0.1754	0.1645	0.1548	0.1462	0.1385
39	0.2564	0.2331	0.2137	0.1972	0.1832	0.1709	0.1603	0.1508	0.1425	0.1350
40	0.2500	0.2273	0.2083	0.1923	0.1786	0.1667	0.1563	0.1471	0.1389	0.1316
41	0.2439	0.2217	0.2033	0.1876	0.1742	0.1626	0.1524	0.1435	0.1355	0.1284
42	0.2381	0.2165	0.1984	0.1832	0.1701	0.1587	0.1488	0.1401	0.1323	0.1253
43	0.2326	0.2114	0.1938	0.1789	0.1661	0.1550	0.1453	0.1368	0.1292	0.1224
44	0.2273	0.2066	0.1894	0.1748	0.1623	0.1515	0.1420	0.1337	0.1263	0.1196
45	0.2222	0.2020	0.1852	0.1709	0.1587	0.1481	0.1389	0.1307	0.1235	0.1170
46	0.2174	0.1976	0.1812	0.1672	0.1553	0.1449	0.1359	0.1279	0.1208	0.1144
47	0.2128	0.1934	0.1773	0.1637	0.1520	0.1418	0.1330	0.1252	0.1182	0.1120
48	0.2083	0.1894	0.1736	0.1603	0.1488	0.1389	0.1302	0.1225	0.1157	0.1096
49	0.2041	0.1855	0.1701	0.1570	0.1458	0.1361	0.1276	0.1200	0.1134	0.1074
50	0.2000	0.1818	0.1667	0.1538	0.1429	0.1333	0.1250	0.1176	0.1111	0.1053

Appendix G

Option Item Analysis of 3 Test Signs

Item	Option/ Score	Diff.	SD	Discri.	Item	Option	Diff.	SD	Discri.
a1	0.0(0.0)	0.0354	0.1851	-0.3491	a6	0.0(4.0)	0.1061	0.3087	0.4502
	1.0(1.0)	0.1717	0.3781	-0.4242		1.0(3.0)	0.2020	0.4025	0.2217
	2.0(2.0)	0.2677	0.4439	-0.1600		2.0(2.0)	0.3434	0.4761	-0.1520
	3.0(3.0)	0.3131	0.4649	0.1899		3.0(1.0)	0.2323	0.4234	-0.2922
	4.0(4.0)	0.2121	0.4098	0.3530		4.0(0.0)	0.1162	0.3212	-0.3007
a2	0.0(4.0)	0.0606	0.2392	0.3388	a7	0.0(0.0)	0.0556	0.2296	-0.2001
	1.0(3.0)	0.1717	0.3781	0.3360		1.0(1.0)	0.1768	0.3824	-0.1022
	2.0(2.0)	0.2273	0.4201	-0.0384		2.0(2.0)	0.2879	0.4539	-0.1687
	3.0(1.0)	0.3535	0.4793	-0.2266		3.0(3.0)	0.3687	0.4837	0.0178
	4.0(0.0)	0.1869	0.3908	-0.3768		4.0(4.0)	0.1111	0.3151	0.2776
a3	0.0(0.0)	0.0152	0.1225	-0.1908	a8	0.0(4.0)	0.1364	0.3440	0.3762
	1.0(1.0)	0.1364	0.3440	-0.3589		1.0(3.0)	0.3283	0.4708	0.1489
	2.0(2.0)	0.2677	0.4439	-0.2378		2.0(2.0)	0.2929	0.4563	-0.1159
	3.0(3.0)	0.3384	0.4744	0.0899		3.0(1.0)	0.1869	0.3908	-0.3711
	4.0(4.0)	0.2424	0.4296	0.3424		4.0(0.0)	0.0556	0.2296	-0.2856
a4	0.0(4.0)	0.1465	0.3545	0.4123	a9	0.0(0.0)	0.0253	0.1573	-0.2335
	1.0(3.0)	0.2273	0.4201	0.2528		1.0(1.0)	0.1111	0.3151	-0.3549
	2.0(2.0)	0.2727	0.4465	-0.1989		2.0(2.0)	0.3030	0.4607	-0.3155
	3.0(1.0)	0.2172	0.4134	-0.2386		3.0(3.0)	0.3636	0.4823	0.1965
	4.0(0.0)	0.1263	0.3330	-0.3861		4.0(4.0)	0.1818	0.3867	0.4002
a5	0.0(0.0)	0.0152	0.1225	-0.1532	a10	0.0(4.0)	0.0808	0.2732	0.0754
	1.0(1.0)	0.0808	0.2732	-0.1688		1.0(3.0)	0.2020	0.4025	0.0630
	2.0(2.0)	0.3081	0.4629	-0.3673		2.0(2.0)	0.2828	0.4515	-0.0103
	3.0(3.0)	0.4091	0.4929	0.1404		3.0(1.0)	0.3182	0.4670	-0.2417
	4.0(4.0)	0.1869	0.3908	0.2724		4.0(0.0)	0.1111	0.3151	0.0133

Item	Option	Diff.	SD	Discri.	Item	Option	Diff.	SD	Discri.
a11	0.0(0.0)	0.0253	0.1573	-0.2975	a16	0.0(4.0)	0.0354	0.1851	0.2501
	1.0(1.0)	0.1061	0.3087	-0.2897		1.0(3.0)	0.1313	0.3386	0.3312
	2.0(2.0)	0.2273	0.4201	-0.2771		2.0(2.0)	0.1364	0.3440	0.0649
	3.0(3.0)	0.4646	0.5000	0.1365		3.0(1.0)	0.1667	0.3736	0.0006
	4.0(4.0)	0.1717	0.3781	0.3435		4.0(0.0)	0.1010	0.3021	-0.1552
a12	0.0(4.0)	0.1717	0.3781	0.2874	a17	0.0(0.0)	0.0303	0.1719	-0.2551
	1.0(3.0)	0.3939	0.4899	0.1074		1.0(1.0)	0.1768	0.3824	-0.3020
	2.0(2.0)	0.2828	0.4515	-0.2820		2.0(2.0)	0.3131	0.4649	-0.1697
	3.0(1.0)	0.1010	0.3021	-0.2277		3.0(3.0)	0.3182	0.4670	0.1201
	4.0(0.0)	0.0505	0.2195	-0.1269		4.0(4.0)	0.1414	0.3493	0.3890
a13	0.0(0.0)	0.0051	0.0711	-0.0169	a18	0.0(4.0)	0.0808	0.2732	0.3587

	1.0(1.0)	0.0657	0.2483	-0.1297		1.0(3.0)	0.3586	0.4808	0.2106
	2.0(2.0)	0.2071	0.4062	-0.3491		2.0(2.0)	0.3889	0.4887	-0.3182
	3.0(3.0)	0.4949	0.5012	0.0268		3.0(1.0)	0.1364	0.3440	-0.2034
	4.0(4.0)	0.2222	0.4168	0.2640		4.0(0.0)	0.0354	0.1851	-0.1763
a14	0.0(4.0)	0.1263	0.3330	0.3443	a19	0.0(0.0)	0.0303	0.1719	-0.1645
	1.0(3.0)	0.3737	0.4850	0.1946		1.0(1.0)	0.1566	0.3643	-0.2809
	2.0(2.0)	0.3131	0.4649	-0.2884		2.0(2.0)	0.2929	0.4563	-0.3264
	3.0(1.0)	0.1313	0.3386	-0.2291		3.0(3.0)	0.3889	0.4887	0.2032
	4.0(0.0)	0.0556	0.2296	-0.2500		4.0(4.0)	0.1313	0.3386	0.3535
a15	0.0(0.0)	0.0404	0.1974	-0.2547	a20	0.0(4.0)	0.1515	0.3595	0.3077
	1.0(1.0)	0.0960	0.2953	-0.2046		1.0(3.0)	0.3838	0.4876	0.1277
	2.0(2.0)	0.1162	0.3212	-0.1092		2.0(2.0)	0.2778	0.4490	-0.1998
	3.0(3.0)	0.2121	0.4098	0.3495		3.0(1.0)	0.1212	0.3272	-0.2736
	4.0(4.0)	0.1061	0.3087	0.4532		4.0(0.0)	0.0657	0.2483	-0.2310

Item	Option /Score	Diff.	SD	Discri.	Item	Option /Score	Diff.	SD	Discri.
b1	0.0(0.0)	0.0936	0.2920	-0.4239	b6	0.0(4.0)	0.0443	0.2063	0.1697
	1.0(1.0)	0.1626	0.3699	-0.3397		1.0(3.0)	0.1773	0.3829	0.3946
	2.0(2.0)	0.2611	0.4403	-0.1929		2.0(2.0)	0.1970	0.3987	0.0565
	3.0(3.0)	0.3153	0.4658	0.2338		3.0(1.0)	0.3744	0.4852	-0.2571
	4.0(4.0)	0.1675	0.3743	0.4180		4.0(0.0)	0.2069	0.4061	-0.3737
b2	0.0(4.0)	0.0640	0.2454	0.1502	b7	0.0(0.0)	0.0443	0.2063	-0.2141
	1.0(3.0)	0.1576	0.3653	0.3893		1.0(1.0)	0.1823	0.3870	-0.3559
	2.0(2.0)	0.2266	0.4197	0.0367		2.0(2.0)	0.2709	0.4455	-0.1650
	3.0(1.0)	0.3695	0.4839	-0.2470		3.0(3.0)	0.3645	0.4825	0.1751
	4.0(0.0)	0.1823	0.3870	-0.3719		4.0(4.0)	0.1379	0.3457	0.2946
b3	0.0(0.0)	0.0542	0.2269	-0.2994	b8	0.0(4.0)	0.0985	0.2988	0.1406
	1.0(1.0)	0.1232	0.3294	-0.3798		1.0(3.0)	0.3645	0.4825	0.3499
	2.0(2.0)	0.2315	0.4229	-0.2394		2.0(2.0)	0.3153	0.4658	-0.1721
	3.0(3.0)	0.3596	0.4811	0.1239		3.0(1.0)	0.1379	0.3457	-0.2750
	4.0(4.0)	0.2315	0.4229	0.3945		4.0(0.0)	0.0837	0.2777	-0.3699
b4	0.0(4.0)	0.0690	0.2540	0.1382	b9	0.0(0.0)	0.0788	0.2701	-0.4378
	1.0(3.0)	0.2562	0.4376	0.4114		1.0(1.0)	0.1084	0.3116	-0.2995
	2.0(2.0)	0.3005	0.4596	-0.1259		2.0(2.0)	0.3054	0.4617	-0.1902
	3.0(1.0)	0.2315	0.4229	-0.2519		3.0(3.0)	0.3941	0.4899	0.2096
	4.0(0.0)	0.1429	0.3508	-0.3439		4.0(4.0)	0.1133	0.3177	0.4121
b5	0.0(0.0)	0.0296	0.1698	-0.2893	b10	0.0(4.0)	0.0640	0.2454	-0.1132
	1.0(1.0)	0.1232	0.3294	-0.3115		1.0(3.0)	0.1872	0.3910	0.2063
	2.0(2.0)	0.2463	0.4319	-0.3026		2.0(2.0)	0.3153	0.4658	-0.1969
	3.0(3.0)	0.4335	0.4968	0.2331		3.0(1.0)	0.3005	0.4596	-0.0466
	4.0(4.0)	0.1675	0.3743	0.2718		4.0(0.0)	0.1330	0.3404	-0.0382

Item	Option /Score	Diff.	SD	Discri.	Item	Option /Score	Diff.	SD	Discri.
b11	0.0(0.0)	0.0493	0.2169	-0.3631	b15	0.0(0.0)	0.0542	0.2269	-0.3335
	1.0(1.0)	0.1379	0.3457	-0.2736		1.0(1.0)	0.0887	0.2850	-0.2397
	2.0(2.0)	0.2906	0.4552	-0.3345		2.0(2.0)	0.1281	0.3350	-0.0557
	3.0(3.0)	0.3842	0.4876	0.2114		3.0(3.0)	0.2020	0.4025	0.2377
	4.0(4.0)	0.1379	0.3457	0.4558		4.0(4.0)	0.0985	0.2988	0.4894
b12	0.0(4.0)	0.0690	0.2540	0.1130	b16	0.0(4.0)	0.0542	0.2269	0.2977
	1.0(3.0)	0.3892	0.4888	0.3541		1.0(3.0)	0.2020	0.4025	0.3799
	2.0(2.0)	0.3103	0.4638	-0.3030		2.0(2.0)	0.1281	0.3350	-0.0047
	3.0(1.0)	0.1675	0.3743	-0.1938		3.0(1.0)	0.1330	0.3404	-0.1909
	4.0(0.0)	0.0640	0.2454	-0.2200		4.0(0.0)	0.0542	0.2269	-0.3292
b13	0.0(0.0)	0.0345	0.1829	-0.1826	b17	0.0(0.0)	0.0640	0.2454	-0.2791
	1.0(1.0)	0.1084	0.3116	-0.1819		1.0(1.0)	0.1823	0.3870	-0.2794
	2.0(2.0)	0.3202	0.4677	-0.2556		2.0(2.0)	0.3448	0.4765	-0.3040
	3.0(3.0)	0.3842	0.4876	0.1358		3.0(3.0)	0.2906	0.4552	0.2748
	4.0(4.0)	0.1527	0.3606	0.2090		4.0(4.0)	0.1182	0.3237	0.3992
b14	0.0(4.0)	0.0542	0.2269	0.0954	b18	0.0(4.0)	0.0739	0.2622	0.0560
	1.0(3.0)	0.3892	0.4888	0.3045		1.0(3.0)	0.3744	0.4852	0.3596
	2.0(2.0)	0.3842	0.4876	-0.2742		2.0(2.0)	0.3892	0.4888	-0.3367
	3.0(1.0)	0.1281	0.3350	-0.2136		3.0(1.0)	0.1429	0.3508	-0.2317
	4.0(0.0)	0.0443	0.2063	-0.1364		4.0(0.0)	0.0197	0.1393	-0.0338

Item	Option /Score	Diff.	SD	Discri.	Item	Option /Score	Diff.	SD	Discri.
c1	0.0(0.0)	0.2135	0.4109	-0.4361	c6	0.0(4.0)	0.0885	0.2848	0.3041
	1.0(1.0)	0.1875	0.3913	-0.2767		1.0(3.0)	0.2656	0.4428	0.2559
	2.0(2.0)	0.1667	0.3737	-0.1351		2.0(2.0)	0.3438	0.4762	-0.2282
	3.0(3.0)	0.2760	0.4482	0.1889		3.0(1.0)	0.2240	0.4180	-0.2782
	4.0(4.0)	0.1563	0.3640	0.4920		4.0(0.0)	0.0781	0.2691	-0.1804
c2	0.0(4.0)	0.0573	0.2330	0.3245	c7	0.0(0.0)	0.0417	0.2003	-0.1882
	1.0(3.0)	0.1458	0.3539	0.3270		1.0(1.0)	0.1823	0.3871	-0.3152
	2.0(2.0)	0.2135	0.4109	0.0582		2.0(2.0)	0.3594	0.4811	-0.1790
	3.0(1.0)	0.3542	0.4795	-0.2567		3.0(3.0)	0.3542	0.4795	0.2491
	4.0(0.0)	0.2292	0.4214	-0.3911		4.0(4.0)	0.0573	0.2330	0.2673
c3	0.0(0.0)	0.0885	0.2848	-0.2616	c8	0.0(4.0)	0.1094	0.3129	0.1810
	1.0(1.0)	0.1875	0.3913	-0.3435		1.0(3.0)	0.3229	0.4688	0.2980
	2.0(2.0)	0.2292	0.4214	-0.1487		2.0(2.0)	0.2865	0.4533	-0.1214
	3.0(3.0)	0.3646	0.4826	0.1838		3.0(1.0)	0.2083	0.4072	-0.3487
	4.0(4.0)	0.1302	0.3374	0.3217		4.0(0.0)	0.0729	0.2607	-0.2804
c4	0.0(4.0)	0.0937	0.2922	0.4006	c9	0.0(0.0)	0.0365	0.1879	-0.2681
	1.0(3.0)	0.2083	0.4072	0.2252		1.0(1.0)	0.2292	0.4214	-0.4549
	2.0(2.0)	0.3594	0.4811	-0.1387		2.0(2.0)	0.2396	0.4279	-0.1065

	3.0(1.0)	0.2396	0.4279	-0.2551		3.0(3.0)	0.3750	0.4854	0.3638
	4.0(0.0)	0.0990	0.2994	-0.3565		4.0(4.0)	0.1198	0.3256	0.1282
c5	0.0(0.0)	0.0573	0.2330	-0.0982	c10	0.0(4.0)	0.0833	0.2771	-0.1076
	1.0(1.0)	0.1146	0.3194	-0.2441		1.0(3.0)	0.2031	0.4034	-0.0256
	2.0(2.0)	0.2917	0.4557	-0.3477		2.0(2.0)	0.4167	0.4943	-0.1022
	3.0(3.0)	0.4115	0.4934	0.1829		3.0(1.0)	0.2552	0.4371	0.0010
	4.0(4.0)	0.1198	0.3256	0.3212		4.0(0.0)	0.0417	0.2003	0.0789

Item	Option /Score	Diff.	SD	Discri.	Item	Option /Score	Diff.	SD	Discri.
c11	0.0(0.0)	0.0625	0.2427	-0.3409	c15	0.0(0.0)	0.1094	0.3129	-0.3444
	1.0(1.0)	0.1823	0.3871	-0.3578		1.0(1.0)	0.0885	0.2848	-0.1857
	2.0(2.0)	0.2188	0.4145	-0.1987		2.0(2.0)	0.1458	0.3539	0.0385
	3.0(3.0)	0.4115	0.4934	0.2533		3.0(3.0)	0.1771	0.3827	0.1781
	4.0(4.0)	0.1146	0.3194	0.3405		4.0(4.0)	0.0833	0.2771	0.4692
c12	0.0(4.0)	0.1458	0.3539	0.4257	c16	0.0(4.0)	0.0625	0.2427	0.2416
	1.0(3.0)	0.3385	0.4745	0.0200		1.0(3.0)	0.1615	0.3689	0.1473
	2.0(2.0)	0.2969	0.4581	-0.2822		2.0(2.0)	0.1667	0.3737	0.1064
	3.0(1.0)	0.1615	0.3689	-0.2547		3.0(1.0)	0.1667	0.3737	-0.2275
	4.0(0.0)	0.0521	0.2228	-0.0303		4.0(0.0)	0.0833	0.2771	0.0402
c13	0.0(0.0)	0.0104	0.1018	-0.1575	c18	0.0(0.0)	0.0677	0.2519	-0.2011
	1.0(1.0)	0.0573	0.2330	-0.1613		1.0(1.0)	0.1563	0.3640	-0.3925
	2.0(2.0)	0.3021	0.4604	-0.1739		2.0(2.0)	0.3438	0.4762	-0.2339
	3.0(3.0)	0.4479	0.4986	-0.1229		3.0(3.0)	0.3437	0.4762	0.2732
	4.0(4.0)	0.1823	0.3871	0.3287		4.0(4.0)	0.0885	0.2848	0.3688
c14	0.0(4.0)	0.1250	0.3316	0.3856	c18	0.0(4.0)	2.3229	1.0877	0.3704
	1.0(3.0)	0.3333	0.4726	0.0457		1.0(3.0)	0.1198	0.3256	0.3298
	2.0(2.0)	0.4115	0.4934	-0.2215		2.0(2.0)	0.3750	0.4854	0.1141
	3.0(1.0)	0.0937	0.2922	-0.2374		3.0(1.0)	0.2812	0.4508	-0.1687
	4.0(0.0)	0.0365	0.1879	-0.2133		4.0(0.0)	0.1563	0.3640	-0.2811

Appendix H

Item Deletion Reliability Statistics

RP.06

Item	If Item dropped					
	λ2	Alpha	F-G	F-B	Raju	λ6
a1	0.842	0.831	0.837	0.836	0.831	0.904
a2	0.844	0.833	0.840	0.838	0.833	0.905
a3	0.847	0.836	0.842	0.841	0.836	0.904
a4	0.844	0.832	0.839	0.837	0.832	0.902
a5	0.850	0.840	0.845	0.844	0.840	0.901
a6	0.845	0.833	0.840	0.838	0.833	0.908
a7	0.855	0.846	0.851	0.850	0.846	0.908
a8	0.846	0.834	0.841	0.839	0.834	0.905
a9	0.843	0.831	0.838	0.836	0.831	0.905
a10	0.862	0.854	0.858	0.857	0.854	0.917
a11	0.845	0.834	0.841	0.840	0.834	0.904
a12	0.852	0.842	0.848	0.847	0.842	0.909
a13	0.852	0.842	0.848	0.847	0.842	0.912
a14	0.848	0.836	0.844	0.842	0.836	0.906
a15	0.842	0.831	0.837	0.836	0.831	0.897
a16	0.847	0.837	0.844	0.842	0.837	0.910
a17	0.847	0.835	0.842	0.841	0.835	0.904
a18	0.849	0.838	0.844	0.843	0.838	0.906
a19	0.848	0.836	0.843	0.842	0.836	0.905
a20	0.850	0.839	0.846	0.844	0.839	0.908
Coefficient	0.854	0.844	0.849	0.848	0.844	0.913

RP.10

Item	If Item dropped					
	λ2	Alpha	F-G	F-B	Raju	λ6
b1	0.825	0.809	0.817	0.815	0.809	0.898
b2	0.834	0.819	0.827	0.826	0.819	0.906
b3	0.829	0.814	0.822	0.820	0.814	0.901
b4	0.836	0.820	0.829	0.827	0.820	0.908
b5	0.833	0.817	0.826	0.824	0.817	0.903
b6	0.834	0.818	0.827	0.825	0.818	0.906
b7	0.835	0.820	0.829	0.827	0.820	0.904
b8	0.835	0.819	0.828	0.826	0.819	0.908
b9	0.826	0.810	0.819	0.817	0.810	0.900
b10	0.854	0.844	0.849	0.848	0.844	0.917
b11	0.827	0.811	0.820	0.818	0.811	0.899

Continue

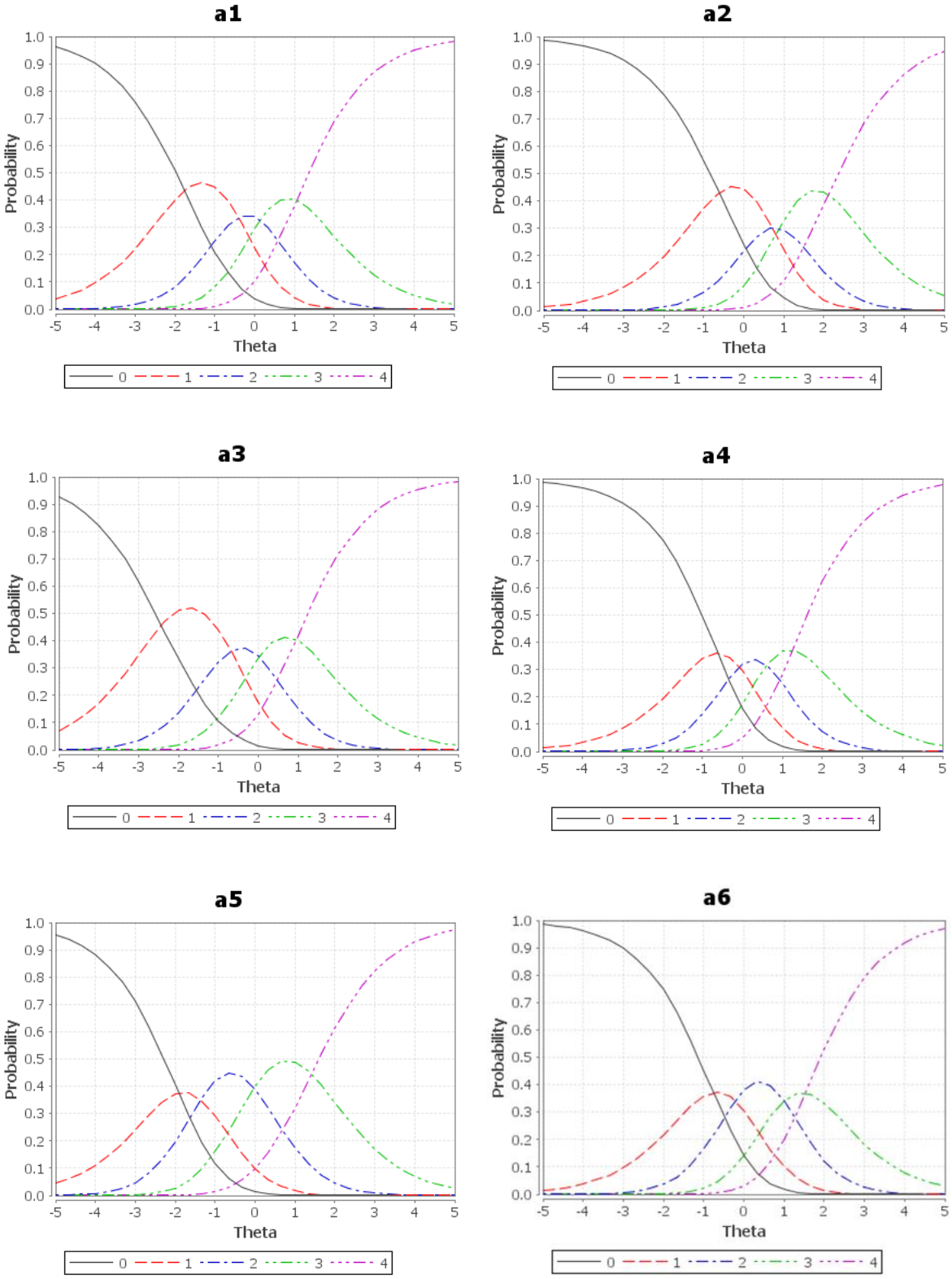
b12	0.840	0.825	0.834	0.832	0.825	0.907
b13	0.841	0.826	0.834	0.833	0.826	0.911
b14	0.841	0.826	0.835	0.833	0.826	0.909
b15	0.831	0.817	0.826	0.823	0.817	0.899
b16	0.832	0.819	0.828	0.825	0.819	0.904
b17	0.831	0.815	0.824	0.822	0.815	0.901
b18	0.841	0.826	0.835	0.833	0.826	0.910
Coefficient	0.843	0.828	0.836	0.835	0.828	0.912

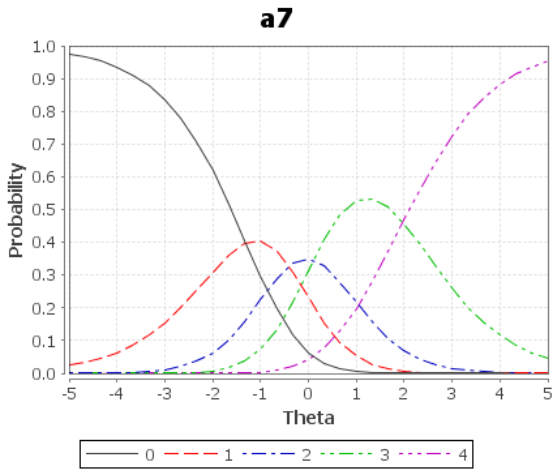
RP.05

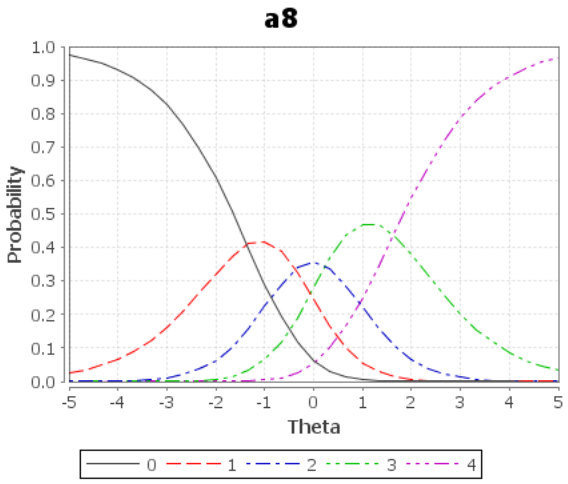
Item	If Item dropped					
	λ2	Alpha	F-G	F-B	Raju	λ6
c1	0.792	0.775	0.784	0.782	0.775	0.823
c2	0.799	0.780	0.790	0.787	0.780	0.828
c3	0.803	0.786	0.795	0.793	0.786	0.832
c4	0.800	0.781	0.792	0.789	0.781	0.828
c5	0.808	0.791	0.800	0.798	0.791	0.855
c6	0.807	0.789	0.800	0.797	0.789	0.839
c7	0.805	0.788	0.798	0.795	0.788	0.830
c8	0.806	0.787	0.798	0.795	0.787	0.837
c9	0.803	0.784	0.794	0.792	0.784	0.842
c10	0.830	0.819	0.824	0.823	0.819	0.864
c11	0.799	0.780	0.790	0.788	0.780	0.828
c12	0.811	0.793	0.804	0.802	0.793	0.838
c13	0.812	0.794	0.804	0.802	0.794	0.838
c14	0.806	0.788	0.800	0.797	0.788	0.838
c15	0.803	0.784	0.795	0.792	0.784	0.824
c16	0.817	0.799	0.809	0.807	0.799	0.850
c17	0.799	0.781	0.792	0.789	0.781	0.822
c18	0.808	0.789	0.800	0.798	0.789	0.840
Coefficient	0.815	0.798	0.807	0.805	0.798	0.847

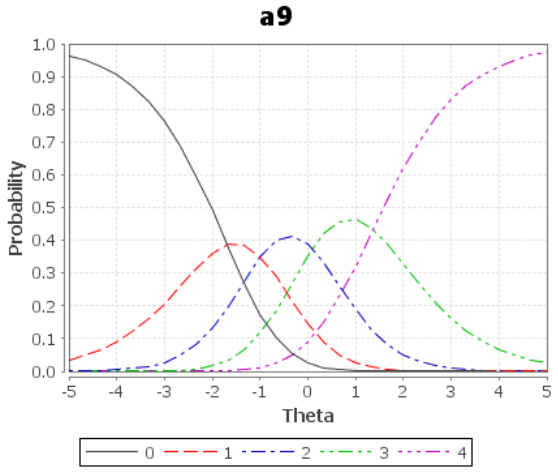
Appendix I

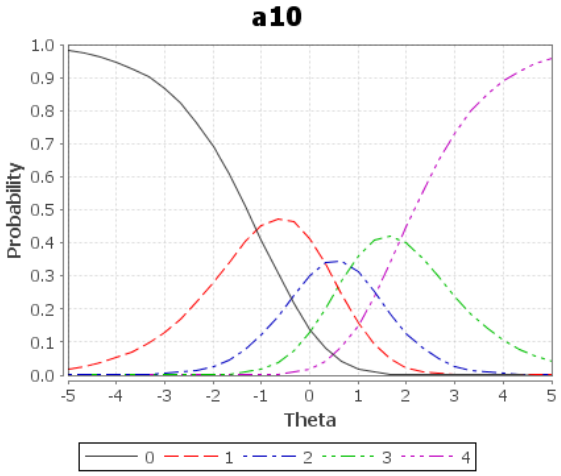
Option Characteristics Curves for Sign A RP.06

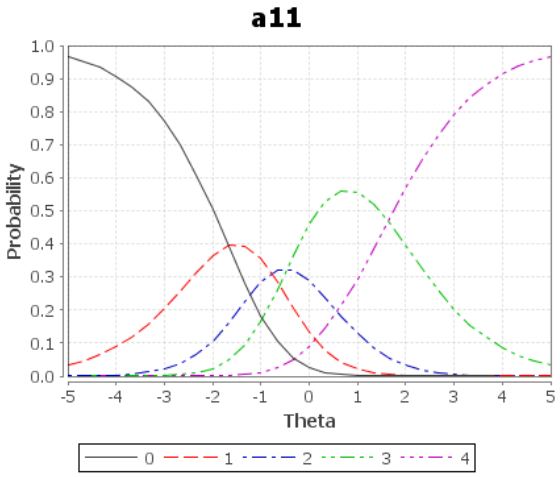


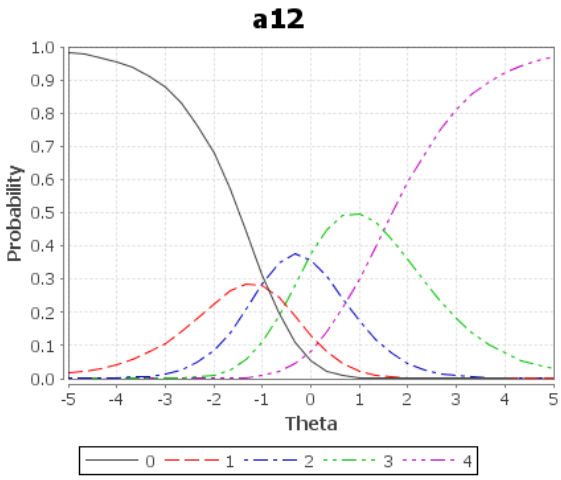

a7
Probability
Theta
0 1 2 3 4

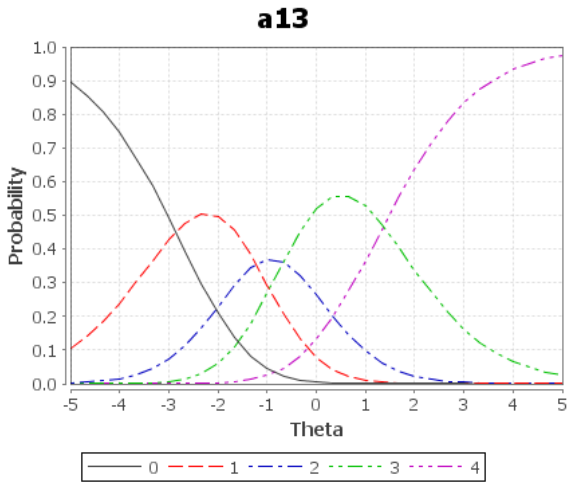

a8
Probability
Theta
0 1 2 3 4

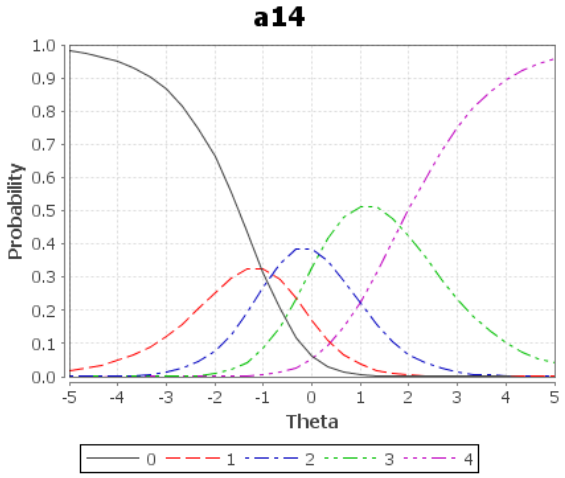

a9
Probability
Theta
0 1 2 3 4

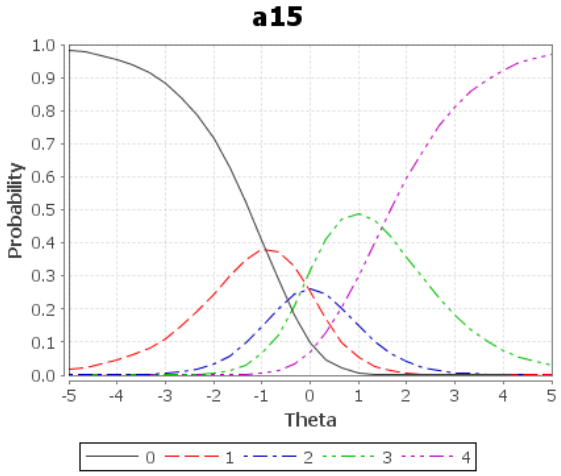

a10
Probability
Theta
0 1 2 3 4

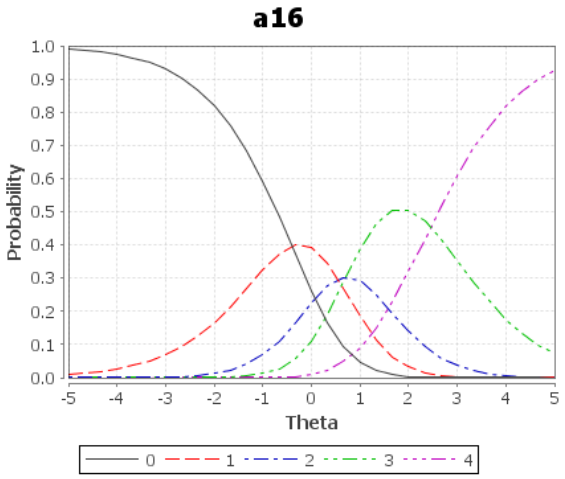

a11
Probability
Theta
0 1 2 3 4

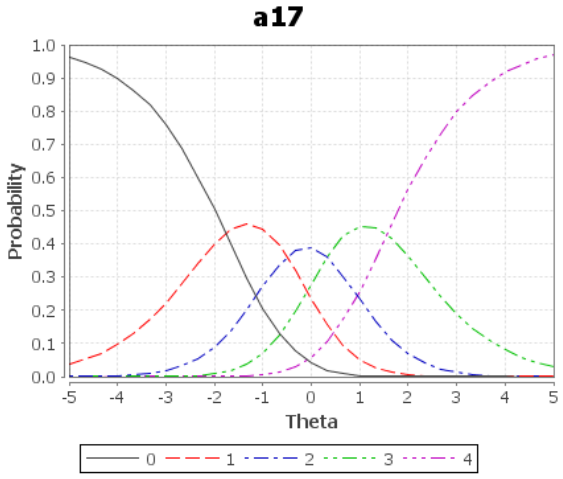

a12
Probability
Theta
0 1 2 3 4

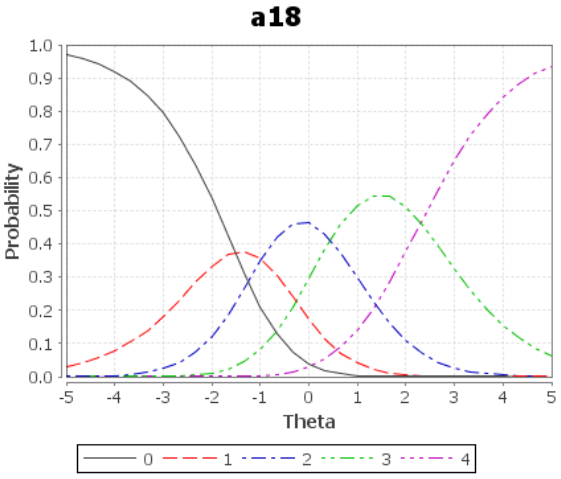

a13
Probability
Theta
0 1 2 3 4

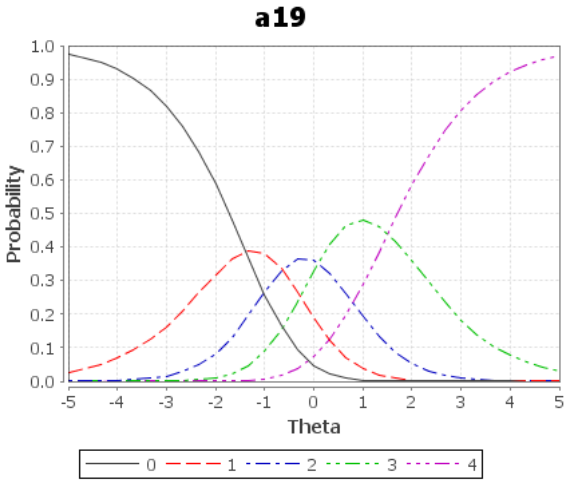

a14
Probability
Theta
0 1 2 3 4

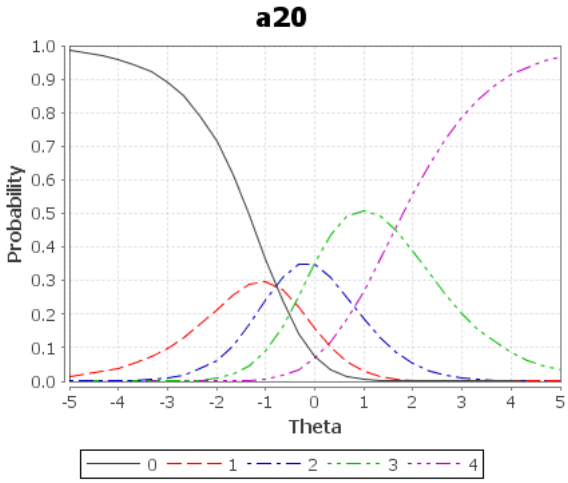

a15
Probability
Theta
0 1 2 3 4


a16
Probability
Theta
0 1 2 3 4


a17
Probability
Theta
0 1 2 3 4

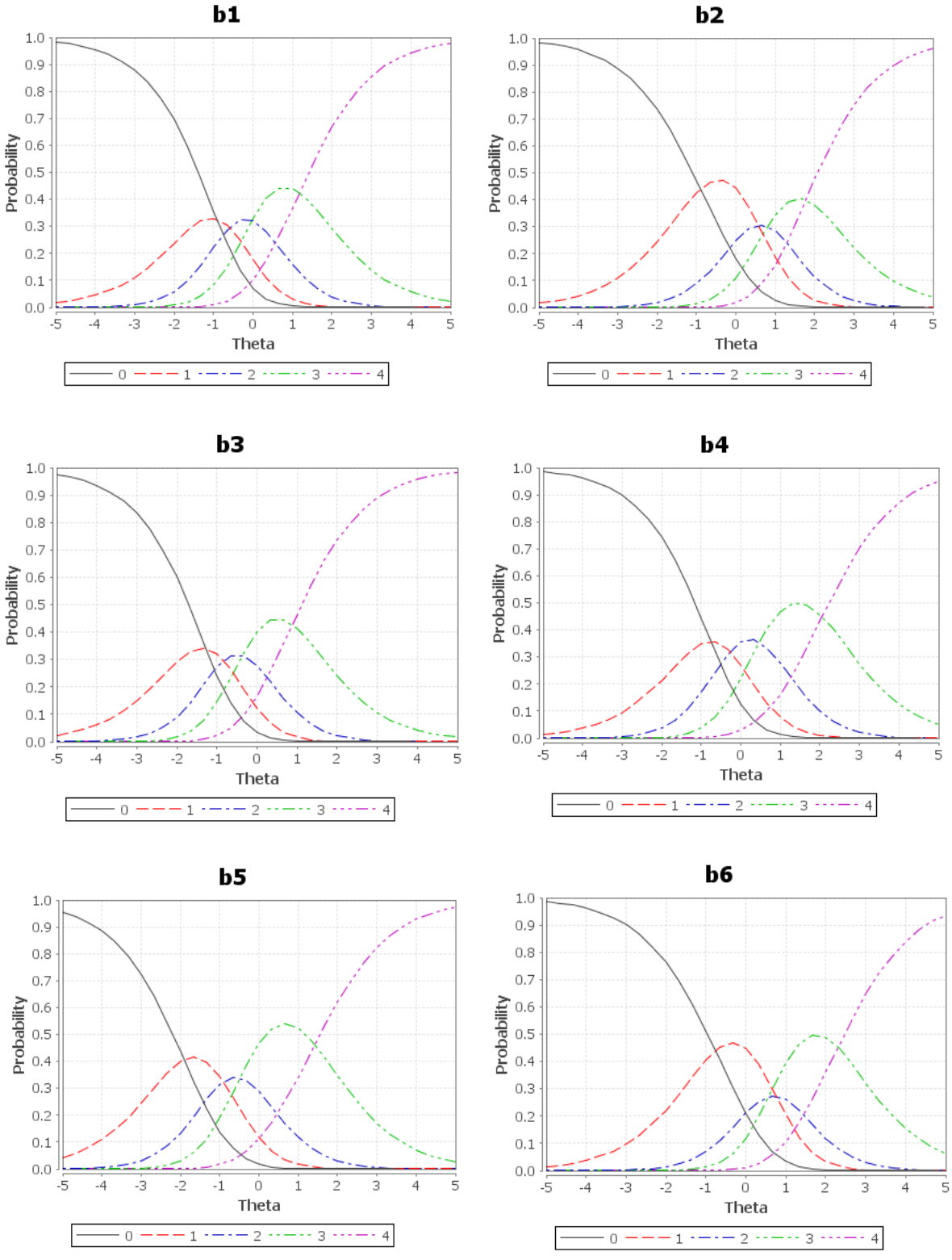

a18
Probability
Theta
0 1 2 3 4

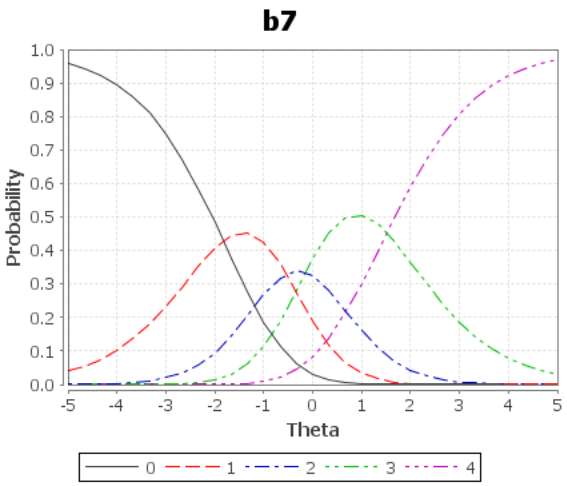

a19
Probability
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-5 -4 -3 -2 -1 0 1 2 3 4 5
Theta
0 1 2 3 4

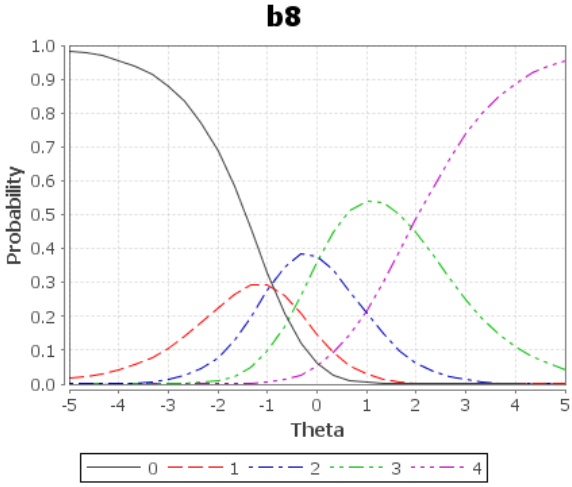

a20
Probability
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-5 -4 -3 -2 -1 0 1 2 3 4 5
Theta
0 1 2 3 4

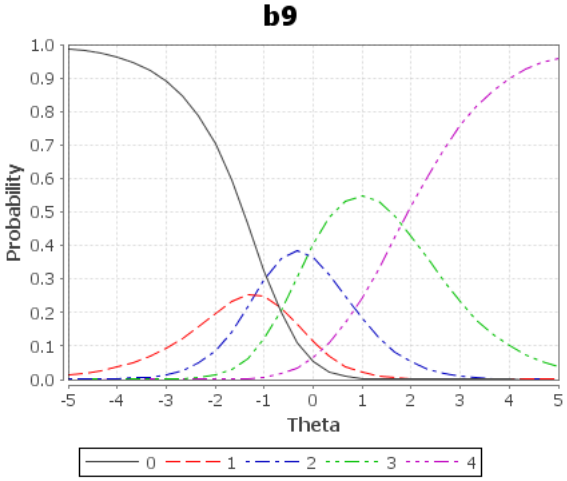
Appendix I

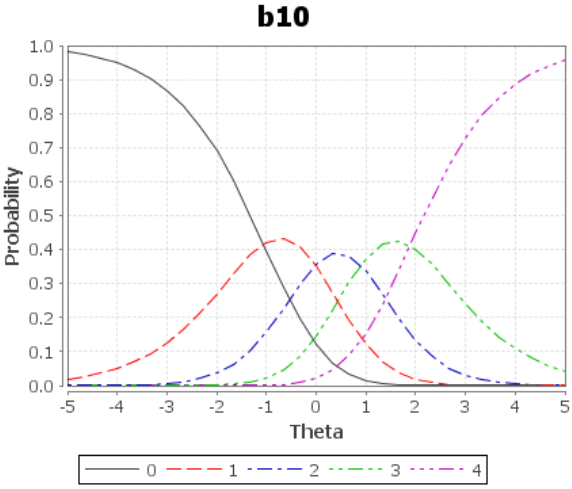
Option Characteristics Curves for Sign B RP.10

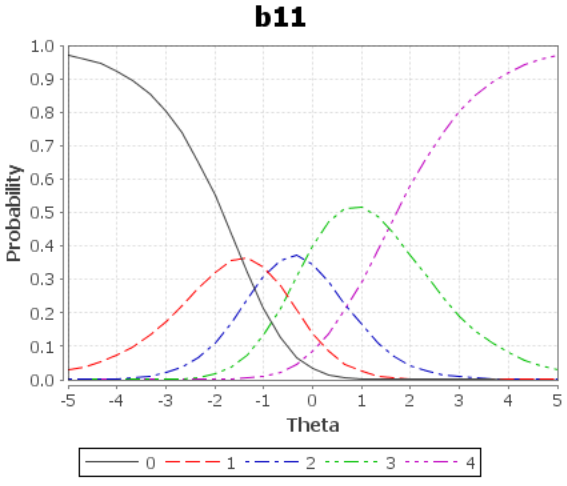


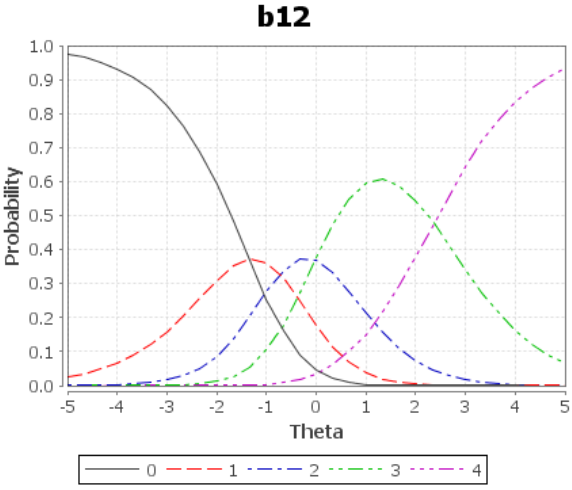

b7
Probability
Theta
0 1 2 3 4


b8
Probability
Theta
0 1 2 3 4

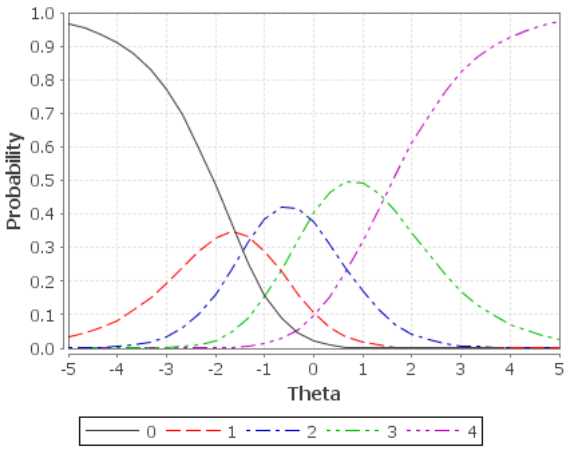

b9
Probability
Theta
0 1 2 3 4


b10
Probability
Theta
0 1 2 3 4

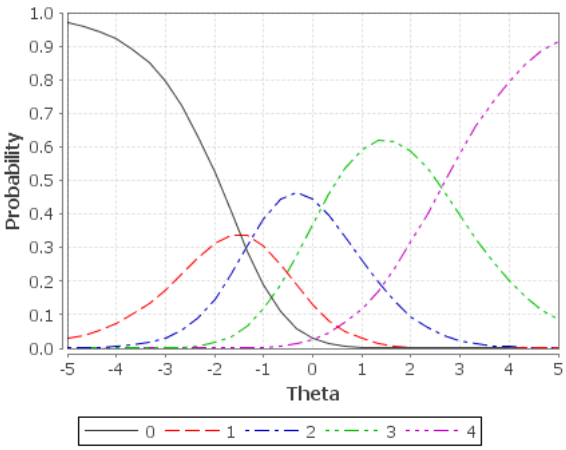

b11
Probability
Theta
0 1 2 3 4


b12
Probability
Theta
0 1 2 3 4

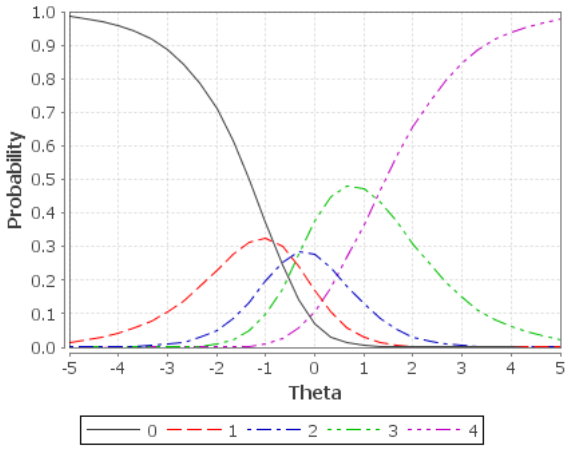
b13



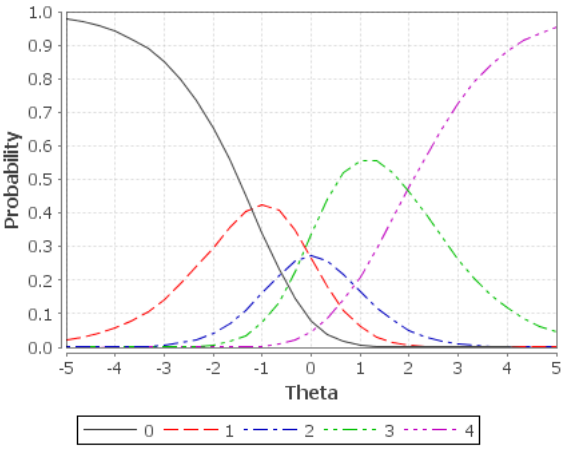
b14



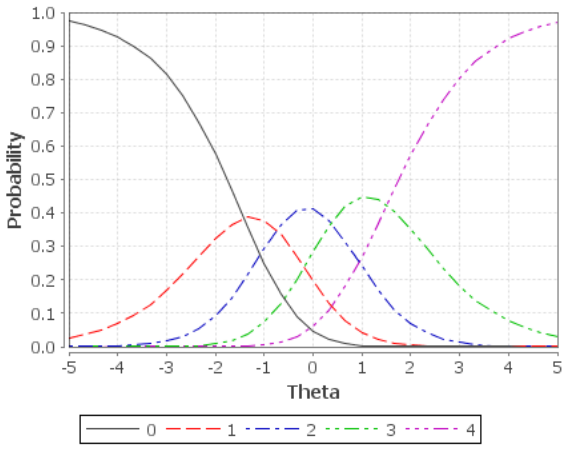
b15



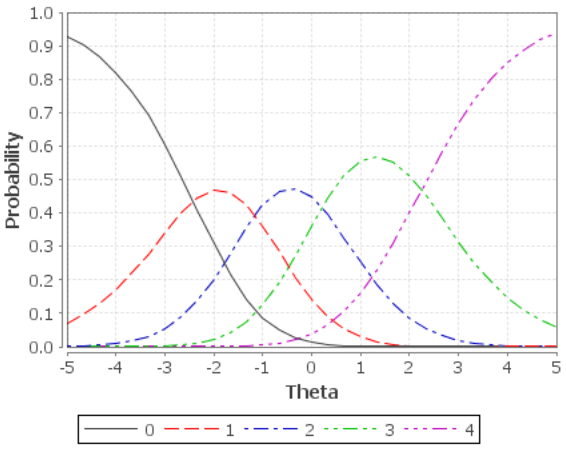
b16



b17

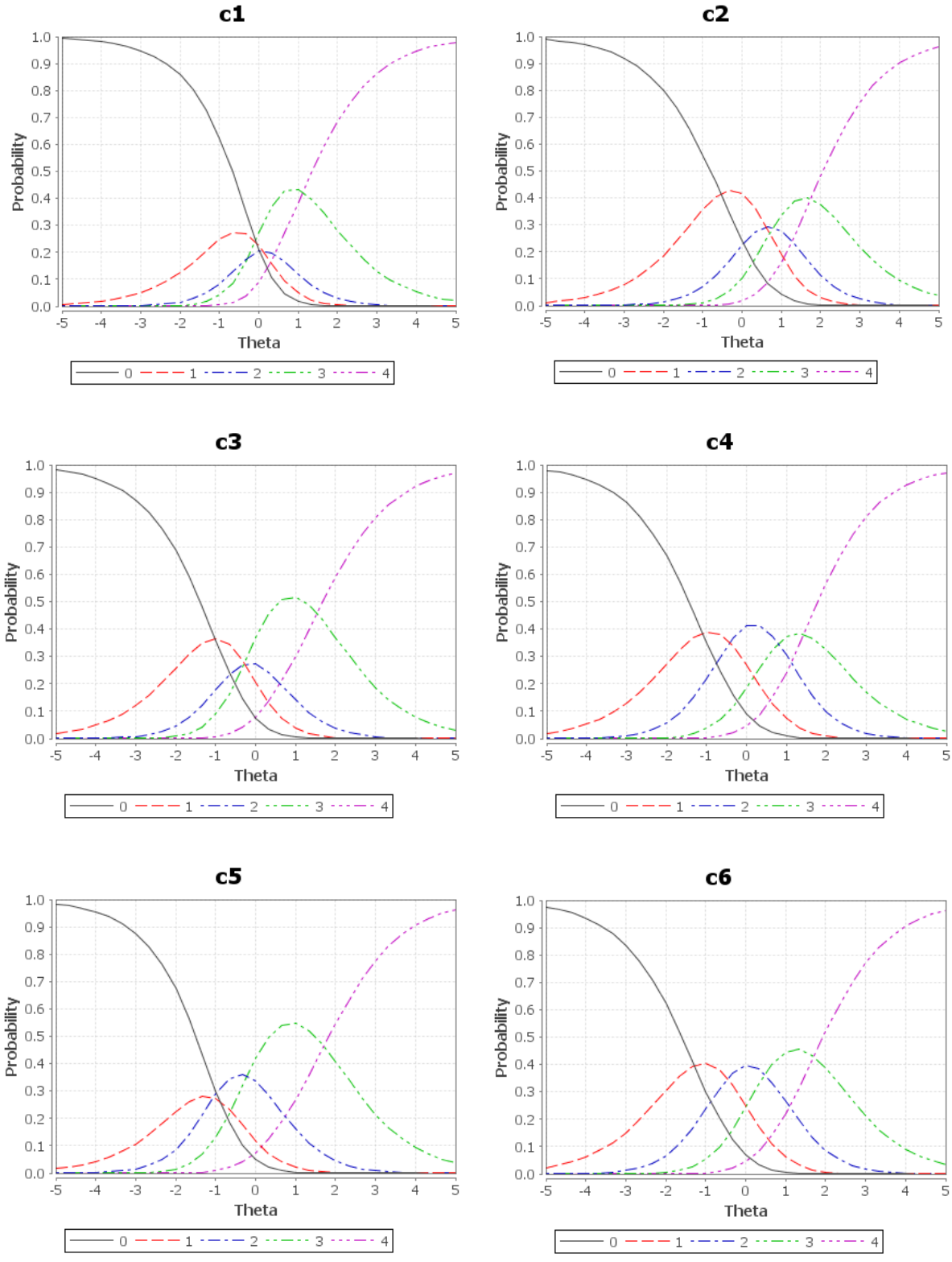


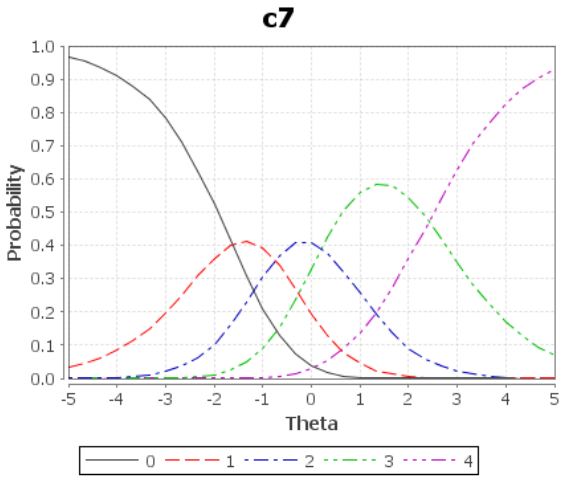
b18

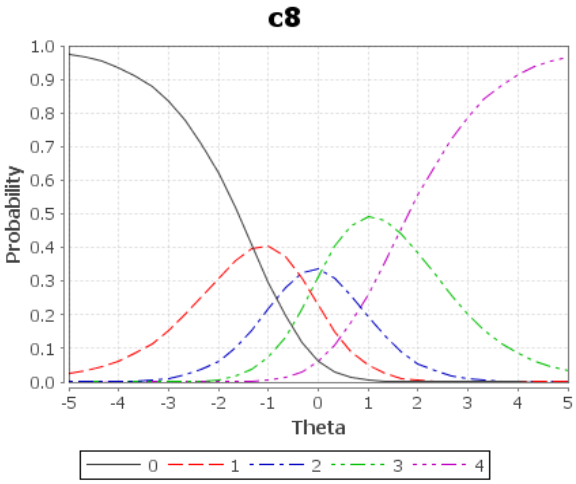


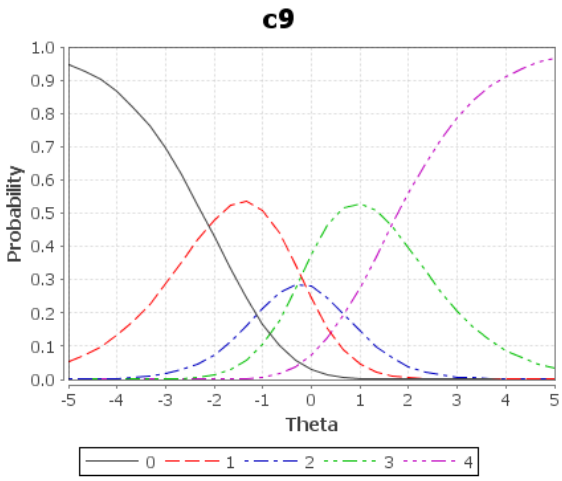
Appendix J

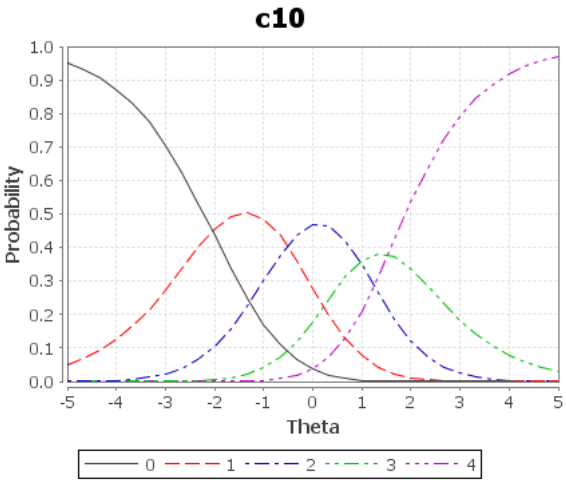
Option Characteristics Curves for Sign C RP.05

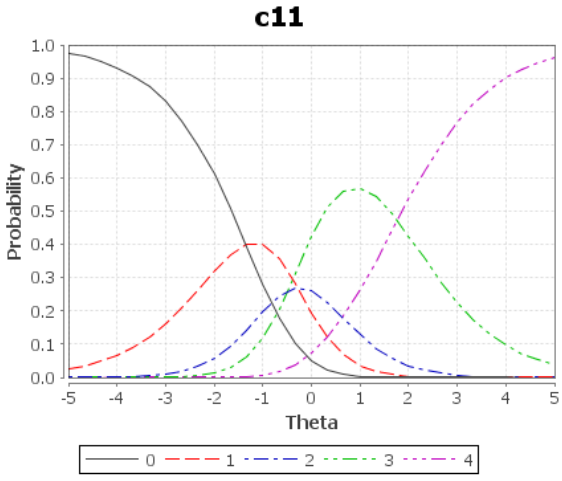


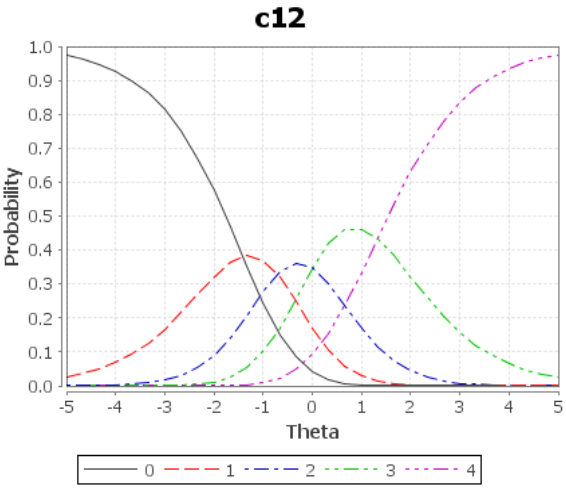

c7
Probability
Theta
0 1 2 3 4

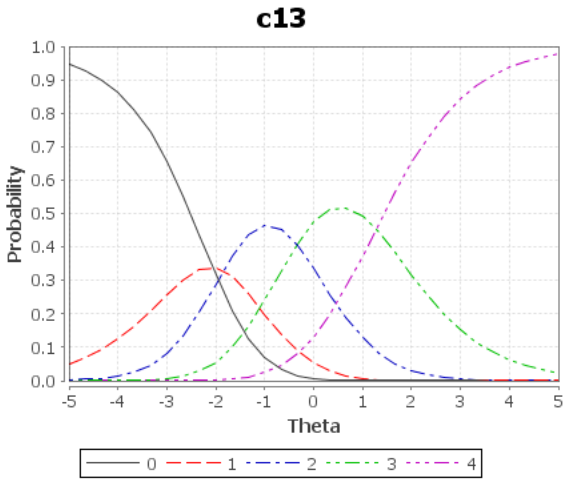

c8
Probability
Theta
0 1 2 3 4

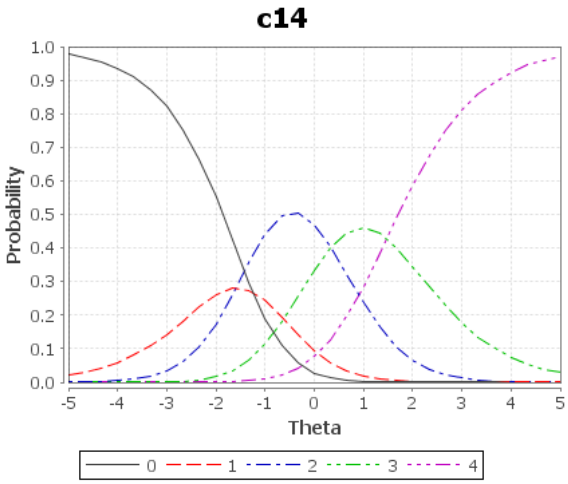

c9
Probability
Theta
0 1 2 3 4

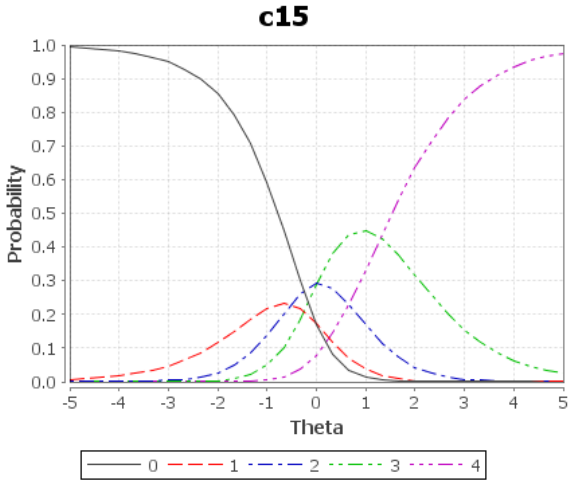

c10
Probability
Theta
0 1 2 3 4

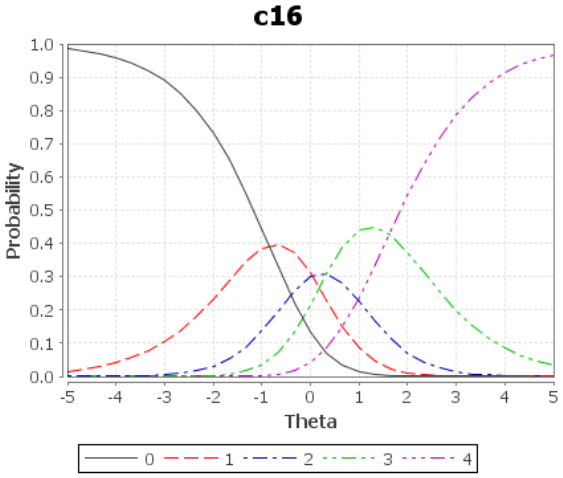

c11
Probability
Theta
0 1 2 3 4

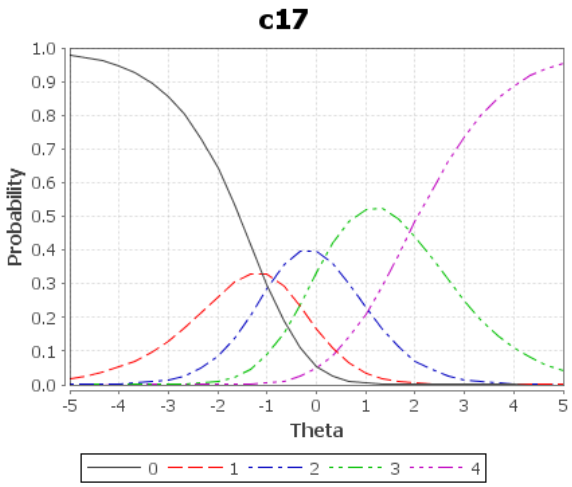

c12
Probability
Theta
0 1 2 3 4

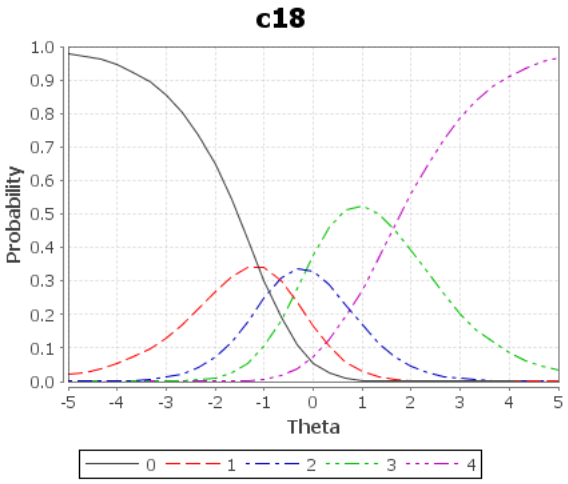

c13
Probability
Theta
0 1 2 3 4


c14
Probability
Theta
0 1 2 3 4


c15
Probability
Theta
0 1 2 3 4

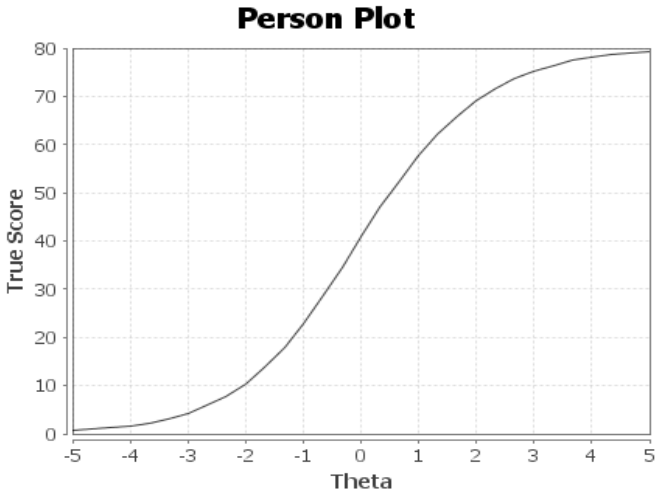

c16
Probability
Theta
0 1 2 3 4


c17
Probability
Theta
0 1 2 3 4

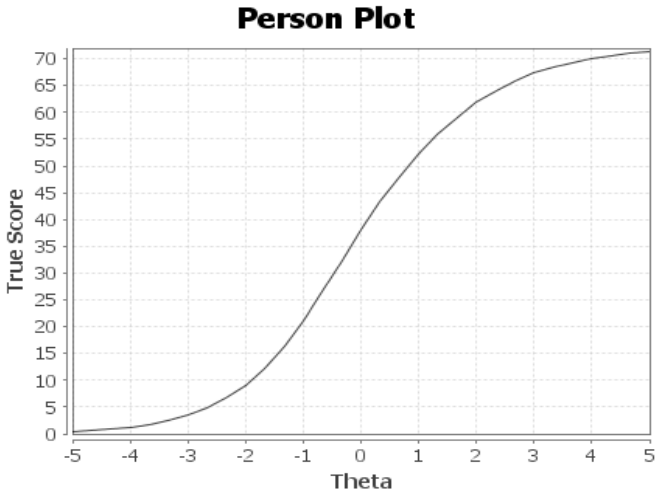

c18
Probability
Theta
0 1 2 3 4

Appendix K

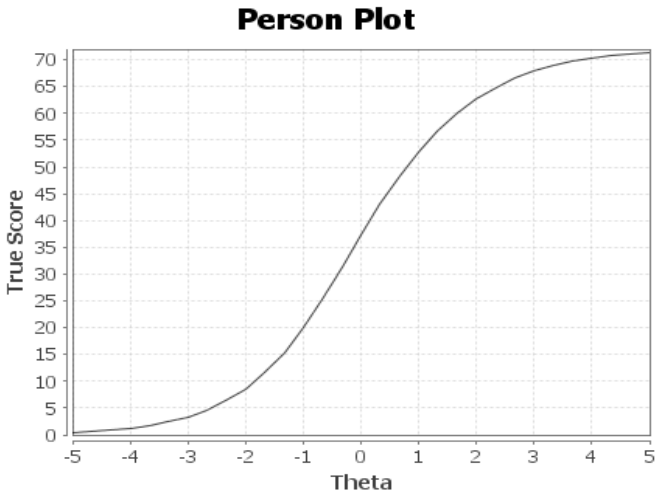
Person Plot for All Test Signs



Person plot for text+symbol sign (sign A, RP.06)



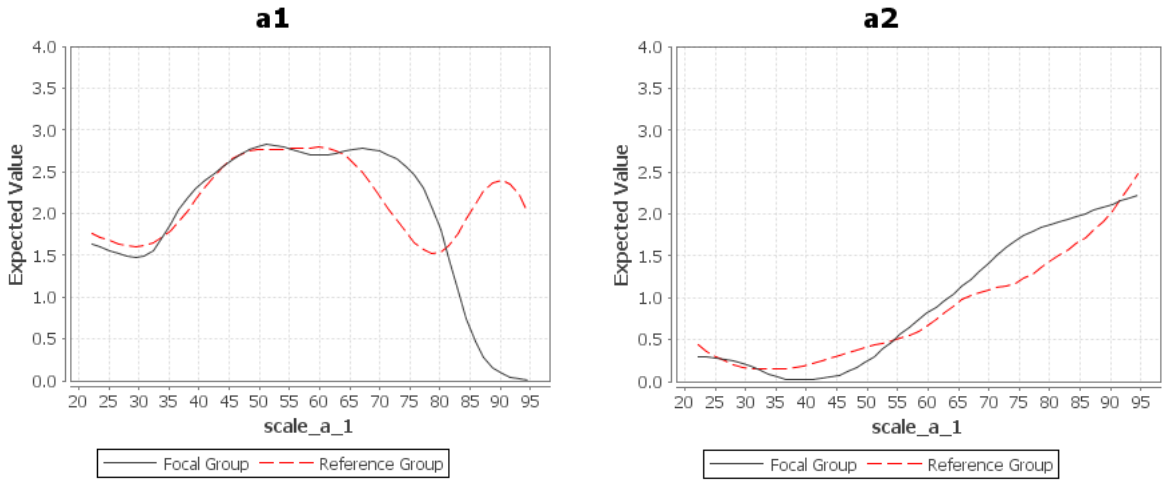
Person plot for symbol-only sign (sign B, RP.10)

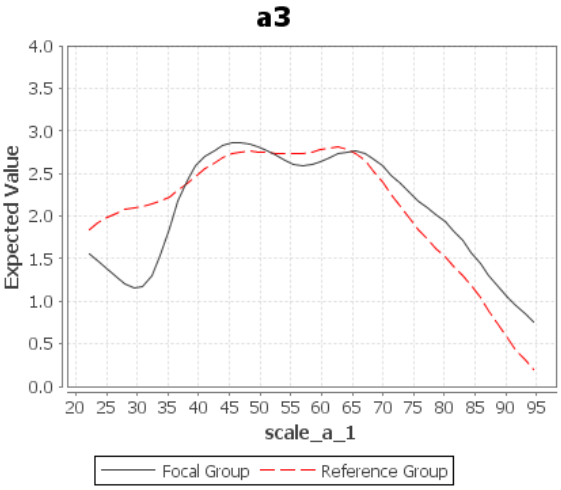


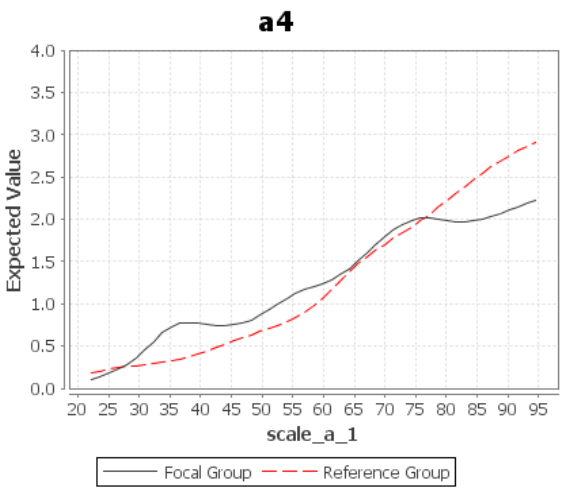
Person plot for symbol-only sign (sign c, RP.05)

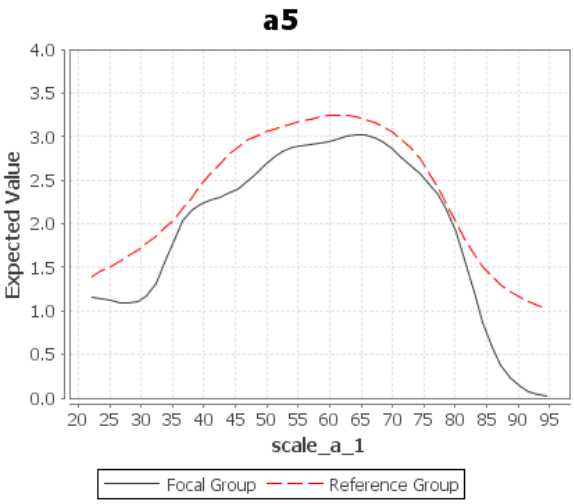
Appendix L

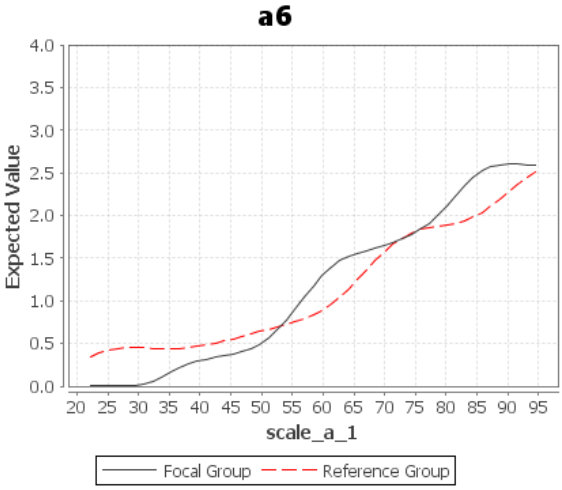
DIF Graph for Sign A RP.06 (Gender)

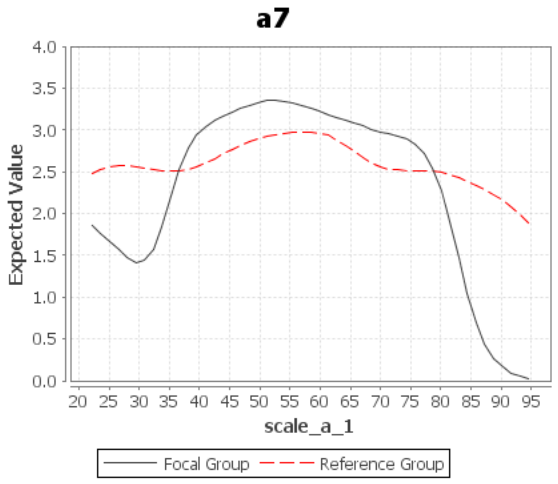


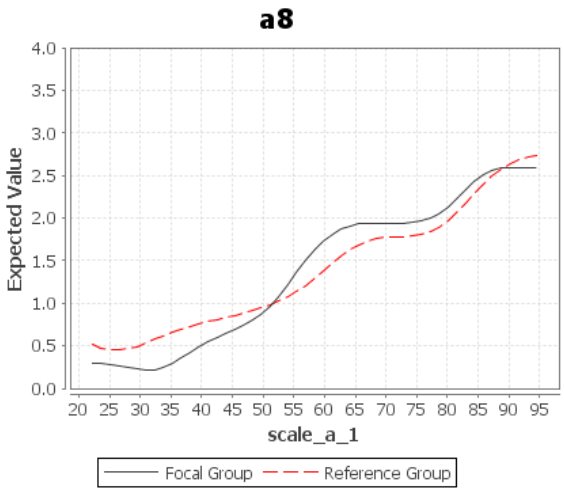


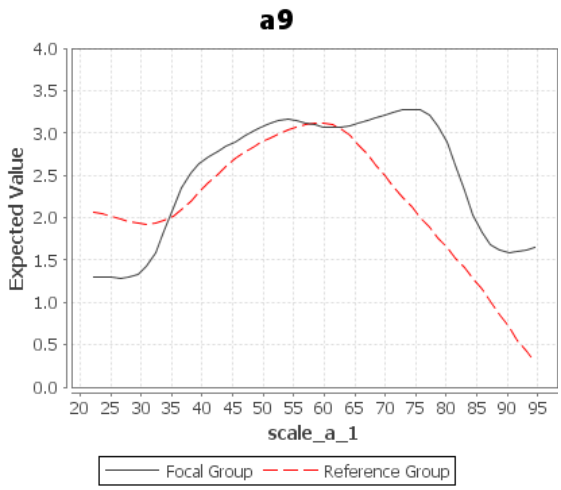


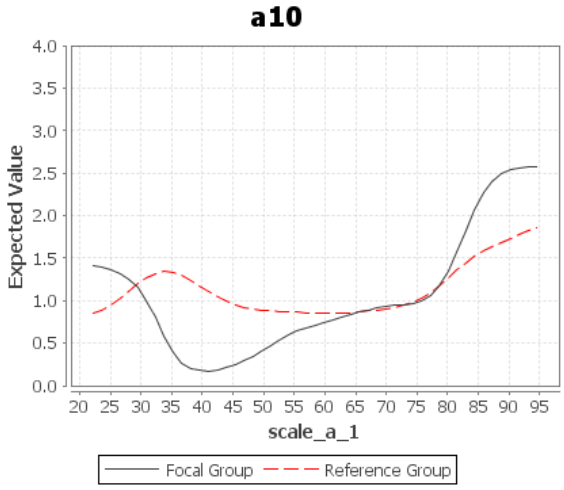


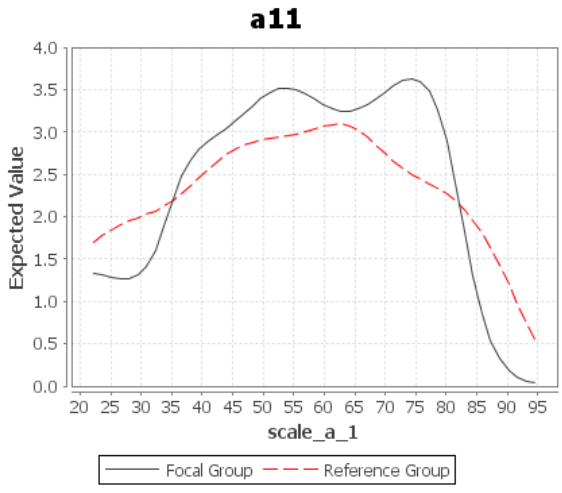


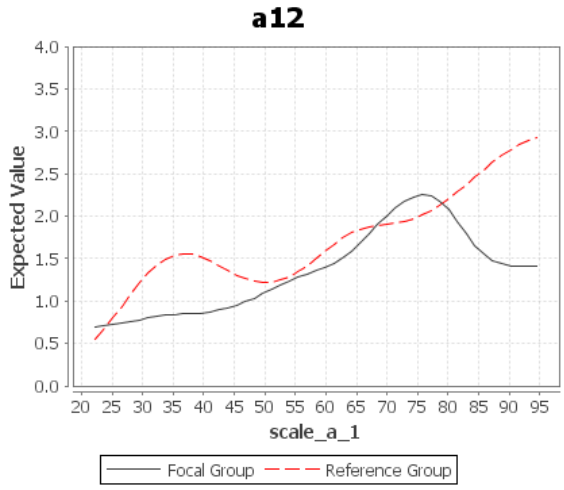

a7
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
scale_a_1
Focal Group
Reference Group

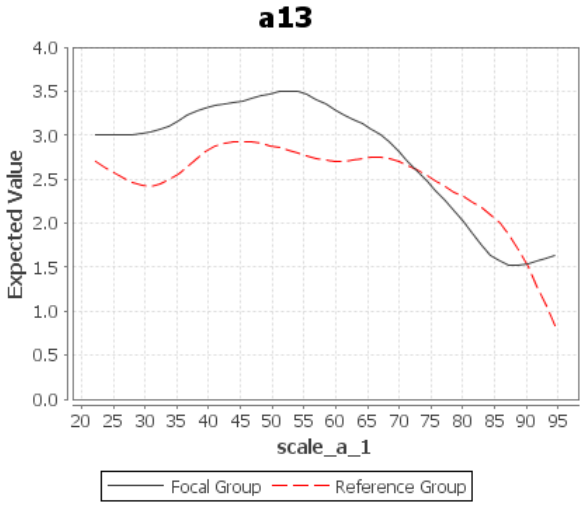

a8
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
scale_a_1
Focal Group
Reference Group

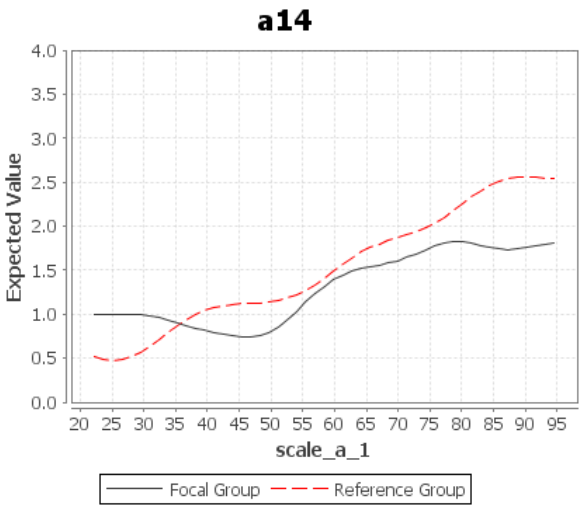

a9
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
scale_a_1
Focal Group
Reference Group

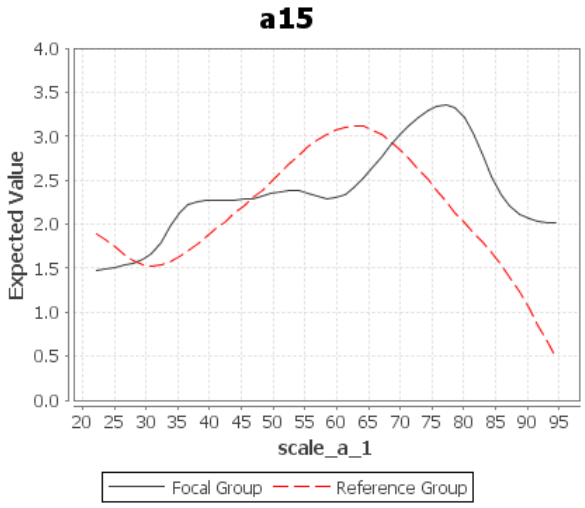

a10
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
scale_a_1
Focal Group
Reference Group

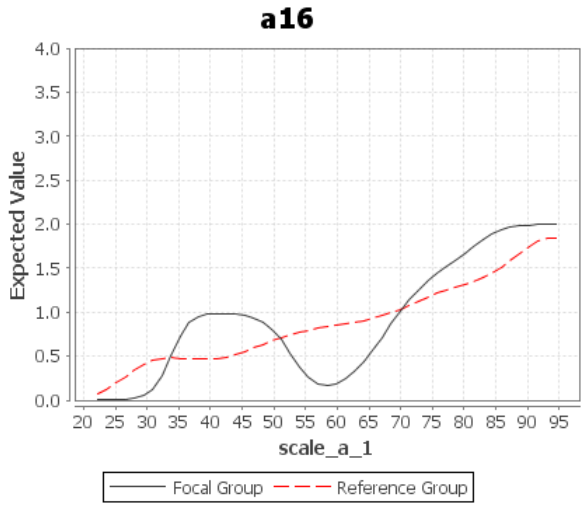

a11
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
scale_a_1
Focal Group
Reference Group

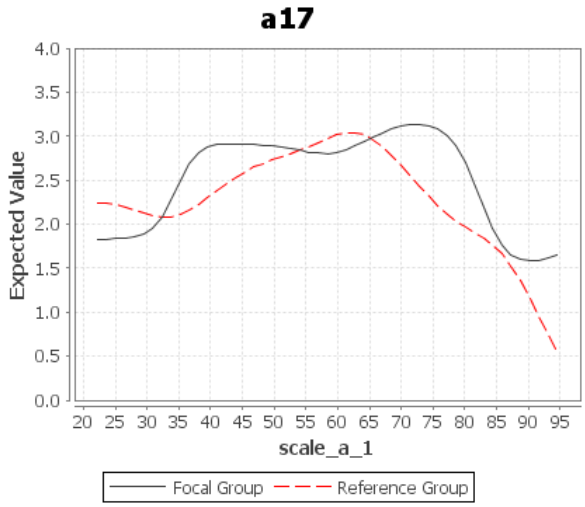

a12
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
scale_a_1
Focal Group
Reference Group

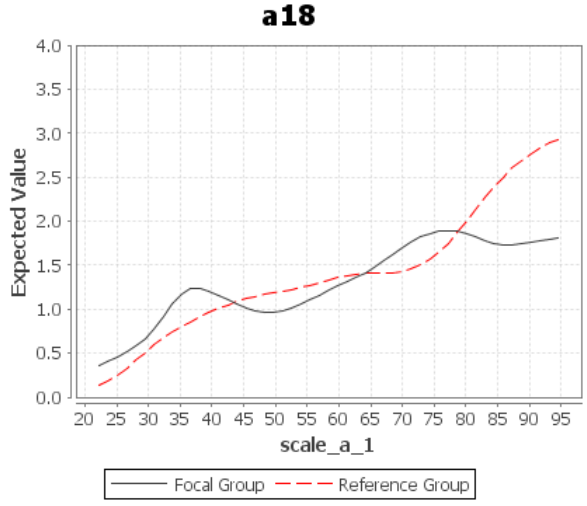

a13
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
scale_a_1
Focal Group
Reference Group

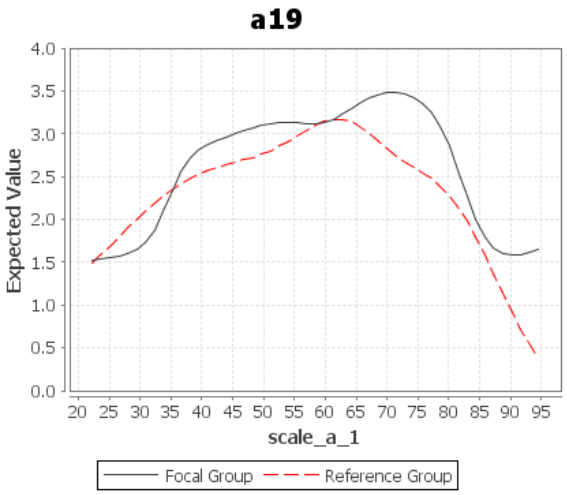

a14
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
scale_a_1
Focal Group
Reference Group

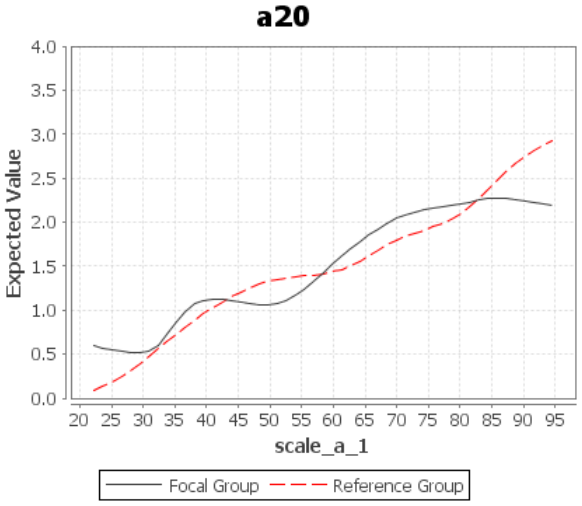

a15
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
scale_a_1
Focal Group
Reference Group


a16
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
scale_a_1
Focal Group
Reference Group


a17
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
scale_a_1
Focal Group
Reference Group

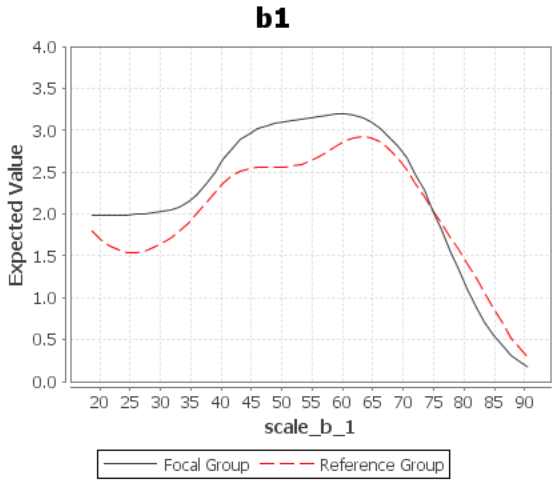

a18
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
scale_a_1
Focal Group
Reference Group

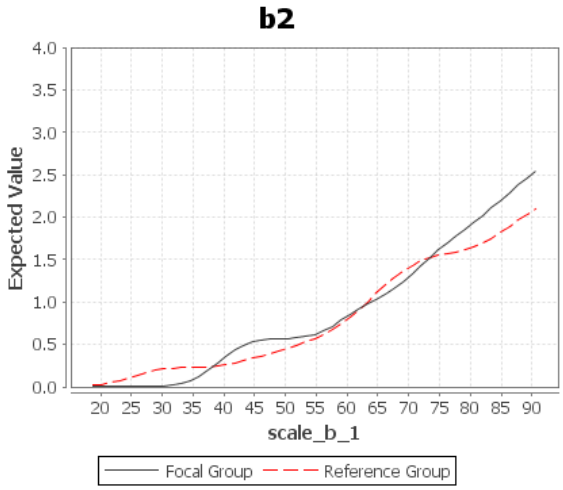

a19
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
scale_a_1
Focal Group
Reference Group

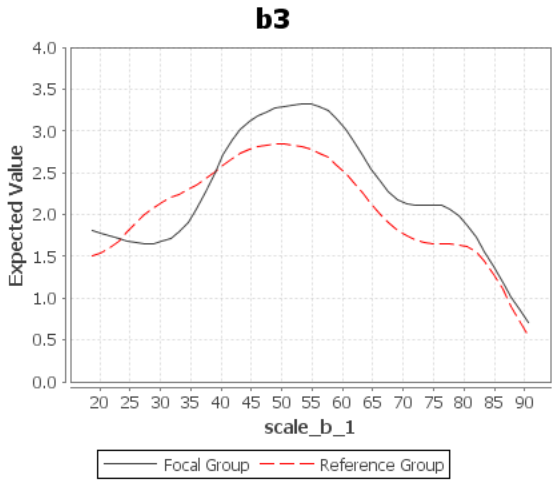

a20
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
scale_a_1
Focal Group
Reference Group

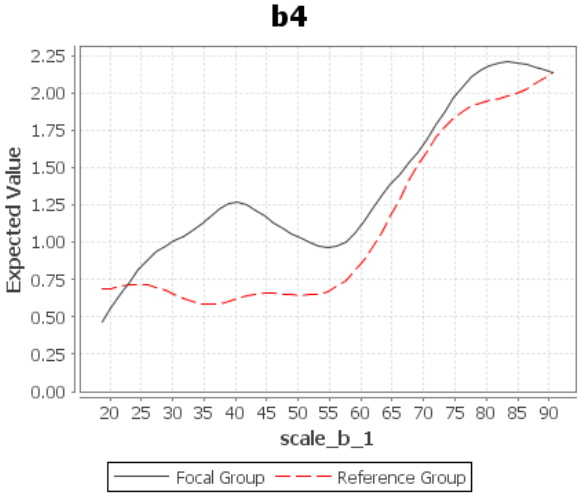
Appendix M

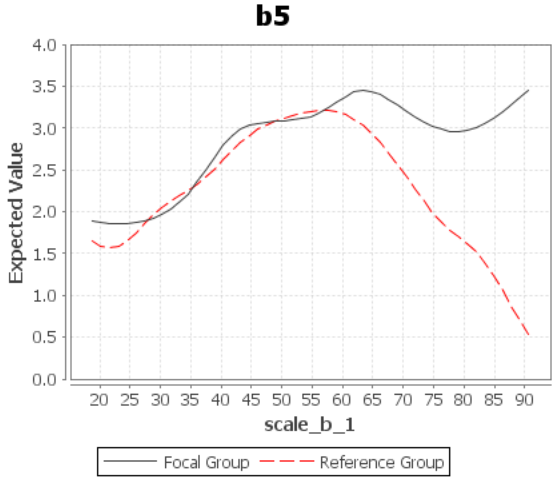
DIF Graphs for Sign B RP.10 (Gender)

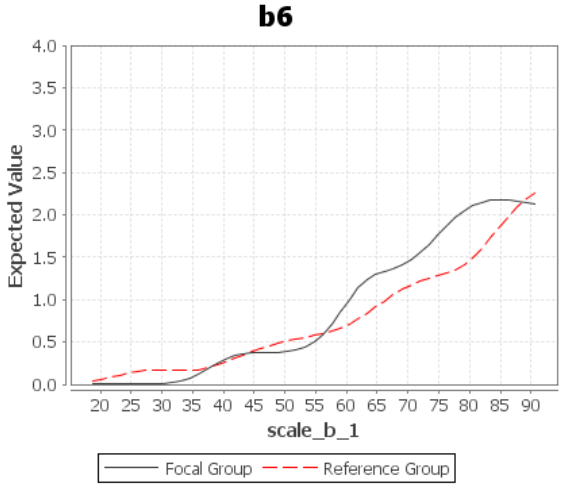


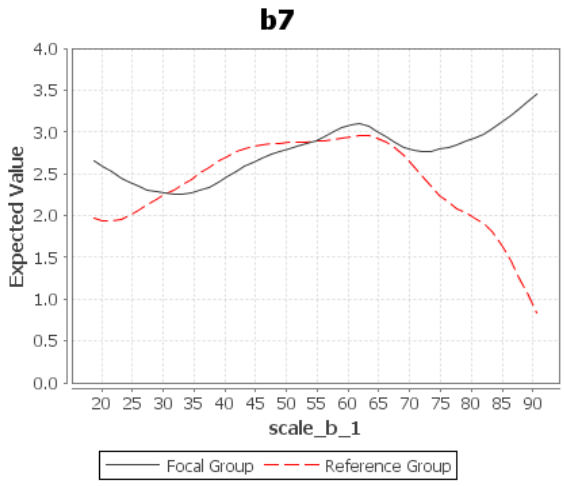


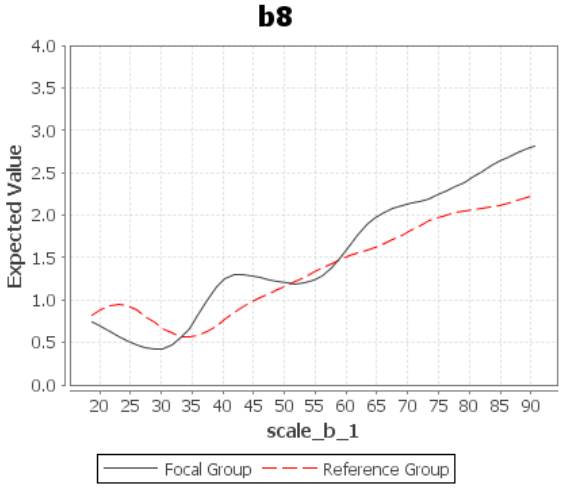


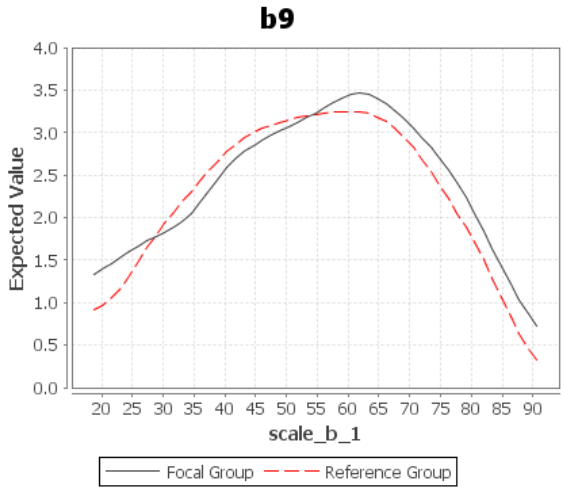


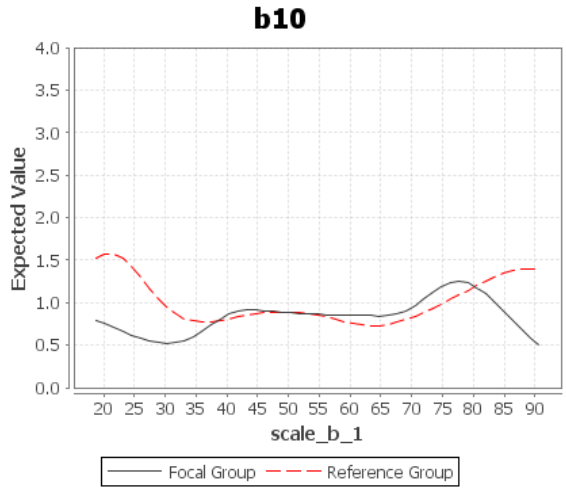


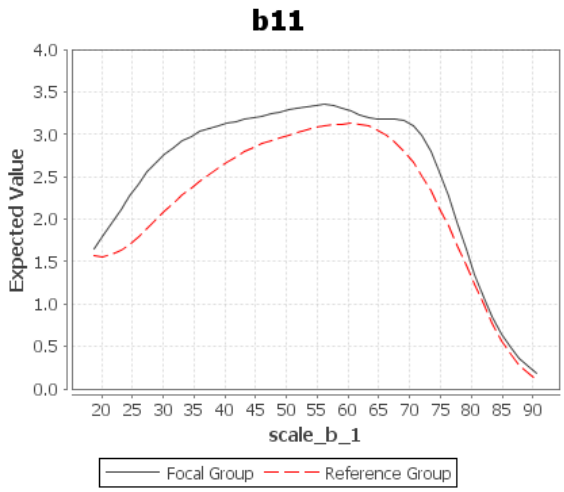


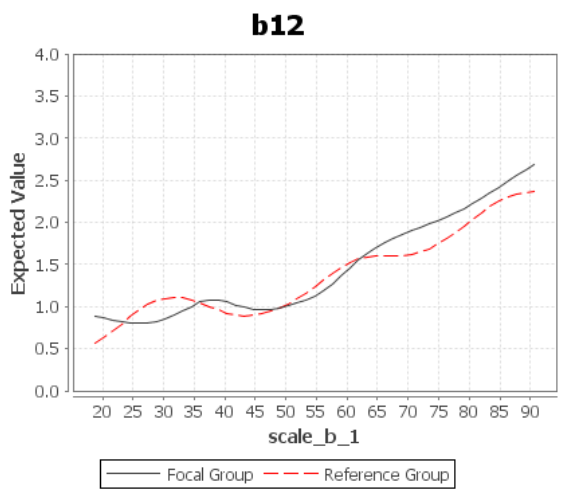

b7
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20 25 30 35 40 45 50 55 60 65 70 75 80 85 90
scale_b_1
Focal Group
Reference Group

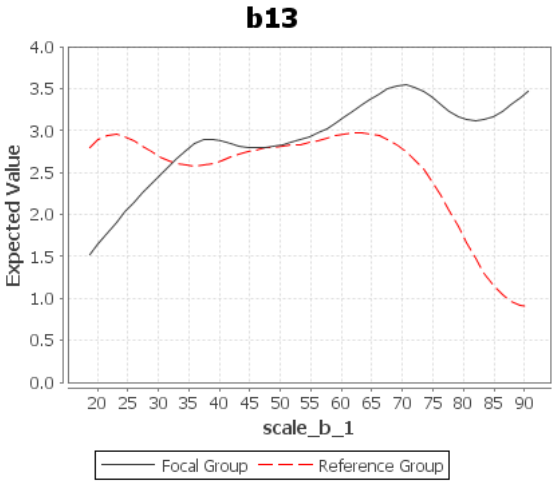

b8
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20 25 30 35 40 45 50 55 60 65 70 75 80 85 90
scale_b_1
Focal Group
Reference Group

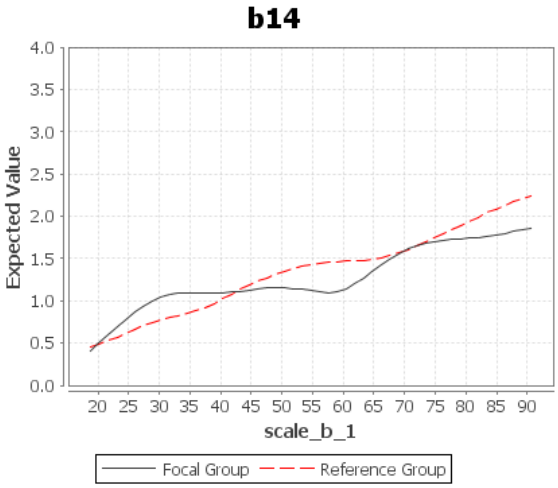

b9
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20 25 30 35 40 45 50 55 60 65 70 75 80 85 90
scale_b_1
Focal Group
Reference Group

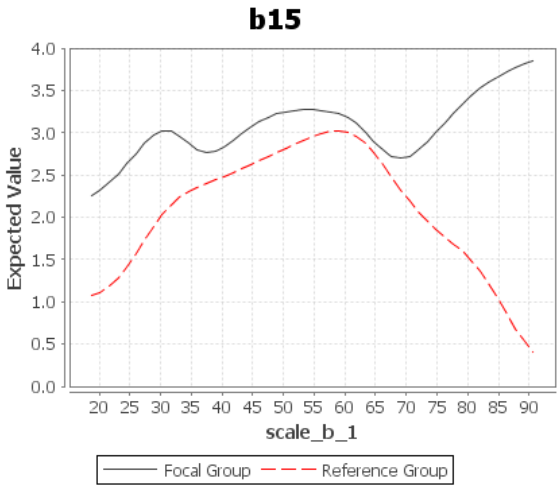

b10
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20 25 30 35 40 45 50 55 60 65 70 75 80 85 90
scale_b_1
Focal Group
Reference Group

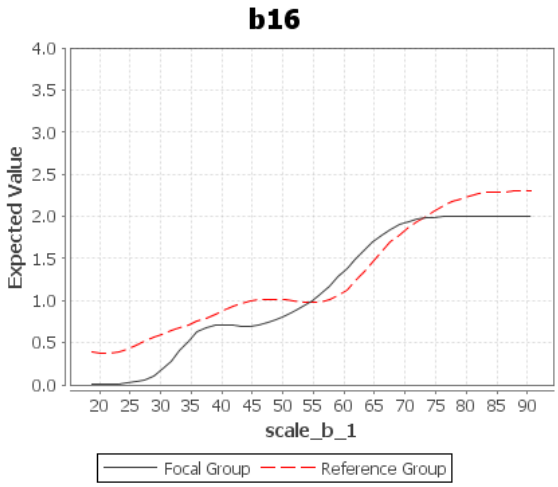

b11
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20 25 30 35 40 45 50 55 60 65 70 75 80 85 90
scale_b_1
Focal Group
Reference Group

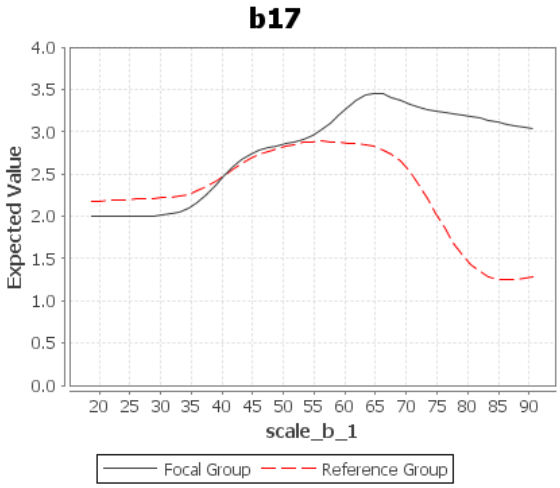

b12
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20 25 30 35 40 45 50 55 60 65 70 75 80 85 90
scale_b_1
Focal Group
Reference Group

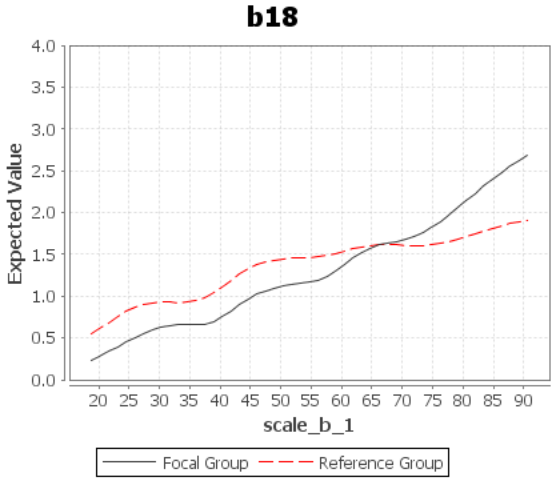

b13
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20 25 30 35 40 45 50 55 60 65 70 75 80 85 90
scale_b_1
Focal Group
Reference Group


b14
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20 25 30 35 40 45 50 55 60 65 70 75 80 85 90
scale_b_1
Focal Group
Reference Group


b15
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20 25 30 35 40 45 50 55 60 65 70 75 80 85 90
scale_b_1
Focal Group
Reference Group

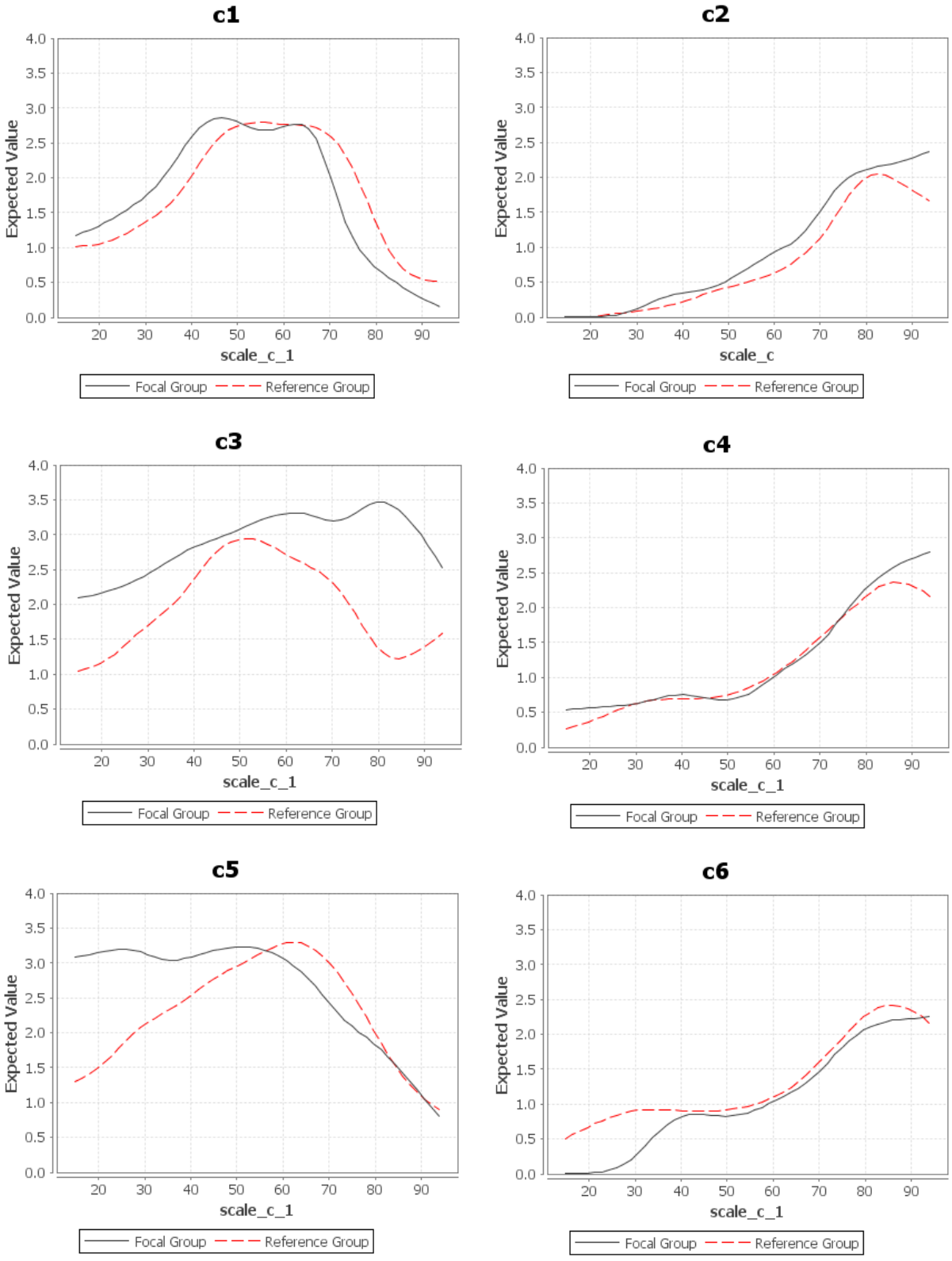

b16
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20 25 30 35 40 45 50 55 60 65 70 75 80 85 90
scale_b_1
Focal Group
Reference Group

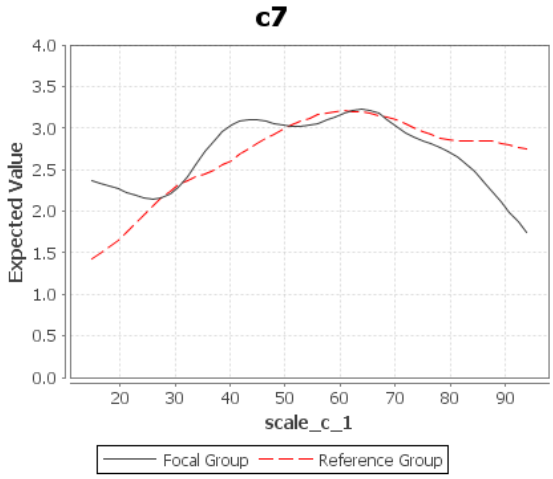

b17
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20 25 30 35 40 45 50 55 60 65 70 75 80 85 90
scale_b_1
Focal Group
Reference Group

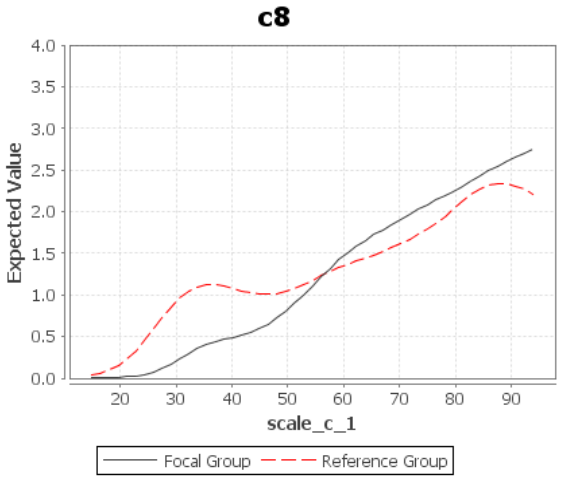

b18
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20 25 30 35 40 45 50 55 60 65 70 75 80 85 90
scale_b_1
Focal Group
Reference Group

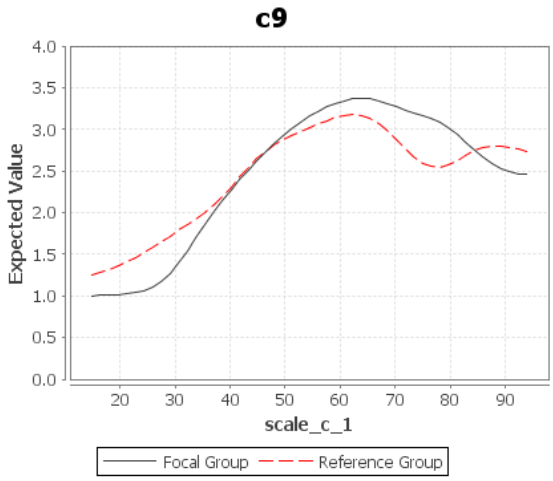
Appendix N

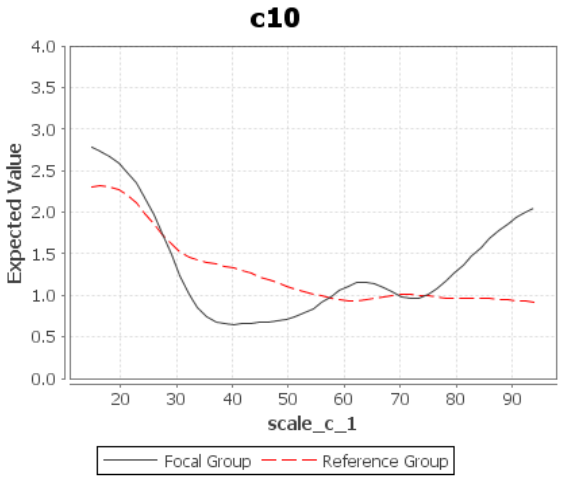
DIF Graphs for Sign C RP.05 (Gender)

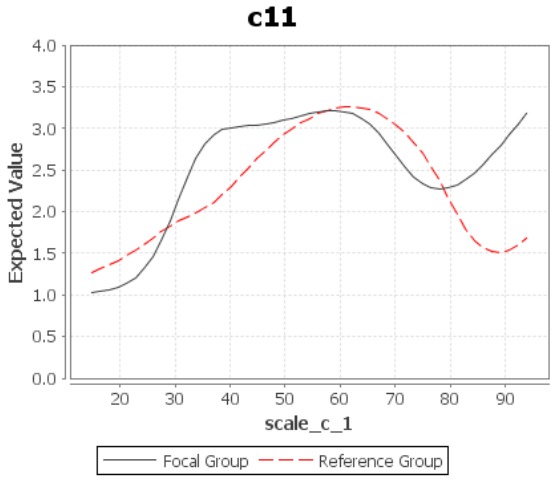


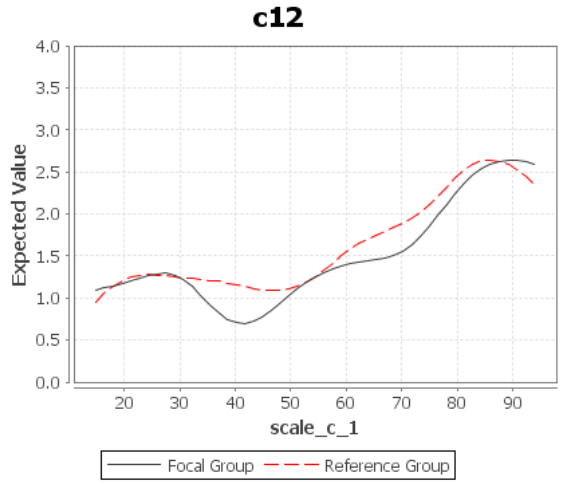

c7
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20
30
40
50
60
70
80
90
scale_c_1
Focal Group
Reference Group

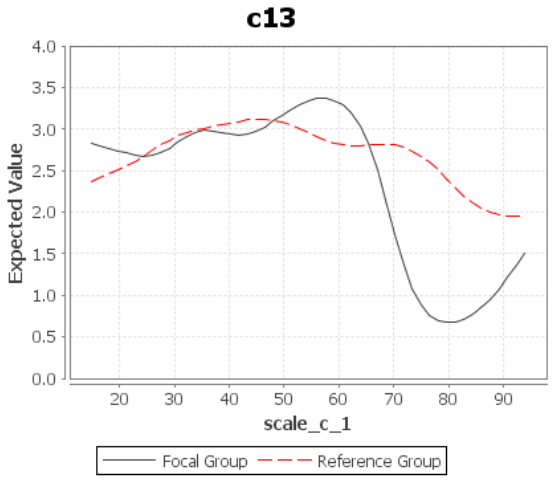

c8
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20
30
40
50
60
70
80
90
scale_c_1
Focal Group
Reference Group

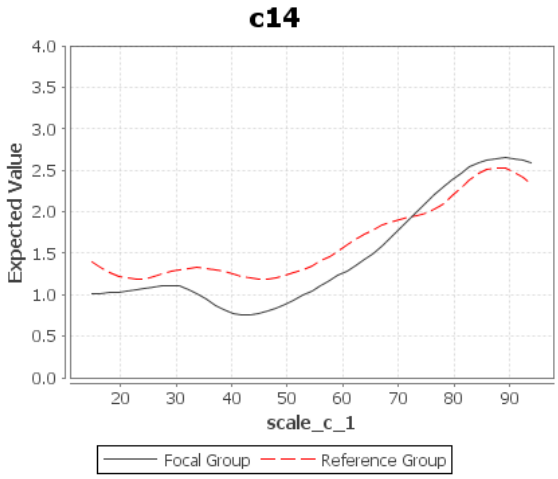

c9
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20
30
40
50
60
70
80
90
scale_c_1
Focal Group
Reference Group

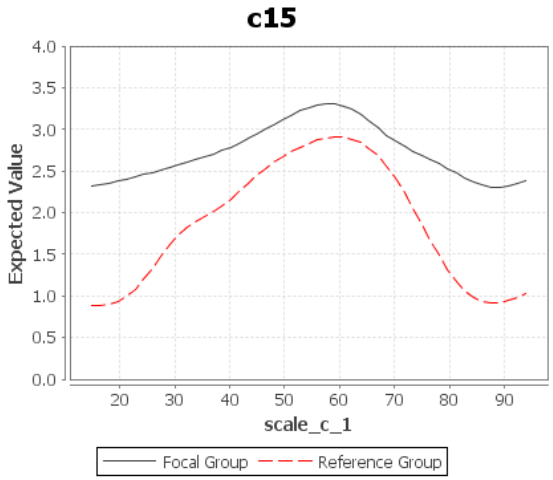

c10
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20
30
40
50
60
70
80
90
scale_c_1
Focal Group
Reference Group

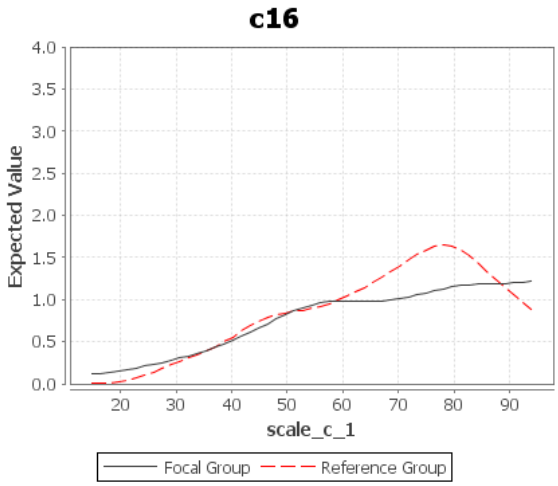

c11
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20
30
40
50
60
70
80
90
scale_c_1
Focal Group
Reference Group

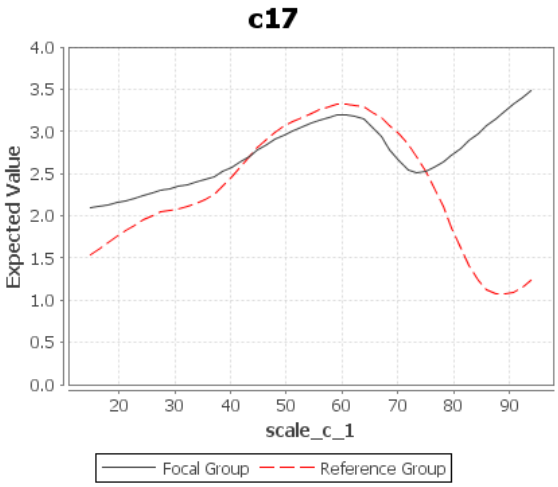

c12
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20
30
40
50
60
70
80
90
scale_c_1
Focal Group
Reference Group

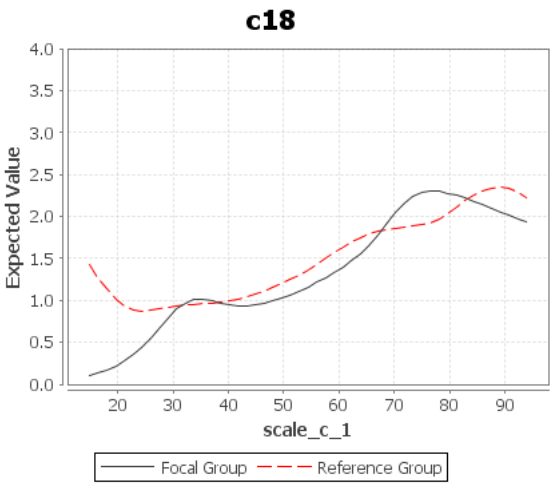

c13
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20 30 40 50 60 70 80 90
scale_c_1
Focal Group
Reference Group


c14
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20 30 40 50 60 70 80 90
scale_c_1
Focal Group
Reference Group


c15
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20 30 40 50 60 70 80 90
scale_c_1
Focal Group
Reference Group

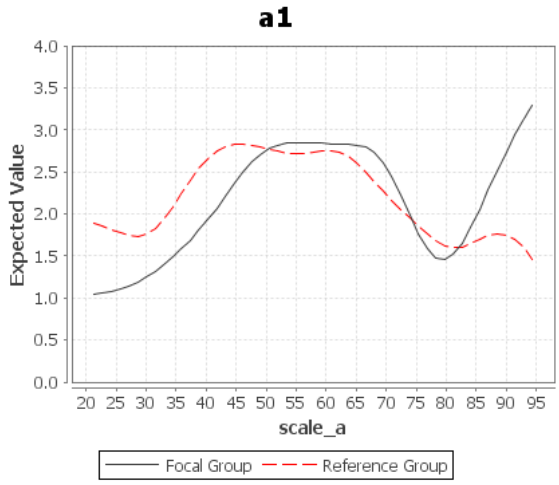

c16
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20 30 40 50 60 70 80 90
scale_c_1
Focal Group
Reference Group

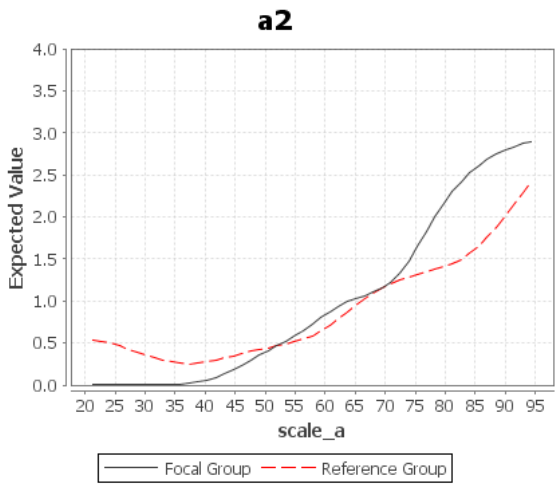

c17
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20 30 40 50 60 70 80 90
scale_c_1
Focal Group
Reference Group

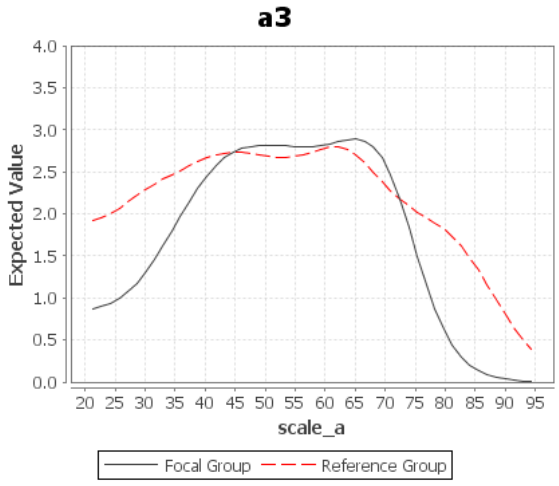

c18
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20 30 40 50 60 70 80 90
scale_c_1
Focal Group
Reference Group

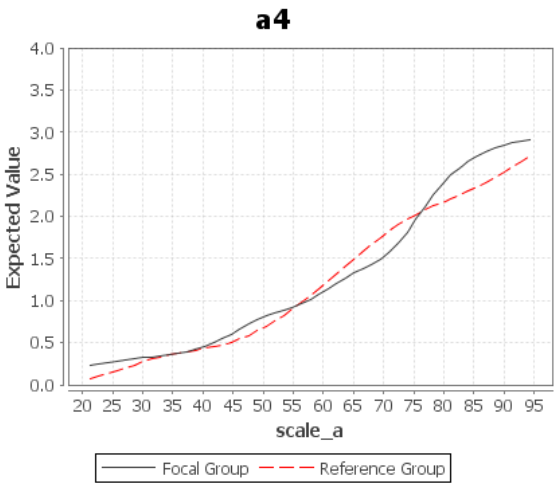
Appendix O

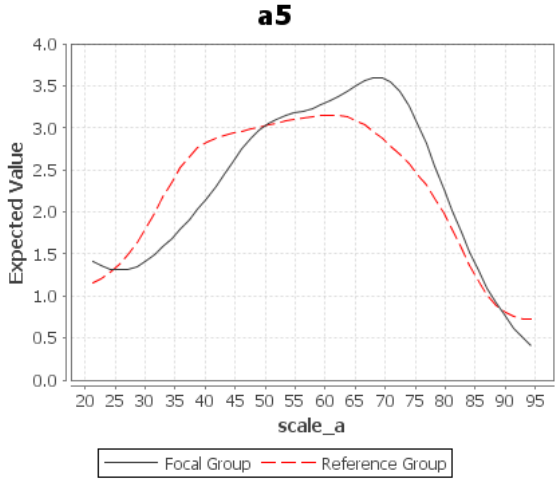
DIF Graph for Sign A RP.06 (License Status)

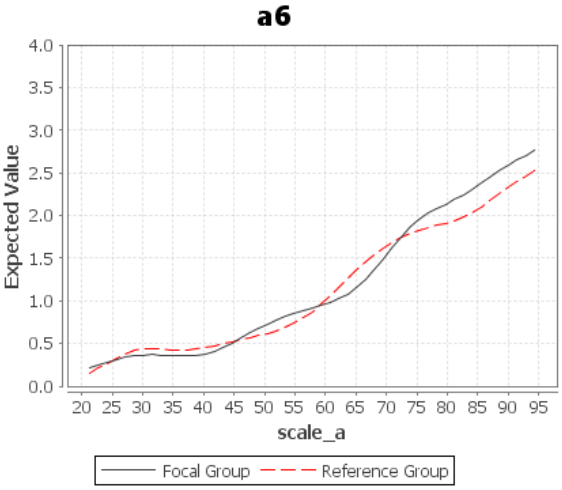


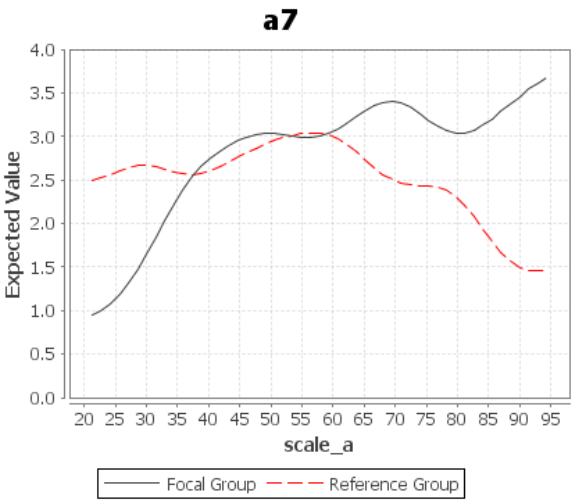


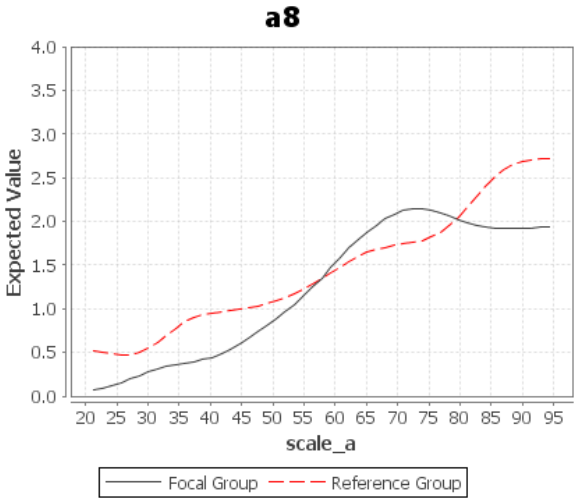


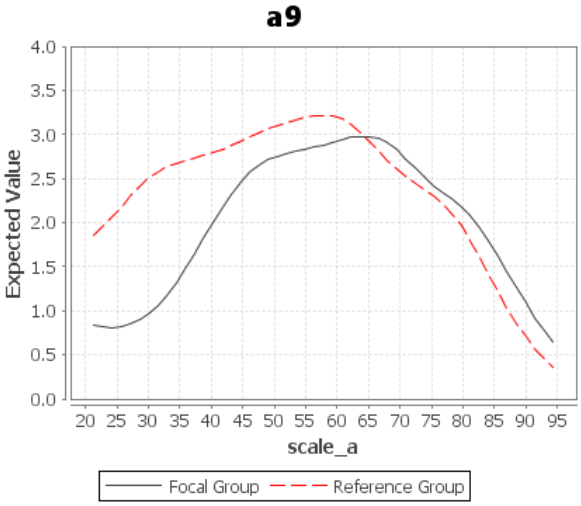


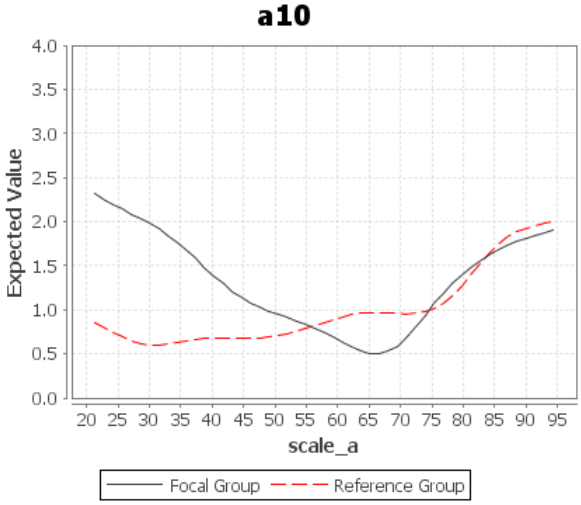


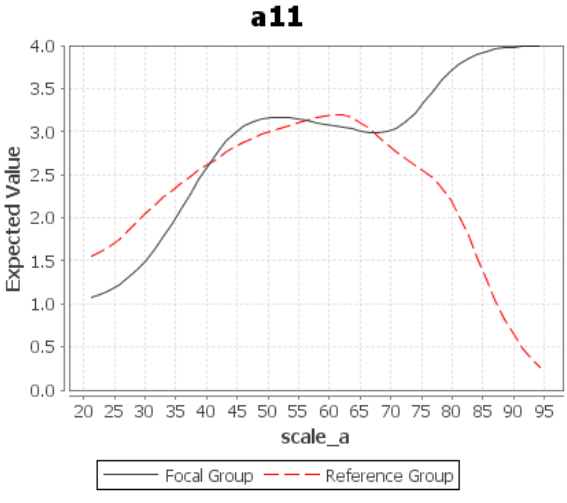


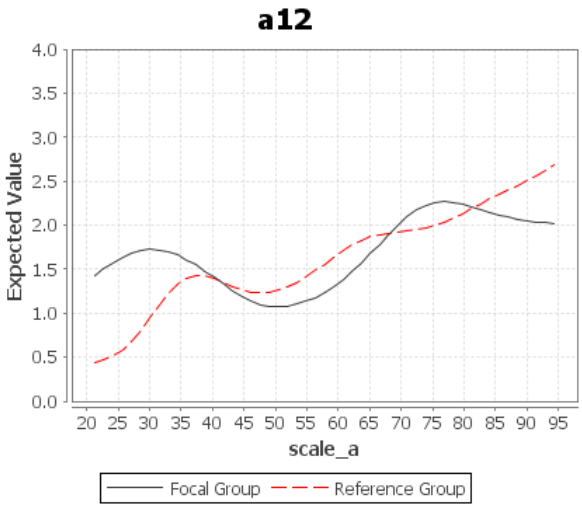

a7
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
scale_a
Focal Group
Reference Group

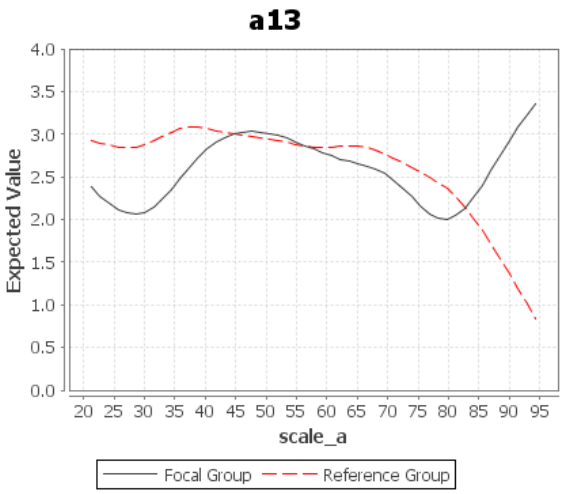

a8
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
scale_a
Focal Group
Reference Group

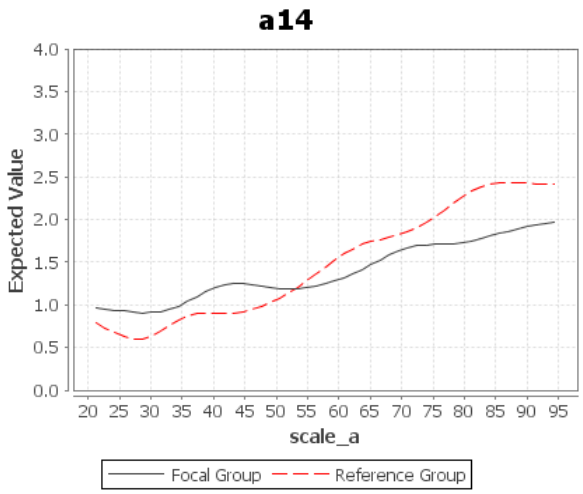

a9
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
scale_a
Focal Group
Reference Group

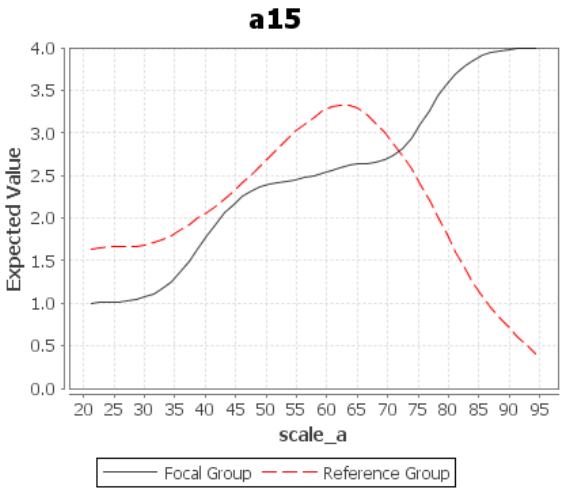

a10
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
scale_a
Focal Group
Reference Group

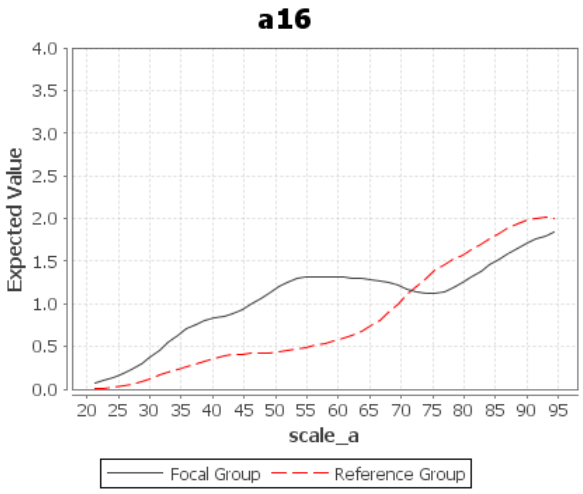

a11
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
scale_a
Focal Group
Reference Group

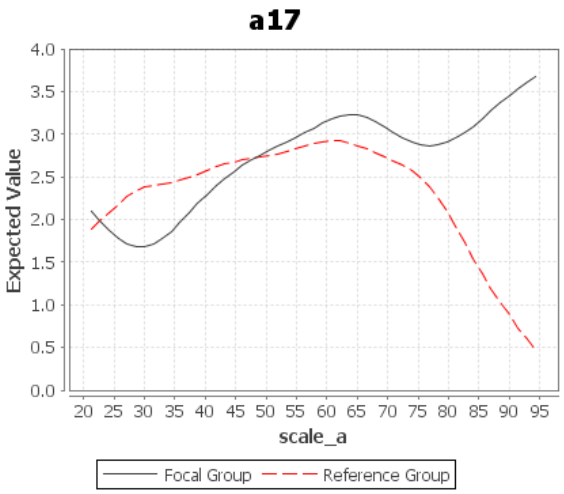

a12
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
scale_a
Focal Group
Reference Group

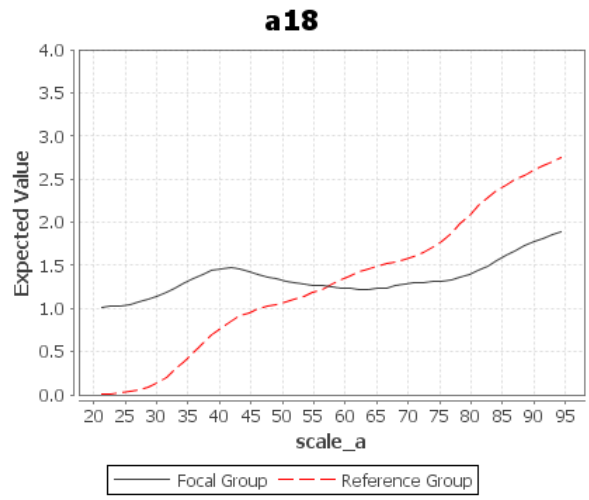

a13
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
scale_a
Focal Group
Reference Group


a14
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
scale_a
Focal Group
Reference Group


a15
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
scale_a
Focal Group
Reference Group

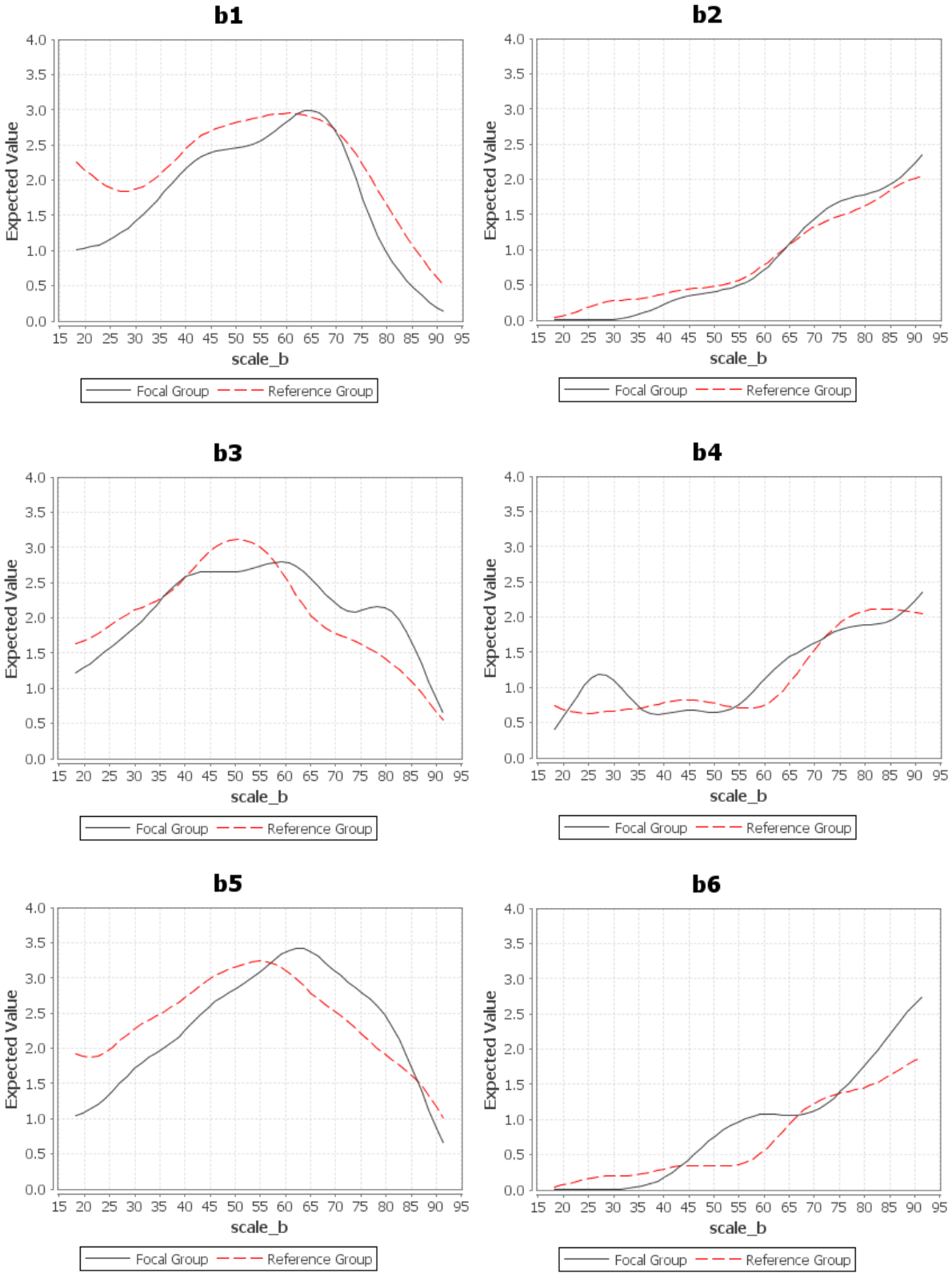

a16
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
scale_a
Focal Group
Reference Group

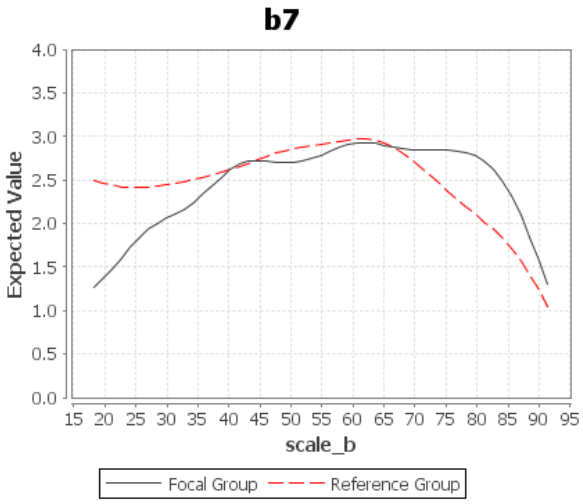

a17
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
scale_a
Focal Group
Reference Group

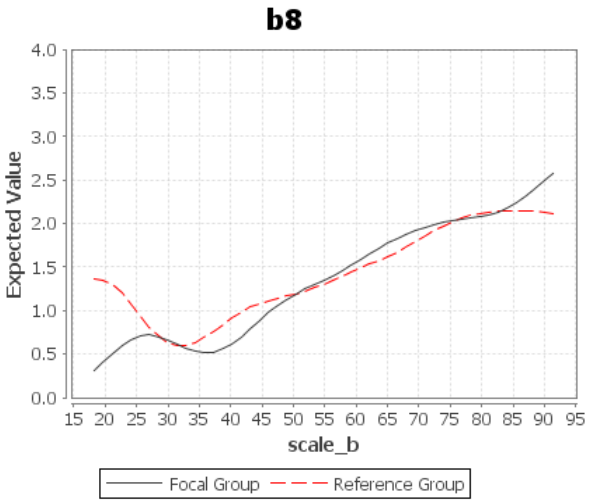

a18
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
scale_a
Focal Group
Reference Group

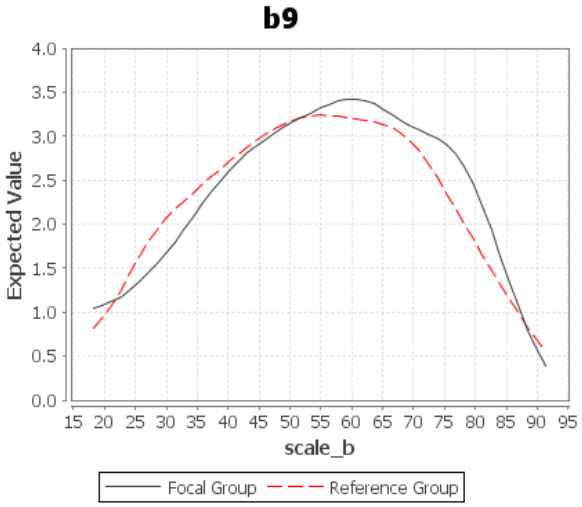
Appendix P

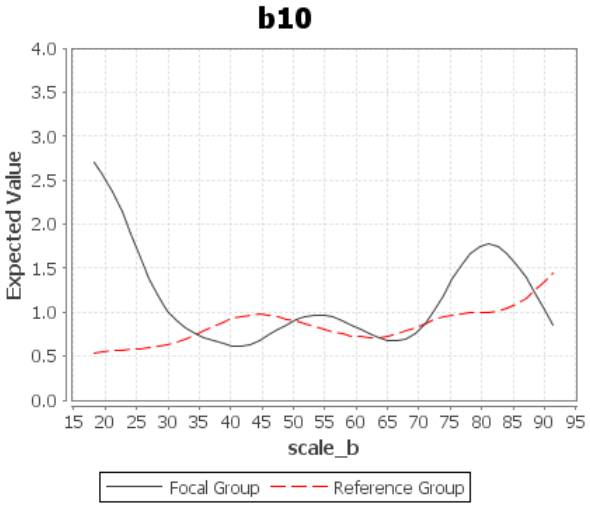
DIF Graph for Sign B RP.10 (License Status)

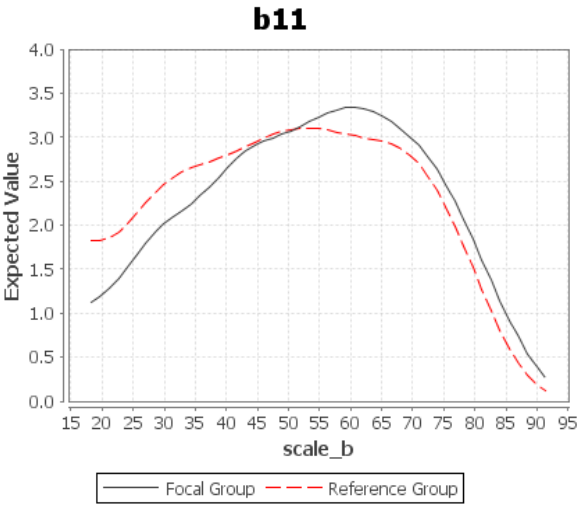


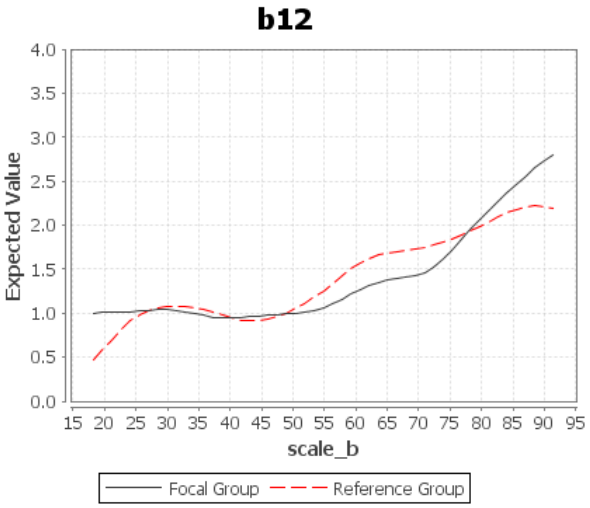

b7
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
scale_b
Focal Group
Reference Group

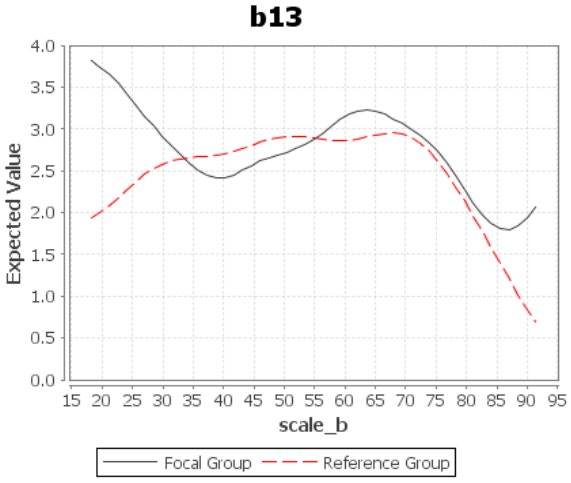

b8
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
scale_b
Focal Group
Reference Group

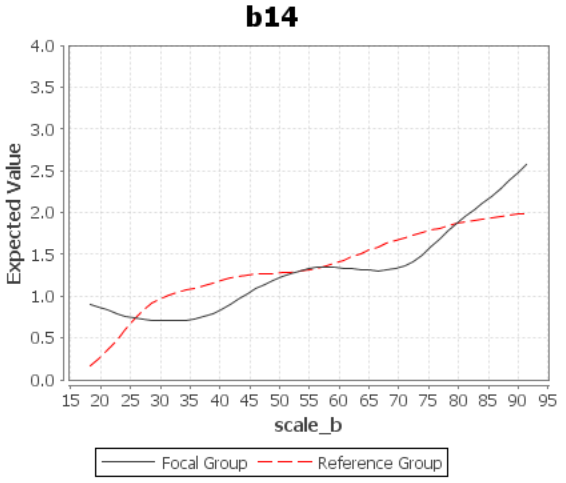

b9
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
scale_b
Focal Group
Reference Group

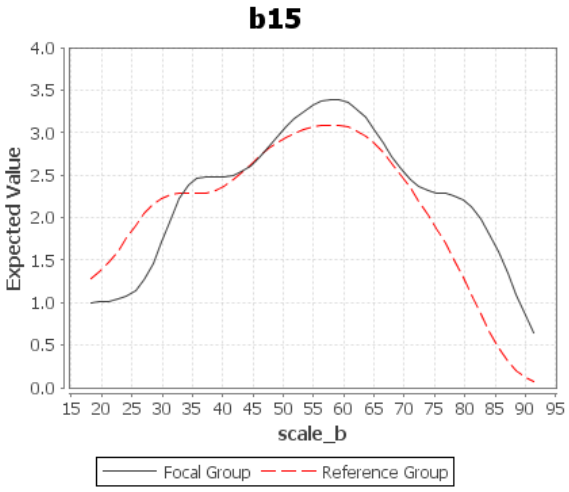

b10
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
scale_b
Focal Group
Reference Group

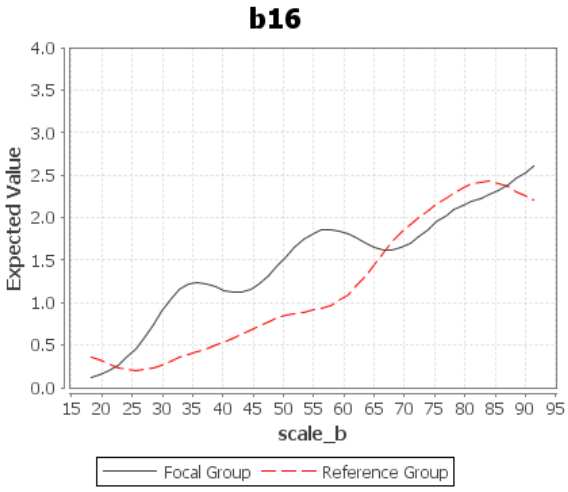

b11
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
scale_b
Focal Group
Reference Group

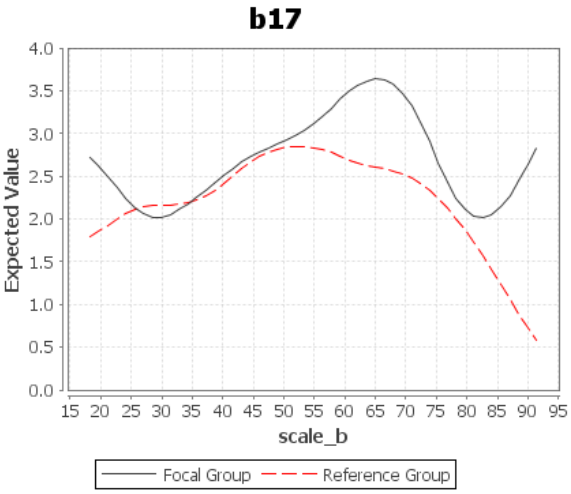

b12
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
scale_b
Focal Group
Reference Group

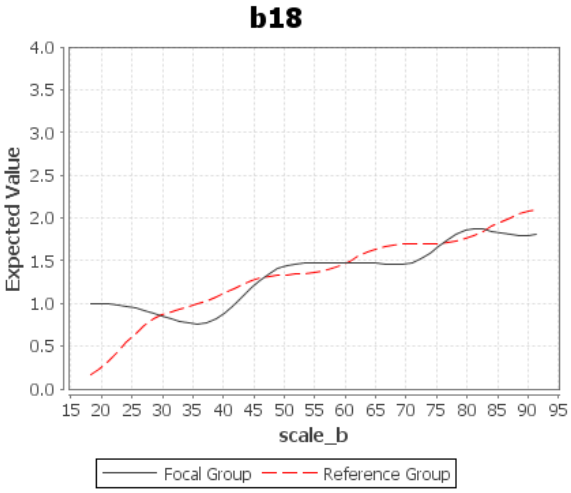

b13
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
scale_b
Focal Group
Reference Group


b14
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
scale_b
Focal Group
Reference Group


b15
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
scale_b
Focal Group
Reference Group

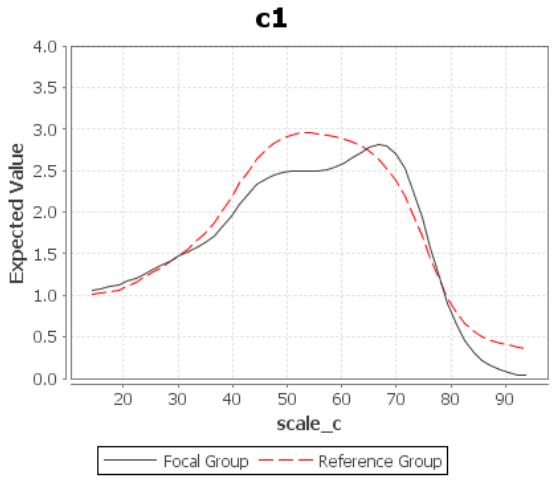

b16
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
scale_b
Focal Group
Reference Group

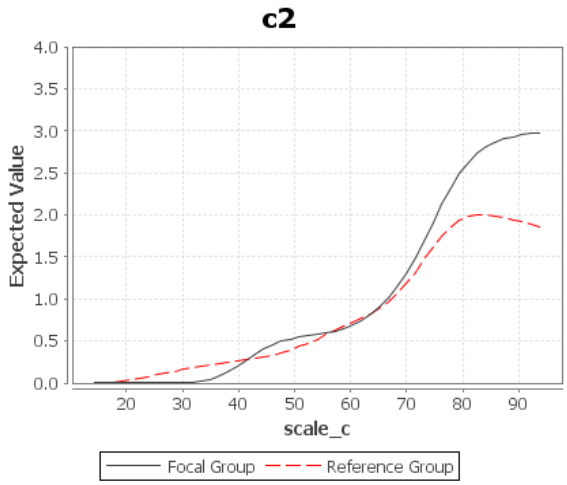

b17
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
scale_b
Focal Group
Reference Group

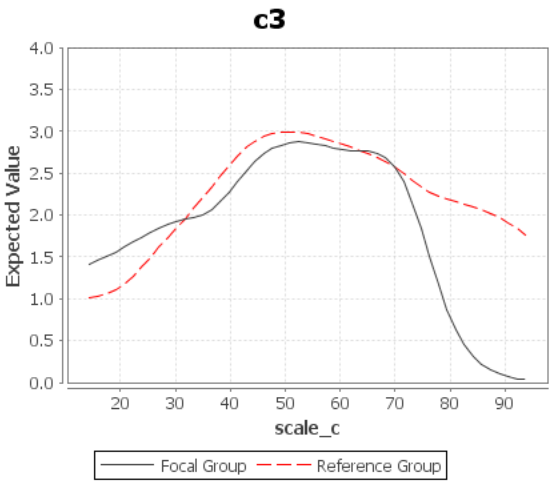

b18
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
scale_b
Focal Group
Reference Group

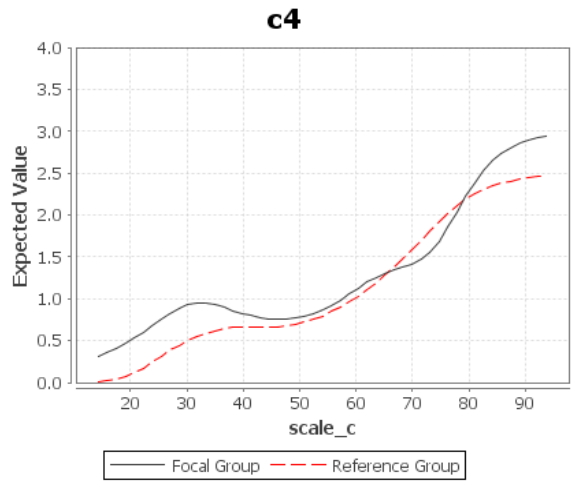
Appendix Q

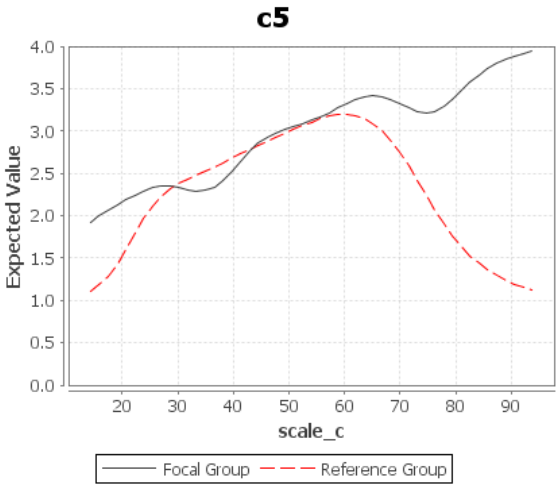
DIF Graph for Sign C RP.05 (License Status)

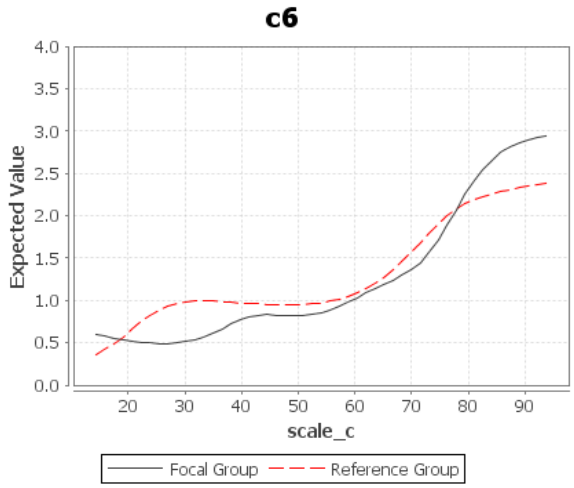


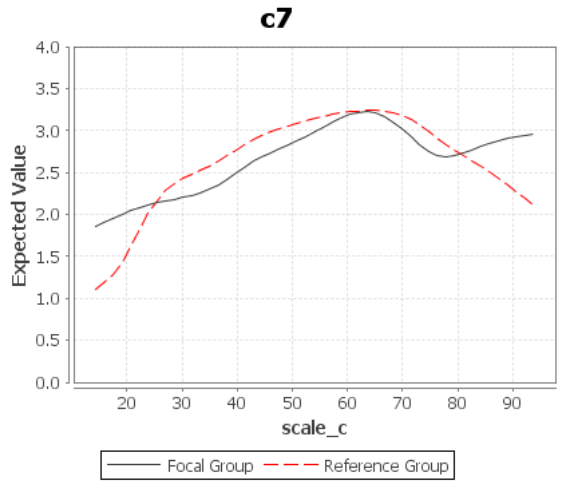


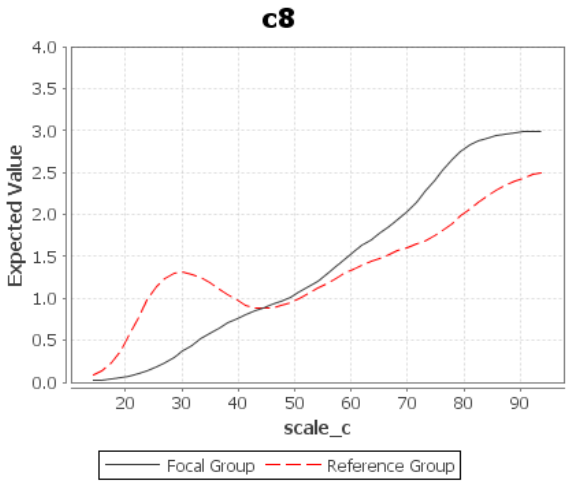


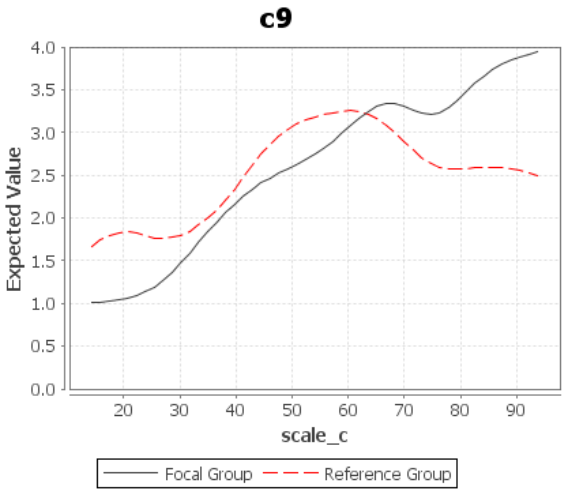


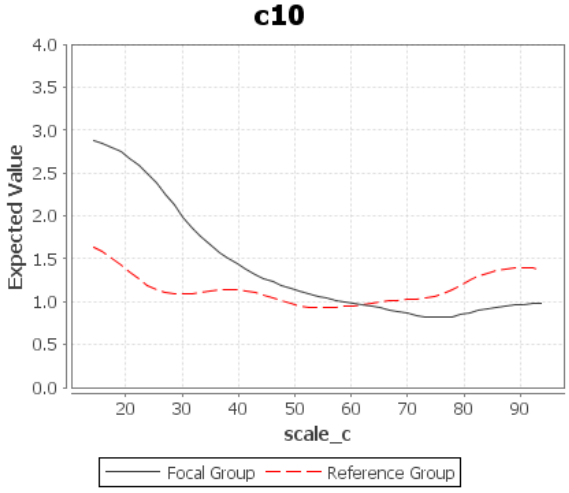


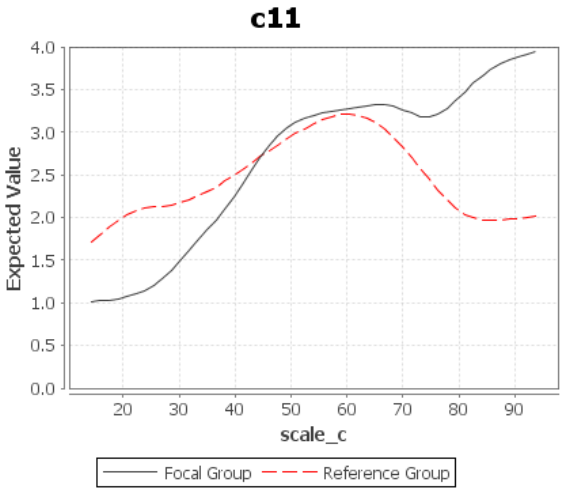


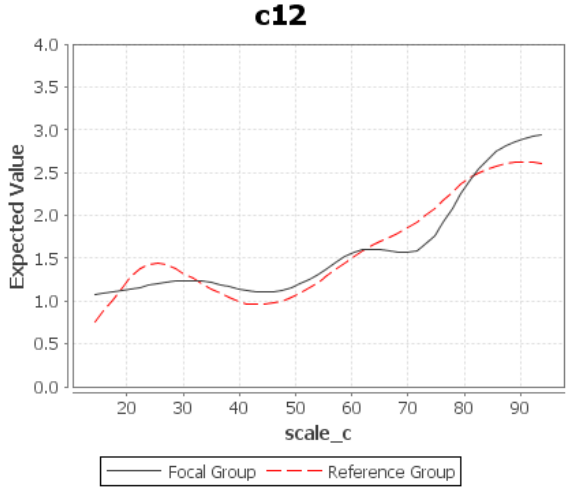

c7
Expected Value
scale_c
Focal Group
Reference Group

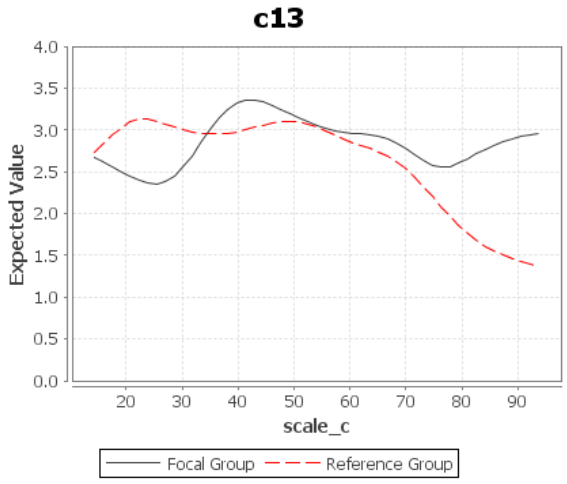

c8
Expected Value
scale_c
Focal Group
Reference Group

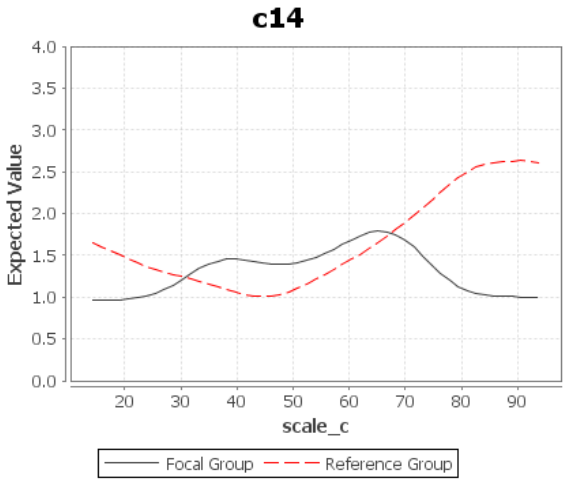

c9
Expected Value
scale_c
Focal Group
Reference Group

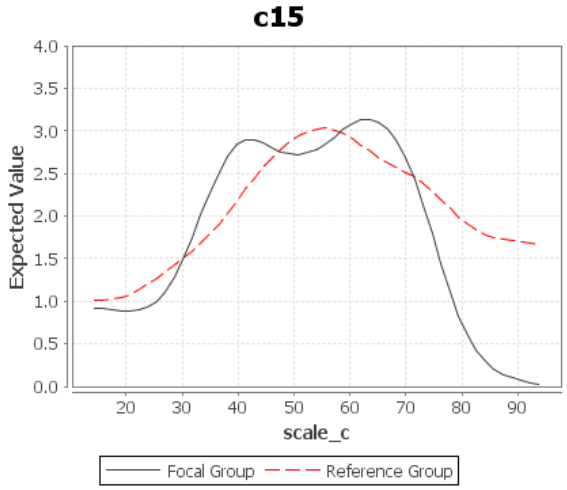

c10
Expected Value
scale_c
Focal Group
Reference Group

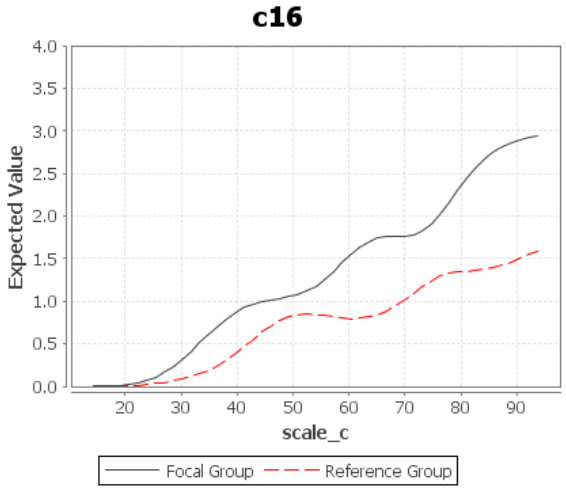

c11
Expected Value
scale_c
Focal Group
Reference Group

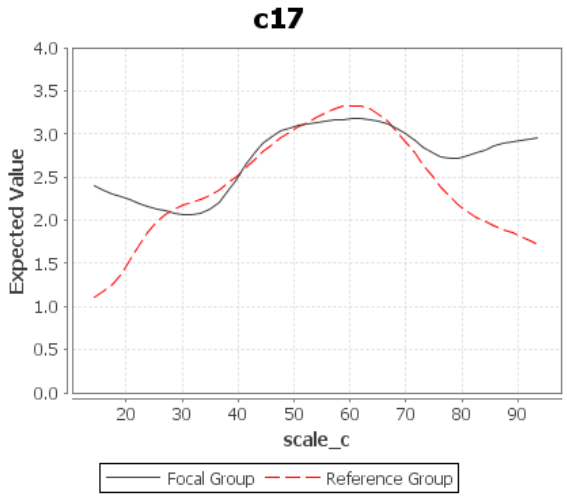

c12
Expected Value
scale_c
Focal Group
Reference Group

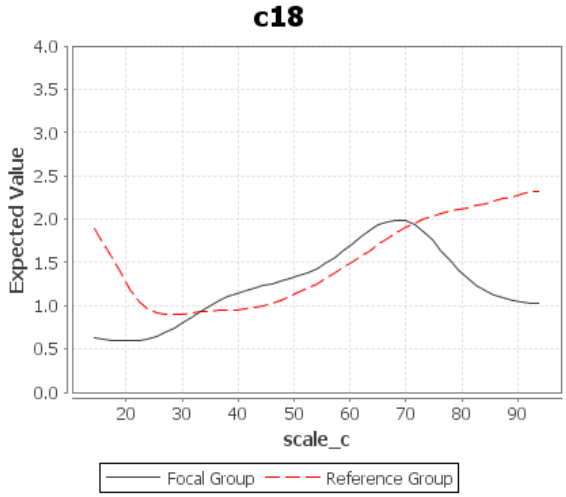

c13
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20 30 40 50 60 70 80 90
scale_c
Focal Group
Reference Group


c14
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20 30 40 50 60 70 80 90
scale_c
Focal Group
Reference Group


c15
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20 30 40 50 60 70 80 90
scale_c
Focal Group
Reference Group


c16
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20 30 40 50 60 70 80 90
scale_c
Focal Group
Reference Group


c17
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20 30 40 50 60 70 80 90
scale_c
Focal Group
Reference Group


c18
Expected Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20 30 40 50 60 70 80 90
scale_c
Focal Group
Reference Group

Appendix R

Inter-item correlation for all items

Inter-item correlation for sign A

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
2	0.42	1.00																	
3	0.59	0.38	1.00																
4	0.43	0.46	0.53	1.00															
5	0.53	0.38	0.47	0.43	1.00														
6	0.31	0.45	0.29	0.31	0.33	1.00													
7	0.23	0.13	0.41	0.25	0.50	0.17	1.00												
8	0.35	0.30	0.31	0.28	0.30	0.29	0.18	1.00											
9	0.46	0.49	0.32	0.26	0.43	0.34	0.07	0.25	1.00										
10	0.00	-0.07	-0.01	0.18	-0.15	0.06	-0.04	0.03	-0.10	1.00									
11	0.50	0.42	0.34	0.25	0.54	0.24	0.27	0.34	0.42	-0.22	1.00								
12	-0.03	-0.06	-0.02	0.18	-0.03	-0.01	0.07	-0.03	0.10	0.22	0.13	1.00							
13	0.01	-0.01	-0.02	0.16	-0.02	0.07	-0.03	-0.15	0.17	0.13	0.20	0.44	1.00						
14	0.06	0.14	0.13	0.29	0.08	0.19	0.16	0.33	0.03	0.14	-0.01	0.41	0.14	1.00					
15	0.57	0.32	0.61	0.60	0.62	0.30	0.33	0.47	0.49	0.07	0.45	0.17	0.18	0.16	1.00				
16	0.17	0.19	0.15	0.24	0.40	0.32	0.30	0.24	0.11	0.01	0.15	0.13	0.06	0.32	0.33	1.00			
17	0.51	0.47	0.48	0.41	0.45	0.38	0.28	0.28	0.38	-0.09	0.49	0.12	0.11	0.27	0.47	0.30	1.00		
18	0.22	0.26	0.20	0.32	0.24	0.33	0.21	0.22	0.32	0.16	0.11	0.38	0.14	0.53	0.37	0.35	0.30	1.00	
19	0.41	0.22	0.29	0.30	0.37	0.36	0.22	0.28	0.44	0.07	0.44	0.20	0.19	0.19	0.45	0.13	0.52	0.27	1.00
20	0.02	0.20	0.07	0.40	0.16	0.27	-0.09	0.20	0.09	0.07	0.17	0.33	0.25	0.24	0.21	0.17	0.20	0.29	0.27

Inter-item correlation for sign B

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
2	0.56	1.00															
3	0.63	0.36	1.00														
4	0.36	0.37	0.18	1.00													
5	0.57	0.31	0.44	0.35	1.00												
6	0.43	0.43	0.27	0.47	0.42	1.00											
7	0.44	0.30	0.47	0.27	0.55	0.40	1.00										
8	0.34	0.18	0.19	0.39	0.34	0.40	0.17	1.00									
9	0.61	0.37	0.48	0.34	0.55	0.42	0.49	0.34	1.00								
10	0.07	-0.14	-0.02	0.10	0.01	0.15	-0.07	0.23	-0.02	1.00							
11	0.53	0.49	0.52	0.34	0.51	0.46	0.54	0.21	0.68	-0.04	1.00						
12	0.17	0.16	0.13	0.14	0.31	0.21	0.21	0.34	0.13	0.08	0.28	1.00					
13	0.15	-0.09	0.22	0.03	0.21	0.12	0.23	-0.01	0.28	0.08	0.34	0.30	1.00				
14	0.28	0.33	0.35	0.27	0.28	0.22	0.07	0.11	0.17	-0.06	0.39	0.25	0.05	1.00			
15	0.69	0.45	0.63	0.35	0.47	0.38	0.41	0.23	0.58	-0.03	0.52	0.22	0.30	0.29	1.00		
16	0.44	0.17	0.41	0.33	0.41	0.35	0.34	0.19	0.37	0.14	0.42	0.18	0.37	0.20	0.50	1.00	
17	0.44	0.30	0.34	0.20	0.49	0.45	0.50	0.23	0.52	0.10	0.50	0.42	0.25	0.25	0.58	0.37	1.00
18	0.14	0.19	0.29	0.05	0.19	0.21	-0.03	0.15	0.10	0.19	0.16	0.37	0.06	0.40	0.21	0.22	0.26

Inter-item correlation for sign C																	
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**
2	-0.56	1.00															
3	0.65	-0.44	1.00														
4	-0.38	0.46	-0.29	1.00													
5	0.13	-0.08	0.20	-0.10	1.00												
6	-0.22	0.30	-0.18	0.54	-0.02	1.00											
7	0.26	-0.01	0.25	0.03	0.08	0.05	1.00										
8	-0.24	0.28	-0.17	.41	-0.07	0.36	-0.09	1.00									
9	0.62	-0.43	0.50	-0.38	0.20	-0.29	0.16	-0.40	1.00								
10	0.28	-0.19	0.26	0.00	0.11	0.12	0.21	-0.04	0.20	1.00							
11	0.55	-0.37	0.43	-0.33	.17	-0.18	0.21	-0.30	0.65	0.26	1.00						
12	-0.04	0.13	-0.07	0.24	0.02	0.29	0.05	0.24	-0.10	0.14	-0.17	1.00					
13	0.19	-0.13	0.02	-0.18	0.04	-0.05	0.18	-0.15	0.24	0.04	0.38	-0.31	1.00				
14	-0.04	0.22	-0.09	0.29	0.03	0.36	0.16	0.29	-0.18	0.17	-0.10	0.47	-0.20	1.00			
15	0.68	-0.48	0.49	-0.48	0.09	-0.20	0.47	-0.33	0.67	0.19	0.58	-0.30	0.33	-0.19	1.00		
16	-0.27	0.27	-0.24	0.36	-0.08	0.21	-0.27	0.21	-0.27	0.02	-0.31	0.21	0.02	0.27	-0.26	1.00	
17	0.54	-0.33	0.52	-0.30	0.18	-0.10	0.30	-0.24	0.50	0.26	0.52	-0.14	0.37	-0.09	0.57	-0.14	1.00
18	-0.08	0.23	-0.20	0.39	0.02	0.31	0.12	0.26	-0.22	0.00	-0.26	0.30	-0.15	0.46	-0.16	0.20	-0.18