Measuring affective responses to vehicle interior textures using Paired Comparisons

Farzilnizam Bin Ahmad

Submitted in accordance with the requirements for the degree of Doctor of Philosophy

The University of Leeds

School of Mechanical Engineering

December 2018

The candidate confirms that the work submitted is his own, except where work which has formed part of jointly-authored publications has been included. The contribution of the candidate and the other author to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

The candidate states to be the primary author and be directly responsible for authoring the work contained within the publications below, which formed the basis for part of the chapters of the thesis as follows.

- Ahmad, F., Holt, R. and Henson, B. (2018). *Measuring Affective Responses To Confectionaries using Paired Comparisons*. Go Green with Emotion. 7th International Conference on Kansei Engineering & Emotion Research 2018, 19-22 March 2018, Kuching, Malaysia. *Linköping Electronic Conference Proceedings* 978-91-7685-314-6, Issue 146, Article 027 pg. 245-254. Publisher Linköping University Electronic Press, Linköpings Universitet.

- Ahmad, F., Holt, R. and Henson, B. (2017). *Measuring Affective Response of Confectionaries using Pair Comparison Method in Rasch Model*. 3rd International Symposium on Affective Science and Engineering (ISASE 2017), 21 – 22nd March 2017, Kogakuin University, Shinjuku, Tokyo, Japan.

Acknowledgements

This thesis has been carried out with tremendous support from my supervisors Dr Brian Henson and Dr Raymond Holt. I am so grateful to have a wonderful supervisors like you both.

I want to express my immense gratitude especially to Dr Brian Henson, who accepting my initial proposal and supervised me till the end of study. He always nourished me with his invaluable wisdom in academics and positive morale support during my difficulties. I am proud to be his student.

I want to express my gratitude to Dr Justin Francis Gallagher from the Institute Design, Robotic and Optimisation (iDRO) who supported me in developing the algorithm and writing the program for the two pair comparisons studies. I am also indebted to Dr Mike Horton from the Psychometric Laboratory for Health Science, who introduce and develop my initial skills in the Probabilistic Rasch Model Software. To my Rasch colleague, Dr Fabio Ribeiro de Camargo, Dr Syakima Ibrahim and Ana Dwiriana Setiati thank you for unwavering supports. Thanks to Dr Thomas Magill (IDRO) for helping me to sort out the theses formatting.

To my friends, colleagues, member of staff iDRO and all volunteers who contributed and supported my works, from initial works to final analysis, thank you very much for helping me.

To my wonderful family, especially to my children Faheen Nur Qaireena, Muhammad Falique Eirfan, Faheen Nur Tihanni, my father Ahmad Sulong, my mother Jamaliah Ibrahim and my siblings thank you for extraordinary support.

A special thanks to my sponsor, the Council of Trust for The People (Mara), Ministry of Rural Development, The Government of Malaysia, for sponsoring my doctoral study at The University of Leeds, UK.

Last but not least to my internal and external examiners, Prof. Dr David Barton and Dr Simon Schutte, thank you so much for accepting and examine my theses.

Abstract

The design trend in the automotive industry is currently moving towards exploring more innovative ways of redefining the craftsmanship quality of the vehicle interior trim through perceived quality. Affective engineering has been widely used as a robust method for understanding people' affective responses to vehicle quality attributes, also being aesthetically and physically pleasing.

However, the research has identified the Semantic Differential scale and Likert scale are subject to biases and errors in assessing perceived quality attributes, resulting in non-linear measurements which ends up with poor reliability outcomes.

In this study, the affective engineering approach has introduced pair comparison technique in order to measure valid and reliable participants' affective responses using the multivariate statistical analyses of the Rasch model — with the objective to establish the linear correlation between participants' affective responses to physical of multisensory cues of touch, vision and feeling of interior vehicle textures.

In this research, the use of Rasch analysis of paired comparisons of products to derive a linear measurement of affective response is tested. Seven pieces of interior vehicle textures and nine unidimensionally fit statements to measure the dimension of perceived quality attributes. A computer-based self-report system presented one hundred and sixty-nine participants with pictures of pairs of stimuli and the evaluative statements in all combinations, and the participants were asked to indicate which stimuli satisfied the statement best.

The analysis demonstrates the viability of using Rasch analysis to obtain measures of affective response from paired comparisons that participants find the choice faster and easier to make paired comparisons compared with evaluating products separately against the Likert scale. It has improved biases and error where the participants no longer make difficult judgements but that in this case, the fit of the data to the Rasch model is very poor.

Keywords: Pair comparison, Affective Engineering and Rasch Model Theory

Table of Contents

Chapter 1 1
1.1 INTRODUCTION 1
1.1.1 Background – What is PQ 3
1.1.2 Way PQ can be measure 3
1.1.3 Motivation of this study 6
1.2 PROBLEM STATEMENTS 7
1.2.1 PQ in New Product Development 8
1.2.2 Promoting AE in accessing PQ in vehicle development 10
1.2.3 Common issues in the self-reported assessment 12
1.2.4 The advantages of PC 14
1.3 RESEARCH AIM 15
1.3.1 Specific objectives: Test of research hypothesis 15
1.4 STRUCTURE OF THESIS 16
1.4.1 Thesis Structure – Diagram 18
Chapter 2 19
2.1 AFFECTIVE ENGINEERING 19
2.1.1 Measuring Affective Responses 19
2.1.2 Common Issue in Measuring Affective Responses 20
2.1.3 Indiscriminating Promote Bias 22
2.2 BIAS AND IMPACT TO MEASUREMENT STRUCTURE 23
2.2.1 Sampling Bias 24
2.3 COMMON ISSUE IN QUESTIONNAIRE DESIGN 26
2.3.1 Items 26
2.3.2 Scales 27
2.3.3 Stimuli 30
2.3.4 Impact of indiscriminating 31

2.4 SOLUTIONS 32
2.4.1 Visual Analogue Scales 33
2.4.2 Rasch Model Theory 36
2.4.3 Rationale the choice of Rasch Model in AE study....................37
2.5 PAIR COMPARISONS 38
2.5.1 Thurstone' PC 40
2.5.2 Bradley-Terry-Luce Model 41
2.5.3 Rank order using paired comparison 44
2.6 The Rasch Dichotomous model 45

Chapter 3 47
3.1 INTRODUCTION 47
3.2 HYPOTHESIS 48
3.3 OBJECTIVES 48
3.4 METHODS 49
3.4.1 Affective Engineering method – Focus Group 49
3.4.2 Generating Affective words 50
3.4.2.1 User experience theory 50
3.4.2.2 Affective Engineering Approach 50
3.4.2.3 Affinity Diagram 51
3.4.3 Data Collections 53
3.5 ETHICAL APPROVAL, CONSENT FORM AND RISK ASSESSMENT 53
3.6 APPARATUS AND EXPERIMENT SETTING 54
3.6.1 The rationale in choosing automotive textures as stimuli 54
3.6.2 Relationship of stimuli to the focus group study 55
3.6.3 Data administration 56
3.6.3.1 Building the SDS items 56

3.6.3.2 Building the LS1 Scale Items 58

3.6.3.3 Data analysis.. 60

3.7 RESULTS OF SDS1 STUDY 61

3.7.1 PCA Analysis 61

3.7.2 Preliminary result - Summary of statistics for SDS1 study 64

3.7.3 Power of fit statistics for SDS1 study 65

3.7.3.1 Reliability – Internal consistency 65

3.7.3.2 Item fit location and fit residuals 66

3.7.3.3 Person fit location and fit residuals.................................... 66

3.8 BIAS IN SDS1 STUDY 67

3.8.1 Lack of invariance 67

3.8.2 Lack of expected probability targeting 67

3.8.3 The high fit residual of person-items 68

3.8.4 A large number of items holds disordered thresholds 69

3.8.5 Items remain disordered after adjustments 72

3.9 RESULT OF LS1 STUDY.. 73

3.9.1 PCA analysis.. 73

3.9.2 Preliminary result - Summary of statistics for LS study 76

3.9.3 Power of fit statistics 77

3.9.3.1 Reliability – internal consistency 77

3.9.3.2 Item fit and fit residuals 78

3.9.3.3 Person fit and fit residuals.. 78

3.10 BIAS IN LS STUDY 79

3.10.1 Lack of invariance 79

3.10.2 Lack of expected probability targeting 79

3.10.3 The high fit residual of person-items 80

3.10.4 A large number of items holds disordered thresholds 81

3.10.5 Items remain disordered after adjustments 82

3.11 DISCUSSIONS 84

3.11.1 Delimitation of the studies 84

3.11.2 The absence of a unidimensionality test 84

3.11.3 Discussion of result of SDS1 and LS1 studies 84

3.11.3.1 Lack of invariance 85

3.11.3.2 Lack of the expected probabilistic targeting 85

3.11.3.3 Higher fit residuals 85

3.11.3.4 Large number items hold disordered thresholds 86

3.11.4 Investigating person-item misfit...86

3.11.4.1 Removing extreme persons 87

3.11.4.2 Extreme items 87

3.11.5 Problematic calibrating process 87

3.11.5.1 Suffering from the calibrations process 87

3.11.5.2 Item calibrations – Rescoring...88

3.12 SUMMARY...89

Chapter 4 91

4.1 INTRODUCTION 91

4.2 HYPOTHESIS OF THE STUDY APPROACH..................................93

4.3 OBJECTIVES...94

4.4 METHODS 94

4.5 ETHICAL APPROVAL, CONSENT FORM AND RISK ASSESSMENT 95

4.5.1 Data collections..95

4.6 APPARATUS AND EXPERIMENT SETTING 96

4.6.1 Data administration 97

4.6.1.1 Building the item pool 97
4.6.1.2 Computer-based survey 97
4.6.1.3 Data coding and analysis process.................................... 98
4.7 RESULTS.. 99
4.7.1 Summary of statistics – All paired confectioneries 99
4.7.2 Reliability and internal consistency 100
4.7.3 Item fit location and fit residuals 100
4.7.4 Person fit location and fit residuals.. 101
4.7.5 Fit statistics for facet analysis – All paired confectioneries..... 103
4.7.6 Fit statistics for facet analysis – Individual confectionery 106
4.7.7 Expected probability targeting 106
4.7.8 Disordered Thresholds 111
4.7.9 Calibrating PC1 processes... 112
4.7.10 Differential item functioning (DIF) analysis 112
4.7.10.1 Examining gender bias of DIF by ICC 114
4.7.10.2 DIF by age group 115
4.7.10.3 Examining age bias of DIF by ICC 116
4.7.11 Data collection speed 117
4.7.12 Calibrated statistics 117
4.7.13 Calibrations Results 118
4.7.13.1 Summary of statistics – Removing one Pairwise.......... 118
4.7.13.2 Summary of statistics – Removing one stimulus.......... 118
4.7.13.3 Summary of statistics – Individual Confectioneries 119
4.7.13.4 PairWise© software results 119
4.7.14 Linear correlation between scales....................................... 125
4.8 DISCUSSIONS 129

4.8.1 Stimuli bias 129

4.8.2 PC1 bias 131

4.8.2.1 The advantages of PC 132

4.8.3 Test of local dependencies 134

4.8.4 Delimitation - Test of unidimensionality.......................................134

4.9 SUMMARY..135

Chapter 5 136

5.1 INTRODUCTION 137

5.2 HYPOTHESIS OF THE STUDY..138

5.3 OBJECTIVES...138

5.4 METHODS..139

5.4.1 Unidimensionality approaches 139

5.4.2 Unidimensionality t-test procedure and process flow.............139

5.4.3 Dual structure results 142

5.5 ETHICAL APPROVAL, CONSENT FORM AND RISK ASSESSMENT...143

5.5.1 Data collections...143

5.6 APPARATUS AND EXPERIMENT SETTING 144

5.6.1 LS2 Items..144

5.7 FACET ANALYSIS RESULTS 147

5.7.1 Summary of fit statistics 147

5.7.1.1 Reliability – Internal consistency 147

5.7.1.2 Item location and fit residuals 148

5.7.1.3 Person-location and fit residuals..148

5.7.1.4 Facet analysis – By location order 148

5.7.2 Calibrated statistics..149

5.7.2.1 Summary of calibrated fit statistics..149

5.7.2.2 Reliability – Internal consistency 149

5.7.2.3 Item location and fit residuals.. 149

5.7.2.4 Person-location and fit residuals 150

5.7.3 Statistical comparison between LS1 and LS2 scales 150

5.7.3.1 Better Chi-square value 151

5.7.3.2 Better person-item separation 151

5.7.3.3 Good reliability 151

5.7.3.4 Better fit residuals 152

5.7.3.5 Individual stimulus logit illustrates high correlation logit . 152

5.8 NON-FACET ANALYSIS RESULTS 159

5.8.1 Summary of statistics 159

5.8.1.1 Reliability – Internal consistency 159

5.8.1.2 Item fit and fit residuals 160

5.8.1.3 Person fit and fit residuals... 161

5.8.2 Response dependencies... 161

5.8.3 Unidimensionality results.. 162

5.8.3.1 First t-test and binomial result 162

5.8.3.2 Second t-test and binomial test 163

5.8.4 Exploring t-test with multiple trials 164

5.8.5 Unidimensionality accepted items 166

5.8.6 Proposing unidimensionality items for PC2 study 168

5.9 DISCUSSIONS 169

5.9.1 Test of local dependencies... 169

5.9.2 Test of unidimensionality.. 169

5.9.3 Test of items fit residual 170

5.10 SUMMARY... 171

Chapter 6 172
6.1 INTRODUCTION 172
6.2 HYPOTHESIS OF THE STUDY..173
6.3 OBJECTIVES...173
6.4 METHODS 174
6.5 ETHICAL APPROVAL, CONSENT AND RISK ASSESMENT 174
6.5.1 Data collections...174
6.6 APPARATUS AND EXPERIMENT SETTING 175
6.6.1 Data administration 176
6.6.1.1 Building the items pool..176
6.6.1.2 Computer-based survey..177
6.6.1.3 Data coding and analysis process 178
6.7 RESULTS 179
6.7.1 Summary of statistics – All paired textures 179
6.7.2 Reliability – Internal consistency 179
6.7.3 Item fit location and fit residuals..179
6.7.4 Person fit location and fit residuals 181
6.7.5 Fit statistics for facet analysis – All paired textures.................182
6.7.6 Fit statistics for facet analysis – Individual textures.................183
6.7.7 Expected probability targeting...185
6.7.8 Disordered thresholds...187
6.7.9 Data collection speed..188
6.7.10 Calibrated statistics...189
6.7.11 Rasch model analysis results..189
6.7.12 PairWise© analysis results 189
6.7.13 Linear correlation between scales...194

6.7.14 Pearson correlation 196

6.7.15 Correlation validation using Matlab 197

6.8 DISCUSSIONS 198

6.8.1 PC2 demonstrates greater discrimination and stability against LS and SDS.. 198

6.8.1.1 Targeting performance of person-items distribution 198

6.8.1.2 Person Ability 200

6.8.2 Delimitation – Bias in PC2.. 201

6.8.3 PC2 excessively discriminates the person-items fit................ 201

6.8.4 Investigating the items and person over- and under-discrimination 202

6.8.5 Source of bias 203

6.9 SUMMARY.. 205

Chapter 7.. 207

7.1 VALIDITY IN AFFECTIVE ENGINEERING 207

7.1.1 Test of local dependencies.. 208

7.1.2 Person-item fit residual 210

7.1.3 The advantages of calibrations procedures in RM 211

7.2 COMPARATIVE ANALYSIS.. 213

7.2.1 Person fit residuals unfit to the model 214

7.2.2 Disordered thresholds 214

7.2.3 Targeting performance 217

7.2.4 Person ability in PC.. 220

7.2.5 Establishing linear correlations and validity........................... 222

7.2.6 Data collection speed 224

7.3 CORRELATION COEFFICIENT WITH PHYSICAL STIMULI......... 225

7.3.1 Correlation of surface roughness 225

7.3.2 Correlation between surface glossiness 228

7.3.3 Correlation with surface colour..230

7.3.4 Surface colour matching between stimuli.................................230

7.4 DICRIMINATING BIAS IN PC 232

7.4.1 Bias violates the response targeting 232

7.4.2 PC over-discriminating more than SDS 234

7.4.3 Stimuli bias in vehicle interior textures 237

7.4.4 Sampling bias in vehicle interior texture..................................239

7.4.5 Stimuli bias in confectionery 240

7.5 LIMITATION OF THE STUDIES 241

7.5.1 Insufficient sample size 241

7.5.2 Higher rejection rate due to misfit persons..............................241

7.5.3 DIF analysis 242

7.5.4 Limitation of the RM application 242

7.5.5 Data handling in RM requires expertise 242

7.6 SUMMARY...243

Chapter 8 244

8.1 CONTRIBUTION TO RESEARCH...245

8.1.1 Contribution to the knowledge 246

8.1.2 Impact of research 247

8.1.3 PC offers interval scales being assessed at nominal scales...247

8.1.4 PC adding value to AE 248

8.2 RECOMMENDATION FOR FURTHER RESEARCH.......................250

8.2.1 First PC study 250

8.2.2 Second PC study 251

8.2.3 Data collection methods, processing and analysing: 252

8.2.4 Other possible PC Methods; 253

8.2.5 Prerequisite suggestion for future studies 254

8.3 CONCLUSIONS 255

References 258
Appendix A - Subtest 1 275
Appendix A – Subtest 2 296
Appendix B 303
Appendix C 309
Appendix D 317
Appendix E 321
Appendix F 327

List of Figures

Figure 1.1.1 PQ from visual aesthetics and haptic properties 3
Figure 1.1.2 Objective and subjective measures in AE 4
Figure 1.1.3 Linear correlation between physical properties and data observations.. 5
Figure 1.1.4 Linear correlation between product and services and data observations.. 5
Figure 1.2.1 Aim of the study to improve PQ validity.................................. 13
Figure 1.4.1 Thesis Structure - Diagram 18
Figure 2.2.1 Three major components construct questionnaire design....... 25
Figure 2.3.1 Common scaling technique used in AE studies 28
Figure 2.4.1 Example of category scales, visual analogue scale and PC (reproduce from (Mattacola et al., 1997) 34
Figure 2.4.2 Scaling Techniques... 35
Figure 2.4.3 Statistical approach between classical methods and Rasch Model 38
Figure 2.5.1 Two methods in assessing Pair Comparisons......................... 40
Figure 2.6.1 Probability of a positive response associated with persons' location on the continuum (Camargo, 2013) 46
Figure 3.4.1 Participants were asked to give their feedback based on the scenario given. 51
Figure 3.6.1 Seven stimuli of vehicle interior texture.................................. 54
Figure 3.6.2 Seven-point bipolar SDS1 items 57
Figure 3.6.3 Five-point LS1 scales items 60
Figure 3.7.1 Scatter plot illustrating PCA composition in SDS1 study......... 63
Figure 3.8.1: Calibrated person-item threshold distribution for SDS1 study 68
Figure 3.8.2 Initial analysis of disordered thresholds before calibrations for SDS1 study 71
Figure 3.8.3: Calibrated analysis of ordered thresholds map after calibrations for SDS1 study.. 71
Figure 3.9.1 Scatter plot illustrating PCA composition in LS1 study............ 75
Figure 3.10.1: Calibrated person-item thresholds distribution 80
Figure 3.10.2: Initial analysis of disordered thresholds before adjustment for LS1 study 82
Figure 3.10.3: Calibrated analysis of ordered thresholds map after adjustment for LS1 study 83
Figure 4.6.1 Four items of confectionery.. 96

Figure 4.6.2 Example of the self-report interface using a computer.............98
Figure 4.6.3 Real-time database...99
Figure 4.7.1 Person-item distribution in LS2011 study...............................108
Figure 4.7.2 Person-item distribution in PC1 which all paired of confectioneries were treated as separate facets................................109
Figure 4.7.3 Person-item distribution in PC1 which individual pieces of confectionery were treated as separate facets 109
Figure 4.7.4 Comparison of persons location logits between LS2011 and PC1...110
Figure 4.7.5: Thresholds map in PC1 111
Figure 4.7.6: DIF by Gender 113
Figure 4.7.7: ICC showing plot by gender for the easiest item to endorse.114
Figure 4.7.8: ICC showing plot by gender for the most difficult item to endorse...114
Figure 4.7.9: DIF by age group 115
Figure 4.7.10: ICC showing plot by age group for the easiest item to endorse...116
Figure 4.7.11: ICC showing plot by age group for the most difficult item to endorse...116
Figure 4.7.12: Relationship between PC1 and LS2011 study....................126
Figure 4.7.13: Relationship between RuMM2030® and PairWise©...........126
Figure 4.7.14: Comparison of locations of stimuli between methods 127
Figure 4.7.15: Facets map for the specialness of confectionaries 128
Figure 5.4.1 Unidimensional T-test Process Flow......................................141
Figure 5.4.2 Dual structure results of facet and non-facet analysis............142
Figure 5.6.1 Five-point LS2 items 147
Figure 5.7.1 Strong correlation between LS1 and LS2 study.....................153
Figure 5.7.2 Stimuli Logit between LS1 and LS2 study.............................154
Figure 5.8.1 Unidimensional Assessment prior to Analysis in T-Test 164
Figure 6.6.1: PC2 Self-report assessment computer interface...................177
Figure 6.7.1: Person-item distribution for analysis of current data in which all pairs of textures were treated as separate facets..........................186
Figure 6.7.2: Person-item distribution for analysis of current data in which individual textures were treated as separate facets 186
Figure 6.7.3: PC2 Thresholds map 188
Figure 6.7.4: Weak relationship between PC2 and LS1.............................194
Figure 6.7.5: Weak relationship between PC2 and LS2.............................195

Figure 6.7.6: Weak relationship between PC2 (PairWise©) and LS1........ 195
Figure 6.7.7: Linear R-square correlation in SPSS 197
Figure 6.7.8: Linear R-square correlation in Matlab 197
Figure 6.8.1: Graphical representation of targeting in LS1 study 199
Figure 6.8.2: Graphical representation of targeting in LS2 study 199
Figure 6.8.3: Graphical representation of targeting in PC2 study.............. 199
Figure 6.8.4: Comparison between person location logits LS1 and PC2 .. 200
Figure 6.8.5. Over-discriminating items with high negative value.............. 202
Figure 6.8.6: Under-discriminating items with high positive value 203
Figure 7.2.1: Category scales with disordered thresholds......................... 215
Figure 7.2.2: Category scales with ordered thresholds 215
Figure 7.2.3: Disordered thresholds map before adjustment..................... 216
Figure 7.2.4: Ordered thresholds map after adjustment............................ 216
Figure 7.2.5: Graphical representation of targeting in previous LS2011 ... 218
Figure 7.2.6: Graphical representation of targeting in PC1 218
Figure 7.2.7: Graphical representation of targeting in LS1 study 219
Figure 7.2.8: Graphical representation of targeting in LS2 study 219
Figure 7.2.9: Graphical representation of targeting in PC2 study.............. 219
Figure 7.2.10: Comparison person locations between LS2011 and PC1 (Ahmad et al., 2018) using confectionery... 220
Figure 7.2.11: Comparison of person location logits between LS1 and PC2 using vehicle interior textures... 221
Figure 7.2.12: Strong relationship between PC1 and LS2011using confectionery... 223
Figure 7.2.13: Weak relationship between PC2 and LS2 study 223
Figure 7.3.1: Weak correlation between LS1 logits and surface roughness226
Figure 7.3.2: Weak correlation between LS2 logits and surface roughness226
Figure 7.3.3: Weak correlation between PC2 logits and surface roughness227
Figure 7.3.4: Weak correlation between LS1 logits and surface glossiness229
Figure 7.3.5: Weak correlation between PC2 logits and surface glossiness229
Figure 7.4.1: Over-discriminating items with high negative value in PC2 study 235
Figure 7.4.2: Over-discriminating items with high negative value in PC1 study 235
Figure 7.4.3: Over-discriminating items with high negative value in LS1 study 236

Figure 7.4.4: Over-discriminating items with high negative value in SDS1 study 236

Figure 7.4.5: Monotone stimuli used during pilot test and focus group 237

Figure 8.1.1: PQ attributes on scale measurement....248

Figure 8.2.1: Likert-paired comparison used with vehicle steering wheels 250

Figure 8.2.2: Five seat sample wrap with different fabrics 252

Figure 8.2.3: PC-CAT software (the graphical representation was reproduced from (Linacre, 2000) 253

List of Tables

Table 2.5.1 Possible rank order sample 45

Table 3.4.1 Twenty-four adjectives for item pool 52

Table 3.6.1 Original pool of items for SDS1 study 57

Table 3.6.2: Coding for seven-point SDS1 scales 57

Table 3.6.3 Original pool of items for LS1 study 59

Table 3.6.4 Coding for five-point LS1 scales 60

Table 3.7.1 Rotated component matrix for SDS1 study according to EPA dimensions 62

Table 3.7.2 PCA were plotted based on combined product matrix.............. 63

Table 3.7.3 Preliminary facet design by location order 64

Table 3.7.4 Summary of fit-statistics for the preliminary analysis for SDS1 study 65

Table 3.7.5 Summary of fit-statistics for the calibrated analysis SDS1 study66

Table 3.8.1 Persons with positive high fit residual 69

Table 3.8.2 Persons with negative high fit residual 69

Table 3.8.3 Items hold disordered thresholds after adjustment in SDS1 study 70

Table 3.9.1 Rotated component matrix according to EPA dimensions 74

Table 3.9.2 PCA were plotted based on combined product matrix.............. 75

Table 3.9.3 Preliminary facet design by location order 76

Table 3.9.4 Summary of fit-statistics for the preliminary analysis 77

Table 3.9.5 Summary of fit-statistics for the calibrated analysis 78

Table 3.10.1 Persons with positive high fit residual 80

Table 3.10.2 Person with negative high fit residual 80

Table 3.10.3 Items holding disordered thresholds after adjustment in LS1 study 81

Table 4.6.1 Pool of Pairwise stimuli 96

Table 4.6.2 Pool of items in PC1 study 97

Table 4.7.1: Items with high fit residuals 101

Table 4.7.2 Participants holding high fit-residuals 102

Table 4.7.3: Participants identified as outliers hold higher location logits.. 103

Table 4.7.4 Age demographic profiles of person factor 103

Table 4.7.5: Preliminary facet analysis of all paired confectioneries 104

Table 4.7.6 Preliminary facet analysis (Individual confectionery) 105

Table 4.7.7 No statistical difference between genders..............................113

Table 4.7.8 No statistical difference by age group 115

Table 4.7.9 Variable Pairwise Matrix..118

Table 4.7.10: Fit Statistics for all pairwise were treated as a separate facet120

Table 4.7.11: Prelimenary facet design (All Pairwise)...............................120

Table 4.7.12: Fit statistics for the initial and calibrated analysis – Removing pair one...121

Table 4.7.13: Preliminary facet design (Five Pairwise) 121

Table 4.7.14: Fit statistics for the initial and calibrated analysis – Removing one stimulus...122

Table 4.7.15: Preliminary facet design (Three Pairwise)...........................122

Table 4.7.16: Fit Statistics for the preliminary analysis of individual confectioneries were treated as a separate facet 123

Table 4.7.17: Average individual logit RUMM 1 – RUMM 5 123

Table 4.7.18: PairWise© results...124

Table 4.7.19: Summary statistics 124

Table 4.7.20: Comparison mean location of confectionery between methods 125

Table 4.8.1: Comparison between PC and SDS and LS methods 132

Table 5.6.1 Coding for texture samples 144

Table 5.6.2 Original pool of items for LS2 study 146

Table 5.6.3 Linguistic quantifier and coding for five-point LS2..................146

Table 5.7.1 Sufficient sample size for Items and person stability (Linacre, 1994)...151

Table 5.7.2 Stimuli mean location between LS1 and LS2 study 152

Table 5.7.3 Fit Statistics of Facet Analysis for LS2 study (*n*145) 155

Table 5.7.4 Preliminary Facet Analysis LS2 study (*n*145)..........................156

Table 5.7.5 Fit Statistic of Facet Analysis for LS1 study (*n*107).................157

Table 5.7.6 Preliminary Data Facet Analysis for LS1 study (*n*107)............158

Table 5.8.1 Chi-Square results 159

Table 5.8.2 Calibrated PSI and α..160

Table 5.8.3 Item fit Residuals across the seven samples 160

Table 5.8.4 Person Fit Residuals across the Seven Samples 161

Table 5.8.5 Response Dependency Items 162

Table 5.8.6 Summary of calibrated unidimensionality test across texture samples 163

Table 5.8.7 Second T-test Results 165
Table 5.8.8 Items Fit Based on Second Unidimensional Accepted 166
Table 5.8.9 Nine unidimensionality fit items were selected for next studies167
Table 5.8.10 Final Items pool for PC2 study 168
Table 6.6.1: Seven stimuli of interior vehicle 175
Table 6.6.2: Preliminary pool of pairwise stimuli 176
Table 6.6.3: Nine unidimensionality fit items pool for PC2 study 177
Table 6.7.1: Items with positive high fit residuals 180
Table 6.7.2: Items with negative high fit residuals.................................... 181
Table 6.7.3: Participants holding high positive fit-residuals....................... 181
Table 6.7.4: Participants holding high negative fit-residuals...................... 182
Table 6.7.5: Preliminary facet analysis of all paired textures..................... 183
Table 6.7.6 Preliminary facet analysis of individual texture 184
Table 6.7.7: Pairwise matrix of individual textures with mirror images of difficult items 187
Table 6.7.8 : Fit statistics for all PairWise© treated as separate facets..... 190
Table 6.7.9: Calibrated facet design (All Pairwise) 191
Table 6.7.10 : Fit statistics for the preliminary analysis of individual textures treated as separate facets.. 192
Table 6.7.11: Average individual logit (RUMM 1 – RUMM 5).................... 192
Table 6.7.12: PairWise© results – Calibrated analysis (n123) 193
Table 6.7.13: Summary statistics 193
Table 6.7.14: Individual textures logits across methods 196
Table 6.7.15: Pearson correlation between PC2 and LS1 studies 196
Table 6.8.1: Missing datasets were recorded.. 205
Table 7.2.1: Comparisons of preliminary statistics 213
Table 7.2.2 Average speed comparison.. 224
Table 7.3.1: Surface roughness measurement 225
Table 7.3.2: Surface glossiness measurement 228
Table 7.3.3: Surface colour measurement 230
Table 7.3.4: Colour matching between stimuli.. 231
Table 7.4.1: Individual person-item fit comparisons 233
Table 8.3.1 Summary Statistics of Non-Facet Analysis for Texture 1 275
Table 8.3.2 Unidimensionality Test on each item performance of Non-Facet Analysis – Texture 1... 276

Table 8.3.3 Local Dependencies Matrix in Texture 1 – Calibrated.............277
Table 8.3.4 Summary Statistics of Non-Facet Analysis for Texture 2 278
Table 8.3.5 Unidimensionality Test on each item performance of Non-Facet Analysis – Texture 2 279
Table 8.3.6 Local Dependencies Matrix in Texture 2 280
Table 8.3.7 Summary Statistics of Non-Facet Analysis for Texture 3 281
Table 8.3.8 Unidimensionality Test on each item performance of Non-Facet Analysis – Texture 3 282
Table 8.3.9 Local Dependencies Matrix in Texture 3 283
Table 8.3.10 Summary Statistics of Non-Facet Analysis for Texture 4 284
Table 8.3.11 Unidimensionality Test on each item performance of Non-Facet Analysis – Texture 4 285
Table 8.3.12 Local Dependencies Matrix in Texture 4 286
Table 8.3.13 Summary Statistics of Non-Facet Analysis for Texture 5 287
Table 8.3.14 Unidimensionality Test on each item performance of Non-Facet Analysis – Texture 5 288
Table 8.3.15 Local Dependencies Matrix in Texture 5 289
Table 8.3.16 Summary Statistics of Non-Facet Analysis for Texture 6 290
Table 8.3.17 Unidimensionality Test on each item performance of Non-Facet Analysis – Texture 6 291
Table 8.3.18 Local Dependencies Matrix in Texture 6 292
Table 8.3.19 Summary Statistics of Non-Facet Analysis for Texture 7 293
Table 8.3.20 Unidimensionality Test on each item performance of Non-Facet Analysis – Texture 7 294
Table 8.3.21 Local Dependencies Matrix in Texture 7 295
Table 8.3.22: Second Unidimensionality T-test for texture 1......................296
Table 8.3.23: Second Unidimensionality T-test for texture 2......................297
Table 8.3.24: Second Unidimensionality T-test for texture 3......................298
Table 8.3.25: Second Unidimensionality T-test for texture 4......................299
Table 8.3.26: Second Unidimensionality T-test for texture 5......................300
Table 8.3.27: Second Unidimensionality T-test for texture 6......................301
Table 8.3.28: Second Unidimensionality T-test for texture 7......................302

List of Abbreviations

1. Affective Engineering (AE)
2. Analysis of variance (ANOVA)
3. Bradley-Terry-Luce (BTL)
4. Computer Adaptive Testing (CAT)
5. Difference item functioning (DIF)
6. Evaluative, Potency and Activity (EPA)
7. Factor Analysis (FA)
8. Gloss Unit (GU)
9. Item Characteristics Curve (ICC)
10. Kaisei-Meyer-Olkin (KMO)
11. Likert Scale (LS)
12. Likert Scale on previous study in 2011 (LS2011)
13. Likert scales study 1 using vehicle interior texture n107 (LS1)
14. Likert scales study 2 using vehicle interior texture n145 (LS2)
15. New Product Development (NPD)
16. Original Equipment Manufacturer (OEM)
17. Pair comparison (PC)
18. Pair comparison study 1 using confectioneries (PC1)
19. Pair comparison study 2 using vehicle interior texture (PC2)
20. Perceived Quality (PQ)
21. Polypropylene plastic (PP)
22. Principal Components analysis (PCA)
23. Principal Components Analysis 1 (PCA1)
24. Principal Components Analysis 2 (PCA2)
25. Principal Components Analysis 3 (PCA3)
26. Proportion Significant Test (PST)
27. Quality Function Deployment (QFD)
28. Rasch model (RM)
29. Research and Development (R&D)
30. Standard Deviation (SD)
31. Semantic differential scale (SDS)
32. Semantic differential scale study 1 using vehicle interior texture (SDS1)

Chapter 1

Eliciting Consumers' Perceptions of Vehicle Designs

It is understood that affect and emotional cues are powerful elements in evoking consumer reactions. People can be sparked with the stunning car design and perfect-quality craftsmanship at the first impression, while product marketing has used these elements to stimulate a more significant point-of-sales value in order to attract car sales. Demand is increasing in this area, with many original equipment manufacturer (OEM) moving towards expanding research and development (R&D) in greater scope to identify the innovative way in redefining the quality of craftsmanship of the vehicle interior trim through perceived quality (PQ). PQ is seen as fundamental research to gauge a better understanding of the user's perceptions and expectations in different markets across the world. In human factor studies, affective engineering (AE) can be subjectively measured using a self-report questionnaire in order to understand users' attitudes, when designing products. However, this method can be subject to biases and errors, which resulted in the non-linear measurement of affective response. One of the biases and errors is associated with the difficulties of using and understanding the semantic differential scales (SDS) and Likert scales (LS). The analysis demonstrates the viability of using Rasch analysis to obtain measures of affective response from paired comparisons (PC). In this chapter, the application of PC in AE methods is rationally explained.

1.1 INTRODUCTION

Cars are not only used as a means of getting around but as an expression of personality and a measure of self-esteem. People are willing to pay for this innovation when the vehicle is designed to fit their taste (Horatiu, 2010). Technology now moved into a new era from practicality and functionality to emotional (Desmet and Hekkert, 2009). In the last few years, there has been a growing interest in AE and it has become widely recognised to be an essential human factor in the design of vehicles (Burnett and Irune, 2009; Abidin et al., 2014).

Affective elements such as attitude, emotion and feelings are essential elements that influence product preference (Hekkert, 2006). AE is important in vehicle development because it is a mature market, with OEM no longer depending on marketing campaigns, but the products should be real and perceived as high quality in order to capture consumers' hearts and minds. The first impressions need to be 'cherished' by their owners, to be kept for a long time and perhaps enjoy second-hand value, or at least extended satisfaction over the car warranty lifetime (Childs et al., 2006; Nissan Motor, 2015).

Quite recently, considerable attention has been paid to study affective and touch and feel perceptions of the vehicle interior. Camargo and Henson (2015) apply AE in developing the fabric features for vehicles, car interiors (Jindo and Hirasago, 1997), vehicle interior craftsmanship (Myung Hwan Yun et al., 2004), vehicle development of the Mazda Miata (Mitsuo Nagamachi, 2011), lift trucks (Schütte et al., 2005) and ergonomic interior cabins for Volvo trucks (Karlsson et al., 2003; Helander et al., 2013).

Vehicle styling development has resulted in a significant impact on positive sensual changes in consumers' perceptions of vehicle exteriors since the 1980s. The design process has evolved from "form follows function" to "function follows the emotions and perceptions of quality" (Palasek and Goodall, 1990). Pressure from the market segment, competitors and demand from the consumers have changed the way a vehicle is being designed, both exterior and interior, because ultimately multi-sensory contacts in the product have affected the overall purchasing decisions (Burnett and Irune, 2009). Dr Peter Tropschuh, a former Head of the Vehicle Concept at Audi, explains: *"A car's haptics are of fundamental importance; the haptic impressions decisively influence the customer's purchasing decision"*(AG, 2001).

1.1.1 Background – What is PQ

The automaker has acknowledged the importance of PQ to improve the vehicle interior design craftsmanship quality. In general, PQ can be defined as perception over the quality attributes; the aesthetics values of the product; the way it looks, the feel of the material, the tactile or haptic response of controls and good ergonomics that give a certain amount of pleasure and satisfaction for customers. The powerful of PQ able to stimulate spark feelings at the first impression in meeting the standard as expected by the consumers.

What makes the research look interesting is due to the input from a subjective response, particularly now product properties and human’s latent traits can be measured using Affective Engineering (AE). AE has been used to assess vehicles’ PQ and haptic control using scales-questionnaire such as SDS and LS.

Complicated and messy switches

Smart simplicity and High precision

Figure 1.1.1 PQ from visual aesthetics and haptic properties

1.1.2 Way PQ can be measure

In automotive industries, PQ can be measured in two methods; objective and subjective measure. Objective measure often used to assess with physical properties in meeting expectation according to automotive standard. One of the objective to improve quality perceptions, surface and gap dimension, haptic and optics reaction to controls elements during and after the interaction. While subjective measures translating the affective responses over the physical properties and needs into the domain of product design. Subjective measures gauge latent properties using survey design.

Objective Measures

Physical Properties

Static

- Visual Aesthetics
- Geometric, tolerance and gaps control
- Fit and Finished
- Touch and Feel Properties (Haptics Feedback)

Dynamic

- Pulling, pushing, shifting, turning, rotating, moving, sliding, touching and gesturing.

Methods

- Five senses checks i.e. visual optics, sounds, odor, touch and feeling
- QC Tools or Calibrated machine and haptic robots

Subjective Measures

Latent Attributes

Design Survey

- Open Ended – Freedom to answer using the subject's own knowledge.

- Closed ended - Scales Questionnaires Likert Scales, Semantic Differential Scales, MCQ, Visual analogue scale and etc.

Methods

- Approx. 84 evaluation methods has been reported

- 15 are most reliable methods Such as Human Machine Interaction, Kano, PReMo, Conjoint, QFD, VOC, Norman Product Emotion, Affective Engineering and etc.

Figure 1.1.2 Objective and subjective measures in AE

In AE studies, PQ activities often use static and dynamic properties to obtain qualitative measures. The physical properties results will then correlate with survey data observation to observe a linear correlation as shown in

Figure 1.1.3. The higher correlation value determined the success rate of the test instruments or data collection. However, some studies were reported demonstrates a poor data correlation because of lack and absence of valid and reliable methods (Cook et al., 1981; Hinkin et al., 1997) (Wellings et al., 2005 and Abidin et al., 2014). The linear correlation between observed data and physical properties were used to monitor product and services performance such as PQ validation and brand and customer satisfaction index as shown in Figure 1.1.4.

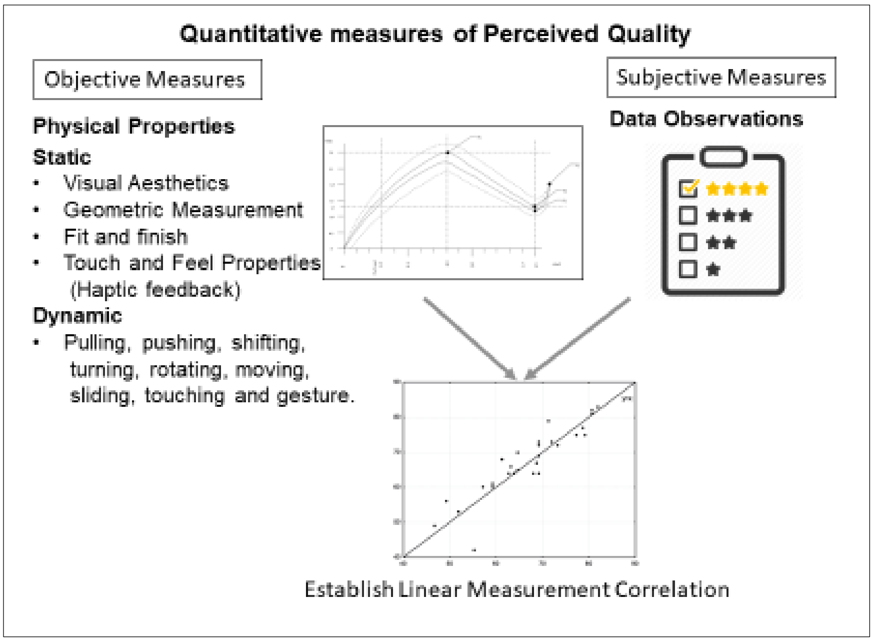


Figure 1.1.3 Linear correlation between physical properties and data observations

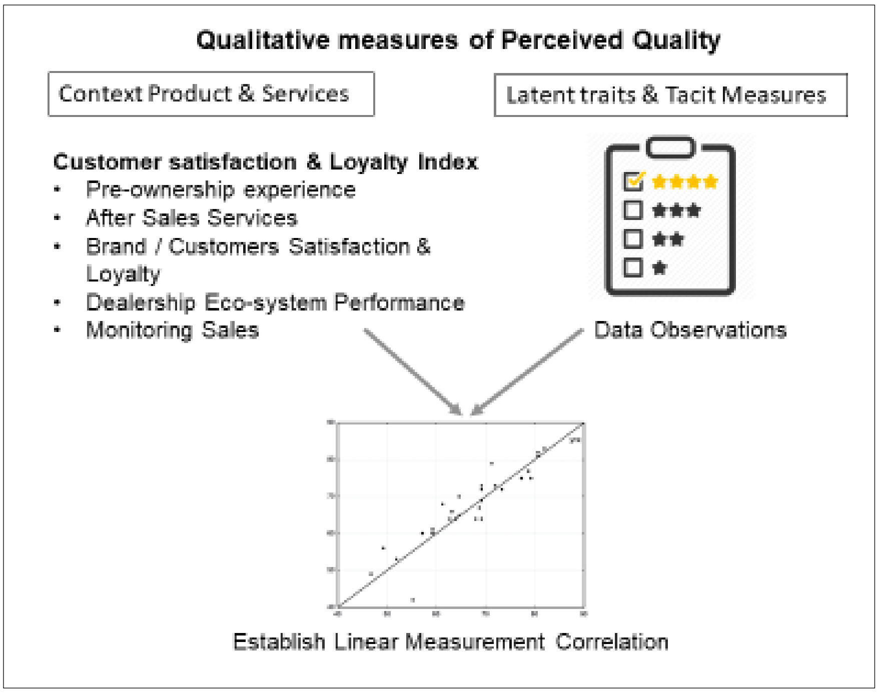


Figure 1.1.4 Linear correlation between product and services and data observations

1.1.3 Motivation of this study

The motivation to start this research due to the absent in design phase in automotive in Research and development which was reported (Wellings et al., 2005 and Abidin et al., 2014). The study was carried out because there are potential techniques can be considered to improve and fulfil the industrial needs and demands.

The existing design guideline or model was insufficient to carry valid and reliable results to support management decisions. Some of the recent studies were reported most design survey of PQ did not focus on the reliability and validity (Cook et al., 1981; Hinkin et al., 1997). The existing model in measuring PQ attributes can be questionable; in a way of how to measure PQ and how well AE these methods provide reliable results (Wellings et al 2015).

AE needs a powerful instrument for quantification and measurement (Schütte and Eklund 2010). The importance of the valid and reliable context is crucial when the industries facing the challenge where an immature decision can be risky when the car needs to be manufactured rapidly to meet very tight lead time and cost-effectively. On other hands, the challenge now automotive decision makers struggle to find an expert, and translating emotions into product solutions is vague and often depends on trial and error (Schütte, 2005).

1.2 PROBLEM STATEMENTS

This research concerns the improvement of the product development process by addressing the quality of the interior trim of production vehicles with the trim through applying AE methodology to texture design development. It has been understood that the aesthetic values of the product - the way it looks, the feel of the material, the tactile of haptic responses of controls and excellent ergonomics - are crucial elements in giving a certain amount of pleasure and satisfaction (Abidin et al., 2014).

AE approach will be used to measure participants's affective responses, including the application of multivariate statistical analyses with the theory of the probabilistic Rasch model (RM). It will investigate the integration of multisensory cues for touch, vision and sound of the texture of trim materials and explore manufacturing trade-offs with a goal of identifying design styling guidelines for vehicles.

Unfortunately, the methodology in assessing affective response is currently absent in the design phase (Abidin et al., 2014). The earlier vehicle development process, the decision-making when selecting design proposals, was executed without concrete evidence and the decision influenced and dominated by soft statistical results, which, it can be argued and lack of reliability. Thus, the decision seems to be weakened because the questionnaire was objectively designed to indicate product preferences rather than testing and validating the questionnaire. Uncalibrated questionnaires have resulted in biases to the design parameters and can be risky.

Therefore, this research is concerned with the improvement of the vehicle interior by proposing a valid and reliable method in assessing and quantifying users' affective responses in order to create a great design. From the automaker's viewpoint, decision-making from a valid and reliable outcome is crucial to support design management in making an accurate and unbiased judgement. However, most of the cases in this context are absent because of the limited number of reliable methods.

1.2.1 PQ in New Product Development

PQ can be defined as users' perceptions or first impressions of the overall quality of a product, service or a brand's ability to fulfil his or her expectations (Antariksa, 2018). PQ in the automotive context refers to the quality that customers acknowledge via the look, the touch and the feel of a car (Nissan Motor, 2015).

In the current economic situation, consumer demand and competitor pressure have resulted in OEM's needing to design and manufacture new vehicles rapidly to remain relevant in the market. A decade ago, the NPD was usually developed within twenty to twenty-four months. However, the trend shrank recently with some of the models developed under sixteen to fourteen months (Winter, 2003).

One alternative to help OEMs meet the need for rapid development is using platform sharing. Today, the giant OEMs from Japan, Germany, the US and Korea are producing new vehicles every year based on shared architecture platforms which use a similar chassis and drivetrain. To give an example, the Audi A3 uses the same chassis platform (PQ34-35) as the Seat Leon, Skoda Superb and Volkswagen Golf, Jetta, Beetle, Bora, and Touran (Nissan and Bernhard, 2006). In a total of 10,126,281 units has been manufactured by the Volkswagen group and this has resulted in this group hold the second largest car manufactured in the world behind the leading Toyota and Hyundai third in 2016.[1]

Similar fashions influenced the designs for OEM in Malaysia, where Proton Holdings Berhad (Proton), Perodua, Honda, Toyota, Nissan, and Mitsubishi have been producing vehicles on a sharing-based platform. To give an example, Proton's Inspira was manufactured under license from the Mitsubishi Lancer GS in Japan, while at same time Proton used their platform from the Proton Preve to develop the Suprema S. This trend rapidly increases every year.

---

[1] International Organization of Motor Vehicle Manufacturers (OICA) – List of manufacturer by motor vehicle production in 2016 documented from various United Nations expert groups

The increase in platform-sharing demand for producing new model developments every year means OEMs have had to face other challenges and complexities in meeting tight R&D lead times. Engineers and designers need to double up the amount of workload in designing, testing and validating each component for a different brand, variance, or various market segments, such as right-hand and left-hand drive, as well niche markets such as special or limited editions that require customised design, colours and materials.

To facilitate the amount of work, most of the NPD uses digital technologies to replacing conventional methods: for instance, robust virtual reality can be used as a simulation in creating various design proposals. The use of technologies will shorten the development time and eliminate unnecessary cost in developing prototypes.

In styling related works, automakers are usually engaged with independent research consultants to work with designers and engineers facilitating the front-end design and engineering studies to enhance customer experience. They are taking tacit knowledge and turn into explicit knowledge, although the back-up from consultants is limited and seldom offers a powerful method (Larson, 1999; Schutte, 2005; Rosenthal and Capper, 2006). The works, such as conducting series of design surveys to understand and determine the specific requirements to match with the design proposals, can be time-consuming and expensive if they need to be done by in-house R&D. But the consultant is needed because they have access to the global database, and sometimes it is important to obtain the information to understand some demographic and cultural differences for various markets and target users.

Apart from common design survey, the OEM’s currently use design survey platforms to gain some useful information about user insight, experiences and perception of quality craftsmanship of the vehicle themed as the PQ. The PQ survey has been classified as a different level of a survey which requires psychophysics and psychology to measure the user’s latent traits and requires reliable methods. The information will be used to improve the weakness of existing products and to set standard guidelines for the next NPD.

1.2.2 Promoting AE in accessing PQ in vehicle development

A PQ study can be the most powerful and effective method in obtaining a wide range of information about the product positioning and benchmarking, derived from the first-hand information by the users. Sensory evaluation study such as AE currently exist is widely used as a reliable methods in assessing PQ in vehicle development (F. Camargo and Henson, 2015).

Marek Reichman, a former Ford's director of interior strategy, stated in July 2004, that *"interiors are important for the next battleground affective to the vehicle sales"*(Wellings et al., 2005). A former Proton CEO's Syed Zainal Abidin Syed Tahir said, "*Every car was designed with a great quality target, and we can meet almost ninety-five percent, but we need five percent of sensual design through perceived quality to buy the customer's soul."* (Hassan and Samsudin, 2012). Wellings et al. (2015) state that the priority in vehicle design now is shifting from exterior appeal to interior emotional appeal; attention is turning to ways to satisfy all the senses.

However, PQ has not been commonly applied within the framework of the NPD process as standard design guidelines for vehicle development, although it was seen as important (Wellings et al., 2005). Automotive decision makers are well aware of this but they struggle with problems to find an expert, and translating emotions into product solutions is vague and often depends on trial and error (Schutte, 2005).

In the bigger scales, AE was used to assess the PQ attributes of a vehicle when in 1987 Nagamachi has applied AE in developing the interiors and exteriors for the production car the Mazda Miata or MX-5. In heavy industry application, the Komatsu PC50 Truck was developed using AE (Mitsuo Nagamachi, 2011) and Simon (2005) used AE when designing the electric forklift, the Toyota-BT Reflex.

In small scales, the application of AE is more specific to developing the small components or parts, such as the fabric features for the seats (Jindo and Hirasago, 1997; F. Camargo and Henson, 2015), seating comfort (Kamijo et al., 1982), steering wheel (Ajovalasit et al., 2013), interior textures (Kikuta et al., 2008; Loss and Jansen, 2015) and audio products (Lu and Petiot, 2014), To promote AE to be accepted as part of standard design guidelines, AE needs to become equipped with valid and reliable methods in accessing the user's PQ, which is currently absent in vehicle development (Wellings et al., 2005). The methods should aid designers to design a better product, not be seen as anti-designer or as a stifling process that disconnects designers' creativity and innovation (Wellings et al., 2005).

The trend in automotive styling is moving towards accepting AE as a design method in designing vehicles. As an example, Kansei Design was used as a method for design development in Toyota Motor Europe (Levy, 2013). More automotive manufacturers have formed dedicated PQ teams who are responsible for measuring perception of quality, including at Jaguar Land Rover (Claudia Newton, 2017). Claudia (2017) stated that PQ needs to deliver two crucial aspects: identifying the influence on the participants during vehicle evaluation; and methods of measuring and assessing the critical vehicle attributes that must have impact on the purchase decision.

Bhise (2012) states that PQ can be measured using objective measurement and subjective evaluation (Bhise, 2012). Okamoto et al. (2013) states the perception of the quality and properties of a material surface by touch comprises two layers: psychophysical to measure surface roughness, and affective to evaluate richness (Okamoto et al., 2013; Claudia Newton, 2017). Wellings et al. (2015) raised the question, how can the information be measured using the best methods as automakers are now ready to use AE, and how well can AE measurement provide reliable results to inflate strong design proposals when dealing with high investment projects, such as in the automotive industry.

1.2.3 Common issues in the self-reported assessment

Schutte and Eklund (2010) emphasised that a major criticism from industries users is that AE needs a powerful instrument for quantification and measurement as well as expertise in statistics and mechanical design. However, most of the AE design survey did not focus on the reliability and validity, which has led to difficulties in interpreting the results (Cook et al., 1981; Hinkin et al., 1997). Also, measures commonly used in the industries have been reported to have psychometric problems (Carman, 1990). Some AE studies provide unstable results in different samples, and the procedures cannot show with certainty that the response is at least on interval scales (Camargo, 2013).

One of the reasons for the item and stimuli being predetermined by the OEMs is to obtain specific information. Either the item was newly developed or was carried over from previous studies or common items. However, most of the assessor do understand and prioritise that the impact of biases affect their results: for example, the number of items being used is kept as low as possible in minimising response bias and that inflates the boredom and fatigue (Hinkin et al., 1997). Also, impact bias from a non-random sampling population is often associated with socially desirable responses, which should be avoided (Paulhus, 1984; Böckenholt and Dillon, 1997; Cattelan, 2012) or a demographic based on the geographic region (Delgado-Rodriguez, 2004).

However, OEMs did not specify which scaling is the best to obtain the most reliable results. A well-established framework to guide researchers through existence scale development methods is lacking; additionally, many researchers might be interested in measurement per se, although they must find a way of studying the importance of items to achieve a statistical fit in a situation where the existing scales are inadequate, inappropriate or unavailable (Hinkin et al., 1997). Some of the researchers understand the impact of the scales of the response are seriously biased (Cliff, 1973; Böckenholt and Dillon, 1997; Cattelan, 2012).

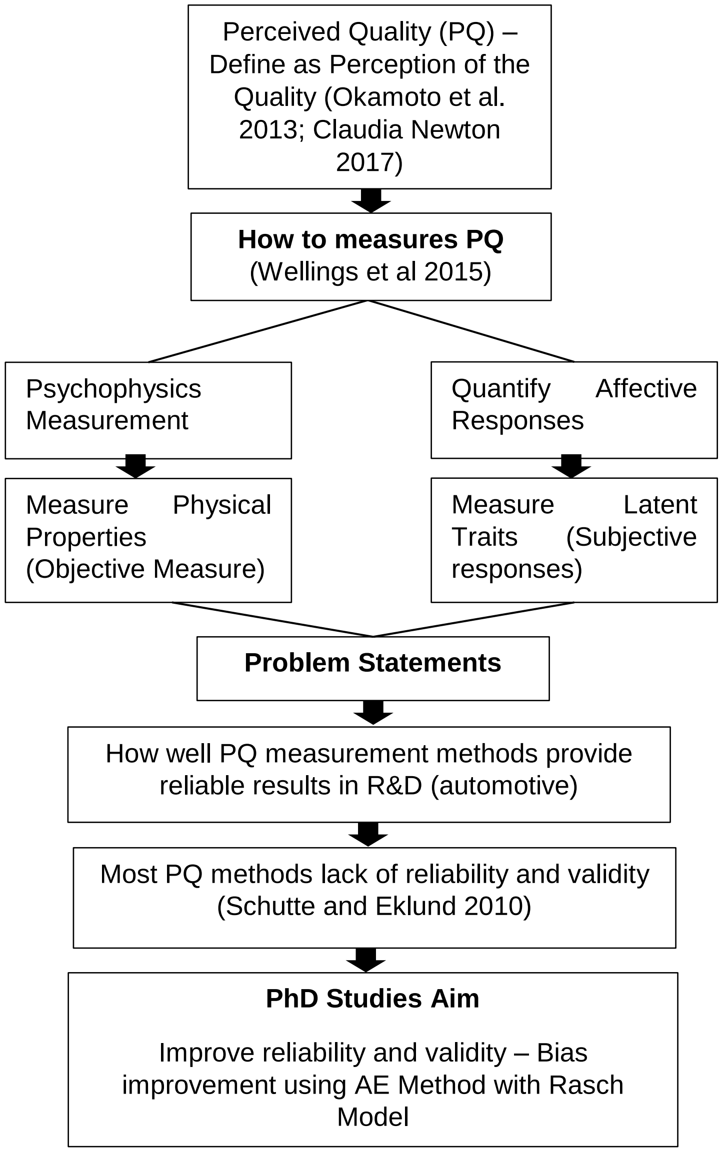


Figure 1.2.1 Aim of the study to improve PQ validity

The accuracy in obtaining information from participants' responses can be questionable and may be exposed to bias and error, which affect poor design in forecasting to support a strong design proposal. This information from immature data can be risky when the car needs to be manufactured rapidly to meet a tight lead time. However, in most cases, items being used rarely undergo appropriate validity tests because of a sense of urgency, confidentiality restrictions and fixed costs.

1.2.4 The advantages of PC

From a psychological perspective, participants find it too easy to discriminate the products along the affective dimension of interest, thus the responses are able to derive linear measurement (Ahmad et al., 2018).

The advantages of PC in statistical outcomes is widely credited by many scholars; PC offers ordinal scales that can be assessed at intervals (Arikan and Gelbal, 2013; Alvin and Adams, 2015). Studies also demonstrate PC is a better method than the LS score (Phelps et al., 2015; Ahmad et al., 2018) and in some situations, PC is more valid than the raw score produced by standard marking (Bramley and Oates, 2011). PC offers high reliability in its statistics output where PC provides the maximum likelihood estimation, calculates a separation index and tests the model fit (Heldsinger and Humphry, 2010)

In the context of data performance, (Choppin, 1968) describes a condition algorithm for estimating the parameters for the dichotomous RM work with the comparison between pairs of items (Garner and Engelhard Jr., 2009). PC offers a maximum likelihood to extract parameters (Garner and Engelhard Jr., 2009). The advantages of using PC includes the ability to avoid unanswerable questions led to missing data (Joseph, 1969) as the data collection can be done using an online platform.

In the context of minimising the bias reduction, PC psychologically offers better readability and faster judgement (Florin, 1999; Ahmad et al., 2018), minimising the burden on participants and enabling instant decisions (Alvin and Adams, 2015). Last but not least, PC can work independently to avoid long waiting times if there is a large group of participants.

1.3 RESEARCH AIM

The aim of this thesis is to assess whether linear measurement of affective responses can be derived from PC. The objective of this study was a structure linked to the thesis structure in Figure 1.4.1.

1.3.1 Specific objectives: Test of research hypothesis

1. **This study aims to examine sources of bias that affect the difficulty of endorsing items in SDS and LS.** Two studies were designed to observe the difficulty outcome from the biases which the extent violate the statistical outcome.
2. **To assess the feasibility of the PC approach by carrying out an AE study using PC for which there is historical data of SDS and LS responses to confectionary**. This study will observe how well the items fit into the model using an SDS and LS scale and how much bias is likely, how successful the calibrations process is and which rating scales best fit in achieving invariance stability.
3. **To develop new items for a PC study of affective responses to the quality of vehicle interiors.** This study examined how unidimensional items affect goodness of fit statistics as expected in RM, then establish new items for a PC study. The study will carry an additional sample size.
4. **Administering an AE study of affective responses to the quality of vehicle interiors using PC.** This study will observe how participants might find it easier and faster to evaluate products using PC, how well PC satisfies the assumption in minimising the effect size of bias and error by determining how much bias in PC is evident and how tedious the calibrations process is to fit the statistics outcome in RM.
5. **To compare and assess SDS, LS and PC approaches to the measurement of affective responses to confectionary and to vehicle interiors.** To determine whether a linear correlation of affective response between PC, SDS and LS can be established. If it can, then PC indicate should not vary within the same context when using SDS and LS scales.

1.4 STRUCTURE OF THESIS

This thesis was divided into eight chapters. These chapters were assigned according to the literature review's focus of area study in AE. Critical evidence was discovered and was addressed by applying methods of PC-Rasch Model. Five studies namely Semantic differential scale study 1 (SDS1), Likert scale study 1 (LS1), Likert scale study 2 (LS2), Pair comparison study 1 (PC1) and Pair comparison study 2 (PC2) were demonstrated within this thesis which synchronises to the thesis framework was illustrated in the framework as shown in Figure 1.4.1.

The first chapter of this thesis is to explain the introduction of AE methods in assessing PQ application in the automotive industries, constructing the logical effect of introducing PC as a solution to overcome the limitations arising from using SDS and LS.

The second chapter of this thesis sets out to learn, understand and define the literature theory of PC while also investigating bias and error from the available sources on the literature and establish the logical structure of the RM theory to the statistical procedures and calibrations process.

The third chapter of this thesis investigates the bias evidence in SDS and LS1 due to biases and errors that affect the difficulty of the task of endorsing the items using interior vehicle textures. It also assesses how well the items fit into the statistical model and whether linear measurement of affective response can be derived using an SDS and LS scale.

The fourth chapter of this thesis tests the hypothesis of PC1 using self-report assessment using confectionery. When making a PC, the participant merely has to indicate which two products they endorse more readily, rather than thinking about which category of responses one would elicit separately. The challenge is then to derive a linear scale of affective responses from such comparisons as to whether participants might find it easier to evaluate products and satisfy the assumptions in minimising the effect size of bias and error.

The fifth chapter of this thesis investigates and determines the unidimensional items for PC2 assessment which are taken from the previous items bank in LS1 study. While retaining the number of items and stimuli, the study performed with additional sample size — the investigation then observing how unfit items are associated with bias and error that could inflate the multidimensional features and corrupt the measurement structure of RM theory.

The sixth chapter demonstrates the statistical stability using PC2 to assess vehicle interior textures deriving the linear measurement of affective responses. This study investigates whether observed data using vehicle interior texture conform to SDS1 and LS2 studies, and offer resemblance logit.

In the seventh chapter, the discussion of the comparative outcomes from the studies is elucidated to explain the logical structure of why comparative scales such as PC perform with better results than non-comparative scale in SDS and LS. On the other hand, it justifies why in certain conditions PC does not work well and produces poor fitting results to the RM.

In Chapter eight the summary and conclusions are drawn from the outcome of the studies.

1.4.1 Thesis Structure – Diagram

Chapter	Focus	Theoretical Approach	Study Approach	Specific Aims
1	Introduction, Aim & Objective	Quantify latent traits in design Human-centered design		Four Objectives; 1 - To determine Bias 2 - Test Pairwise model 3 - Unidimensionality item 4 - Measure Pairwise stability
2	Literature review Measuring AE Response	- Affective Engineering - Inaccuracies evidence in AE - Bias and Error - Rasch Measurement Theory		
3	Determine sources of Biases and error Hypothesis Objectives Methods Results	AE Rasch Model®	**Study 1 - (SDS1)** **Study 2 - (LS1)** using Interior Vehicle Textures	*To determine source of Biases and error *Test Linearity *How well item fit into model using SD & LS
4	Propose PC model Hypothesis Objectives Methods Results	Rasch Model® & PairWise©	**Study 3 - (PC1)** Using Confectionaries	*Test Pairwise Model *Test good fit statistics *Test Bias reduction performance *Linear correlation
5	Determine Unidimensionality items for PC2 Hypothesis Objectives Methods	Rasch Model® & T-test	**Study 4 – (LS2)** Interior Vehicle Textures	*Additional sample size affect statistical fit *Test unidimensionality items *How unfit items corrupt measurement structure of RM
6	Applying PC Hypothesis Hypothesis Objectives Methods	Rasch Model® & PairWise©	**Study 5 - (PC2)** using Interior Vehicle Texture	*Validate the PC stability using Textures *Test Bias reduction performance *Linear correlation
7	Discussion Comparative Analysis	Discussion analysis were explain: 1) Test of fit statistics for Facet & Non-Facet Analysis 2) Different Item functional test (DIF) 3) Unidimensionality, Independent T-Test & Binomial 5) Defining linear measurement using PairWise analysis 6) Comparative results 7) Contribution to knowledge 8) Future works		
8	Conclusion			

Figure 1.4.1 Thesis Structure - Diagram

Chapter 2

Pair Comparisons in Affective Engineering

2.1 AFFECTIVE ENGINEERING

The noun of *"affect"* defined by Henson and Livesey (2006) referring to several psychological states such as emotions, feelings, moods, sentiment and passions; each of which differs in duration and impact. In a similar viewpoint, the word 'affective' as defined by Nagamachi (1995), is the impression that somebody gets from certain artefacts, environment or situation using all sense of vision, hearing, smell and taste as well as cognition (Mitsuo Nagamachi, 2011). Different notion given by AE scholars has complimented the mission and vision of AE to enhanced systematic judgement whether good or bad, safe or dangerous and it makes value judgement looks better (Helander and Khalid, 2014). A similar notion was to define the AE more precise as quantify users' reactions and defines their relationship to physical parameters of design (Barnes and Lillford, 2009). In a similar view, Warell (2001) defined. Affective domain existed in several stages, which refers to receiving, responding, valuing, organising and characterising (Warell, 2001). To summarised all the meaning, *'affective'* can be simply translated as consumers' feelings for a product into design elements (Henson and Livesey, 2006).

2.1.1 Measuring Affective Responses

Many researchers across different research disciplines have now studied the challenge in meeting tomorrow's consumers. AE has been used as marketing tools, such as purchasing insight to capture consumer's attention, pleasurable and desirable; influence consumers in making affectionate purchase decision to the products. Today most consumer products especially automaker has acknowledged the proven AE methods in measuring subjective response and translating into product properties (Nagamachi, 2010). However, AE methods required precision methods to quantify the subjective responses to satisfy the statistical outcome as briefly explained in the first chapter.

2.1.2 Common Issue in Measuring Affective Responses

Survey research remains the most popular source in academic and market research studies because this is a reliable tool to learn about consumer behaviour, preference and perceptions (Dolnicar and Grun, 2013). Most researchers believe that what makes the research look interesting is due to the input from a subjective response, particularly now that humans' latent traits can be measured (Aziz et al., 2013). In AE, the product properties can be evaluated subjectively using the scales-questionnaire. This method is important in measuring PQ towards products and services.

Questionnaires are a popular way of gathering information from participants and responses can be quantified using various sophisticated techniques, and the result presented (Munn and Drever, 2004). The questionnaire is so important because it is the heart of the survey (Krosnick and Presser, 2009) and with the number of traits, abilities and attitudes, psychologist have attempted to design better methods (Snider and Osgood, 1969). This will increase measurement precision at the level of individual ability as well as reduce the burden of the respondent (Bjorner et al., 2003).

However, a potential difficulty for a practitioner-researcher is whether participants are likely, to be frank when responding to the questionnaire as it will affect the reliability of dataset (Munn and Drever, 2004). Therefore to minimise response errors, questionnaires should be designed by best practices (Krosnick and Presser, 2009). These statements are reinforced by (Likert and Likert, 1976) that when completing the questionnaire, it is important that every question needs to be answered thoughtfully and frankly as possible. The questionnaire is not a test, and there are no right or wrong answers.

The most tricky part of the survey, participants were acting the way they see things and reflect the way they act (Likert and Likert, 1976). If the perceptions are distorted, the distortions are reflected in their behaviour, and this may affect missed targeting judgement to distinguish between items, stimuli and rating scales (Camargo, 2013).

Researchers have studied the issue relating to the targeting issue and attempt to recommend based on findings over the year. However, in reality, it is quite difficult to control because in most cases, the participants were asked to respond the survey when the item was not designed for the participants' ability. At the same time, the items were probably make them feel disappointing, annoying, sensitive and meaningless to response than choose to answer but not as what they felt. Sometimes participants refused the answer or just leave it blank. However, the most common case, the unanswered items was happened due to difficulties factor lead to misinterpreting items and unexpected response (Camargo, 2013).

In the early phase of the study, the focus is to learn and understand what is the common ground related to the test questionnaire, especially how it was developed. How well the items, stimuli and the scales were designed according to the participant's discrimination ability and how well the available scaling technique provide equal distance threshold to discriminate effectively.

The literature findings have drawn many robust ideas and technique, from technique to designing the questionnaire, screening the sampling, diagnostic the result of interpreting the analysis. However, one of the most notable studies is emphasising the importance of accuracy in the survey is determine how well the responses to the items in meeting the accuracy in statistics (Tennant et al., 2007; Camargo and Henson, 2011). Good ratting score determined how well the item traits on scaling can be understood and discriminate by the participants as expected by the statistical model (Charles and Suci, 1955).

On another perspective, AE study has reported the accuracy in a slightly different angle with justify the logical underlying behind the poor statistical performance. Recently, many criticisms have been raised amongst AE scientists about the indiscriminating applications in SDS and LS methods which inflates the inaccuracies due to biases and error from grammatical ambiguity or language barrier construct the items, stimuli and number of scaling (Henson and Camargo, 2014). The most notable study reveal the evidence that the use of category scales introduce lack of discriminating and biases among the items because participants understand and use a scales using their subjective unit to judge and interpreting differently (Poulton, 1989). The most common drawback is that some participants have difficulty to understand and use a rating scale (Krosnick and Presser, 2009).

2.1.3 Indiscriminating Promote Bias

The discriminate refers to how well the participants distinguish between the object or in this study refer to category scales based on information given from the items and the stimuli's effectiveness. However, indiscriminating is explained opposite direction, refer to a negative connotation. A critical assumption is one of the major sources introduce some degree of bias and error which the regression would make a fair test seem unfair (Wright et al., 1976; Grider and Malmberg, 2008).

However, some study has reported that indiscriminating is one of the major factors leading to inaccuracy that potentially will impact the overall good fit in statistical analysis in the Rasch Model (Camargo, 2010). The Rasch model generally refers to a family of probabilistic models developed by Danish Mathematician George Rasch (1960) (Andrich, 1988; Henson and Camargo, 2014). Indiscriminating often associates with biases and error where commonly affect the participant' fatigue, disinterest or distraction when the questionnaire is poorly designed can happen at random or without careful judgement (Krosnick and Presser, 2009; Camargo, 2013). While some study has pointed out sources of anomalies in data sets in AE such as redundancy, misrepresentation, misinterpretation, bias and ambiguity that affect poor statistical properties (Heise, 1969).

2.2 BIAS AND IMPACT TO MEASUREMENT STRUCTURE

In general, bias relatively happen when the experiment undergoes incorrect of the association between an exposure and an effect in the target population which statistically introduce low internal validity that does not equal to the true value.

Bias can define in Cambridge Dictionary as the action of supporting or opposing a particular participants or things in an unfair way and allowing personal opinion to influence the judgement. For some scholar, the interpretations is more precise to the context of psychophysics study, where according to Stevens' (1946) dealing with stimuli magnitude states that bias may happen when observers do not know how to use familiar units, and the judgements are made by their own subjective units to discriminate (Poulton, 1989). In a similar view, bias means someone is not accurately evaluating the evidence, and other things are affecting the decisions (Field, 2013). Focalism is an opposite connotation of bias because is a straightforward (Wilson and Gilbert, 2005).

The impact of bias was corrupt the measurement structure and affected the poor interval validity (Delgado-Rodriguez, 2004; Camargo, 2013). Bias can happen in many ways and bias is not only introduced when designing the questionnaire, but it can happen at any stage (Sabin, 2010). Thus it is important to predict the bias beforehand with sufficient understanding about the characteristic of bias in before the questionnaire was designed.

Bias relatively happen when the experiment undergoes with inappropriate of the association between non-random sampling population and the data instruments which tends to be bias known as item bias, scale bias and stimuli bias that affect the rating scales methods and statistical outcome. (Böckenholt and Dillon, 1997; Cattelan, 2012) Have outlined that the response categories scales may seriously bias. Bias potentially affect the good fit in statistical analysis.

2.2.1 Sampling Bias

In survey design, bias is synonym with poor target sampling within the population and to the test instruments which can be exposed to bias (Böckenholt and Dillon, 1997; Delgado-Rodriguez, 2004). In the selection of population, for example, bias could minimise the impact to the accuracy when the recruitment of sampling is taken from an inappropriate population known as Non-random sampling example when the sampling is recruited from voluntary and convenience participants which likely introduce a good chance of getting the bias (Jeff. A, 2016). In the real study, the selection of sampling taken from the appropriate simple random sample within the population, sampling from stratified population or sampling from clustered within the population can be challenging and time consumed.

In this research, however, is not focusing on the effect biases from the sampling which in require the greater scope of works such as required additional services, expertise and deep literature. However, the literature just gives some surface preview how sampling bias could impact the response pattern in the questionnaire. Three prominent response biases often studied in the psychological literature usually violate the data analysis namely social desirable responding, acquiescence and extreme response bias (Paulhus, 1991).

The social desirability define bias as a tendency to make oneself look good in term of prevailing cultural norms when responding to questionnaire items (Mick,1996). Unintentional distortion from acquiescence bias may happen when there is a tendency to agree rather than disagree with items regardless of item content with positive connotation example *“Yea-saying bias”* (Brown and Maydeu-Olivares, 2017). While response tendency bias, this participants tend to endorse extreme response categories on a rating scales. Example choose to pick an extreme degree in category scales either 1 or 5 on a 5-point LS scale regardless of any content. Central tendency bias was also reported as bias when the response only pick central or neutral category scales (Bertram, 2007).

In this study, however, emphasising the investigate of another level of bias to data instruments which associate with item bias, scale bias and stimuli bias

(Figure 2.2.1) which affect the targeting potentially distorted when using rating scales methods and statistical outcome (Wilson and Gilbert, 2005).

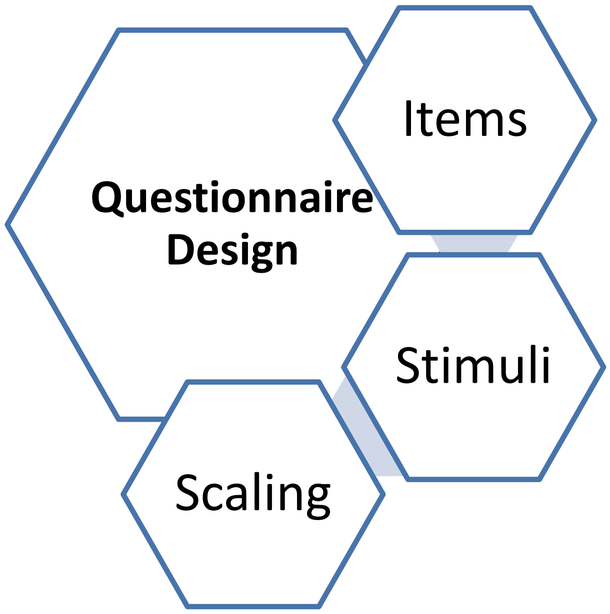


Figure 2.2.1 Three major components construct questionnaire design

2.3 COMMON ISSUE IN QUESTIONNAIRE DESIGN

The inaccuracies of the self-assessment report depended on how well it was designed which can be clustered into three components. First were look at how the items are constructed, second, the sample or stimuli and third type of scaling or category scales being used to construct the questionnaires.

2.3.1 Items

Dimensionality issue is a major issue in constructing a variable. The items being developed based on affective words usually associated with multidimensionality which carried full dimension of attributes with the facts that is difficult to quantify statistically. Thus in AE, the individual differences can be mapped on a single genuine number line such as variable in unidimensional (Andrich, 1988).

The process was reduced dimensions of the adjectives that account for the maximum amount of variance underlying a dataset while maintaining the most critical variance. However, these adjectives sometimes remain as multidimensional although adjective reduction has a reducing the size of the dimension. The adjectives may sometimes violate the measurement structure of unidimensionality and degrades the statistical outcome because RM works when the items are unidimensional best (Camargo, 2013).

One of the reasons why data could be troublesome because of the items developed exposed from grammatical ambiguity which carries different interpretation — unfamiliar words or statements being used such as medical, legal or technical terminology often becoming a burden for laymen participants to understand. Item bias is one of the major sources of inaccuracies and can be troublesome to responses when the items carry multiple interpretations, connotation or vagueness such as grammatical ambiguity, complicated vocabularies and bipolarity resulted the response can be distorted (Floyd J. Fowler, 1995). Adjective without true linguistic contrasts may introduce ambiguity that may introduce missed interpreting (Camargo, 2013). For example, adjective hard – a soft, participant may feel challenging to estimate how hard is hard, and how soft is soft? The contrasts are not the same distance.

On other hands, local dependencies often associate with bias and error where the item was too predictable (Purya Baghaei, 2008).

Bias in items resulted in the item is difficult to understand by the participants. Survey becomes a burden for them when the items so tedious or not enjoyable. Bias will impact the poor respond pattern becoming inconsistency and can potentially avoid that resulting missing data which affect the statistical reliability. Bias can due to difficulties of the item promote sources of variance associated with reliability, factor scores and group means, can cause bias (Borsboom, 2006). Mean value of distribution deposit around in the middle value of the scales because participants did not understand the adjective correctly or could be the adjectives is meaningless thus prefer to check middle scales (Schutte and Eklund, 2010).

2.3.2 Scales

A questionnaire using SDS and LS is the most popular and reliable methods to obtained participants emotion attitude towards a product and services. SDS and LS are widely used to estimate a quality index of customer satisfaction, measuring perception and attitude, estimate scientific properties, cognitive behavioural analysis, psychological and psychophysics assessment (Figure 2.3.1). Although questionnaire using SDS and LS offer high validity due to low manipulative quality (Cortes, 2013), however, a study has reported that SDS and LS are drawn some disadvantages where these scales associated with bias and error.

In previous studies have observed many instances in which the probability of participants endorsing each category on a response scale is not sequentially ordered when we would expect it to be (Camargo and Henson, 2011). While categorical response scales are intuitive, participants are not able to easily discriminate the categories of the response scale because participants have difficulties to understand and use the scales (Krosnick and Presser, 2009). The category scales affect poor targeting because the scales offer greater choices of response answer which have a higher risk to be missed targeting to certain ability participants which resulted in bias in statistics (Henson and Camargo, 2014).

A similar study has reported that LS tends to distort every option choice in category scales which inflates poor targeting (Brown and Maydeu-Olivares, 2017). The terms “extremely”, “quite”, “somewhat” and “slightly” denoted denotes to linguistic quantifiers that have associated with the more or less equal degree of intensity may in introduced ambiguity and haziness to understand (Charles and Suci, 1955; Osgood et al., 1957).

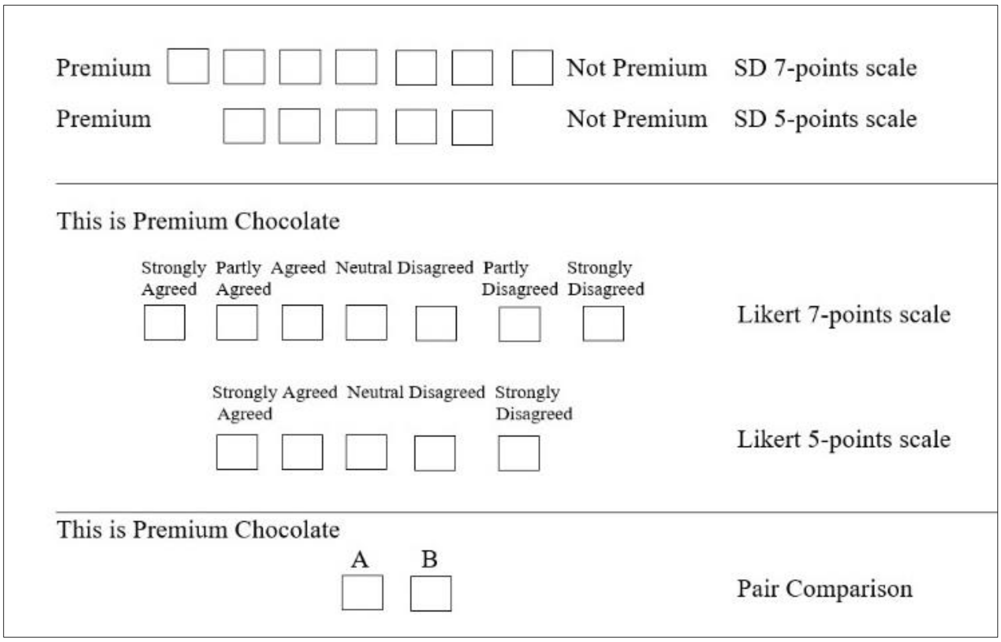


Figure 2.3.1 Common scaling technique used in AE studies

Along similar lines, participants find it challenging to deal greater category scales. Greater scale means the task is becoming tedious, and the distraction inflates counterproductive since the survey aims to get an instant answer that reflects a ‘truer’ evaluation in the sense that it gets to a participants’ underlying beliefs and attitudes (Al-hindawe, 1996). Gulliksen (1958) claim that the seven-category scale results in a far too coarse grouping have proposed a finer unit in SDS scale by merely using the 20-point or 30-point scale instead of the seven-point (Gulliksen, 1958).

In contrast, Winstanley (2013) and Bourke (2013) opinions seem more plausible, the smaller numbers of scaling will evoke better discrimination (Winstanley et al., 2013; Bourke-Taylor et al., 2013) as the participants could not discriminate a large number category scales. In the literature associated with some scales, it remains a subject of deliberation as for how many category scales that works “best”. As a general rule, the scale recommended is in the region of five to seven point (Cox, 1980; Winstanley et al., 2013). Some study recommends seven better than five-point scales for more sensitive ratings while five-point scales appear too narrow (Schutte and Eklund, 2010).

Osgood (1971) demonstrated that the use of the SDS technique with the same stimuli have shaped the large numbers failed to respond and decreased the speed of responses. On other hands, biasing effect on scales would introduce unfair decision (Joubert et al., 2015). The scales promote extreme person where SDS and LS scales allow the participants consistently choose extreme scales, while another participant tended to avoid extreme response but chosen neutral point which introduces central tendency bias and makes trying to determine knowledge level of the participants and to estimate product properties hierarchy (Bertram, 2007; Talikoti, 2016).

Scales-questionnaire often introduces some bias with the unfair decision and distorted the category scales to be treated equally (Cliff, 1973; Joubert et al., 2015). Bias in category scale are well known as unequal stimuli spacing where gives a non-linear response (Poulton, 1989). In similar connotation, a greater alternative to the response answer often introduces risk to bias and error (Ekman and Lennart, 1964; F.R. Camargo and Henson, 2015). Impact of bias would cause the rating scales projecting non-linear measurement. The psychological impact would see bias introduce an amount of blocking to think freedom and response spontaneously. One of the factor source bias in data collection is participants have difficulty to understand and use a rating scale (Krosnick and Presser, 2009), participants do not have familiar units to estimate (Poulton, 1989) and problem in guessing in scaling (Bond and Fox, 2015).

2.3.3 Stimuli

Ambiguity stimuli derived participant's different perception. Stimuli used may expose to bias and error. The basis of the about stimuli judgement, the outcome of the observation was defined how far the stimuli were separated on it (Cliff, 1973). How well the difficulty of the items affects the separation on the scales. The contrast of the stimuli reflects the way it behaves; If the products are too different and participants find it too easy to discriminate because greater contrast would discriminate clearly (Ahmad et al., 2018). However, if the product has a similar contrast then the stimuli are difficult to discriminate, distinguish and understand by the participants. The difficulties of the items play an essential role in participants' justification that affects the response style because of the fundamental assumption that the participants behave according to instruction (Ekman and Lennart, 1964).

Giving an example participants were asked to evaluate which grey colour is the best for cars' exterior. The evaluative might be difficult if the participants do not know or familiar with how to use common units, for example, twenty-five per cent grey, fifty per cent grey and seventy-five per cent grey to discriminate amongst the colour specimens, then the judges are biased because the decision could be based on own subjective units or guessing. Steven's (1989) states that stimuli could be sources of bias if the judgement is made without common units and using participants' judgement (Poulton, 1989).

One reason why category responses might be troublesome could be because participants are asked to evaluate products separately, without reference to a benchmark product. Discrete stimuli is another can be a source of biases, in some cases, the data collection performed in a small or large group of participants where the stimuli need to share with others. Discrete stimuli can potentially introduce bad judgement where the subsequent stimulus was performed better or worse against the first sample. In most cases, participants would skip passed stimulus rather than evaluating back. On other hands, the participants may feel difficult to remember the attributes when comparing with the earlier stimulus against the subsequence stimulus.

In a real survey, most of the participants were used personal opinion to distinguish the stimuli rather than to use objective measurement to discriminate and this affected bias decisions. In evaluating the PQ of tactile surfaces, the bias would potentially violate the measurement structure in a condition where participants struggled to use familiar units to estimate the magnitude of the products and do not know how to quantify the quality of the surface. The evaluation might potentially carry some degree of bias. Jansari et al. (2000) reveal a lateral bias that the perceiver or participants show a bias to the stimuli, the so-called “perceived bias” (Jansari et al., 2000).

2.3.4 Impact of indiscriminating

Indiscriminating leads to inaccuracy and would potentially affect the good-fit in statistical analysis (Suhonen et al., 2013). Common solutions when calibrating statistical analysis would target the outlier such as extreme person and items would introduce the problem that affects overall aggregate scores especially fit-residual. In most of the cases, the outlier will be removed from their original dataset. Although this is the best approach to spot and improve misfit in statistical analysis, it has some implication to fit statistics because fit statistics are dependent on the number of items and participants (Henson and Camargo, 2014).

2.4 SOLUTIONS

In this study, one of the possible alternative to overcome shortcoming problem from category scales, items and stimuli bias is introducing force-choice format known as PC which scientifically able to reduce the response biases (Joubert et al., 2015). The problem in category scales would promote inequality in targeting which derived from a greater option in category scales. This may promote bias in bigger scales. Although PC work differently, PC able to reduce response-style bias because PC is not offering inequality to the binary scales which is major troublesome in SDS and LS.

In AE domain, most recent studies aim to assess humans' latent traits which by nature the attributes at an ordinal level that need to assessed at interval scales (Arikan and Gelbal, 2013). In contrast, some experts claim category scales are usually treated as interval data when they are at best ordinal (Stevens, 1946; Wright and Linacre, 1989). Despite debating the scaling related issues, it is good to understand fundamental of scales and which scales perform best.

One of the objectives of this work is to establish which scaling techniques offer to the participant's capabilities to evaluate the products using PC rather than SDS and LS. The scaling often associates with the level of measurement. Scaling refers to the various technologies to assign the numbers to a series of objects (Manheim, 1977) and numerals can be assigned under different kind of rules, scales and measurement (Stevens, 1946; Sarle, 1995). The earlier development of scaling visualised the complexity to the scales becoming explicit due to various rules to assign the number and mathematical properties or group structure resulting scales which can be measured (Stevens, 1946).

To discuss which scales perform best is ideal to look at how the scaling was developed. Scales system introduced since the late 1920s where the studies were designed to measure attitudes and to a less extent, psychometric and psychophysical research (Manheim, 1977; Lorraine M. Uhlaner, 2005). Leon Thurstone's in 1927s has pioneered and responsible for introducing the law of comparative judgement to measure attitudes (Thurstone, 1927). Thurstone's works have been credited many researcher as precursors of the modern psychometrics due to its contribution of proposing measuring the separation between two opinions on altitudes in psychometrics studies and plot the means on scale continuum (Thurstone, 1927; Camargo, 2013).

Rensis Likert in 1932s came after the Thurstone and was responsible for introducing a popular technique for the measurement of attitudes of LS (Likert and Likert, 1976). Stevens (1946) introduced the theory of scales of measurements where he defines category scales of Nominal, Ordinal, Interval and Ratio to determine the level of equality. In early 70s Osgood introduce SDS as a tool in measuring meaning (Snider and Osgood, 1969; Osgood, 1971). A decade later, Nagamachi (1980) improved the SDS scales and introducing opposite bipolar scales which was popular in AE (Nagamachi, 2008; Lokman and Nagamachi, 2010).

George Rasch (1960) introduced Probabilistic RM for some intelligence in measuring person and item abilities (Rasch, 1980). Andrich in 1988s introduces the “Green” Book for the RM measurement (Andrich, 1988), Fischer and Molenaar in 1995s introduce the “Yellow” book (Fischer and Molenaar, 1995).

2.4.1 Visual Analogue Scales

The visual analogue scales (VAS) is continuous lines scale instead of category scales to obtain more variability that widely used in measuring sensory rating such as satisfaction, perception magnitude and often to measure pain in physical and rehabilitation medicine (Kersten et al., 2012; Sung and Wu, 2018). VAS used a line continuum with the extreme magnitude on both directions such as “*not at all”* and *“very much”* on opposite direction similar like semantic space to determine the responses position of each Kansei word using KESo software (Schutte and Eklund, 2010).

Despite the advantageous of VAS making the choice of preference free from interval effects like SDS and LS scales, some study was reported that the scales using in VAS are not equally spaced and response find it difficult to judge how to rate their pain on VAS scale resulting the finding not very accurate, sort of random, almost guesswork and having work into number first (Jackson D, H et al 2006) (Kersten et al., 2012). Some study was also reported VAS is ordinal scales resulting in non-linear measurement using Rasch analysis and very low test-retest reliability (Kersten et al., 2012) (Carlsson AM, 1983).

In this study, VAS scales were not carried out with PC study, because the objective of PC study is to compare between comparative scales and non-comparative scales. Which scales are performed less bias effect in the context of greater discriminating contrast, instantly easy to response and valid and reliable statistical fit using RM. VAS technically considered Non-comparative scales because it provides resembles scales structure of category rating scale like LS and SDS, where respondents need to determine their choices of preferences in a continuum scale. While PC does not require an interval scales. Thus VAS was not used to compare the scale with PC, because VAS is not comparative scale because VAS has the same scale structure with SDS and LS.

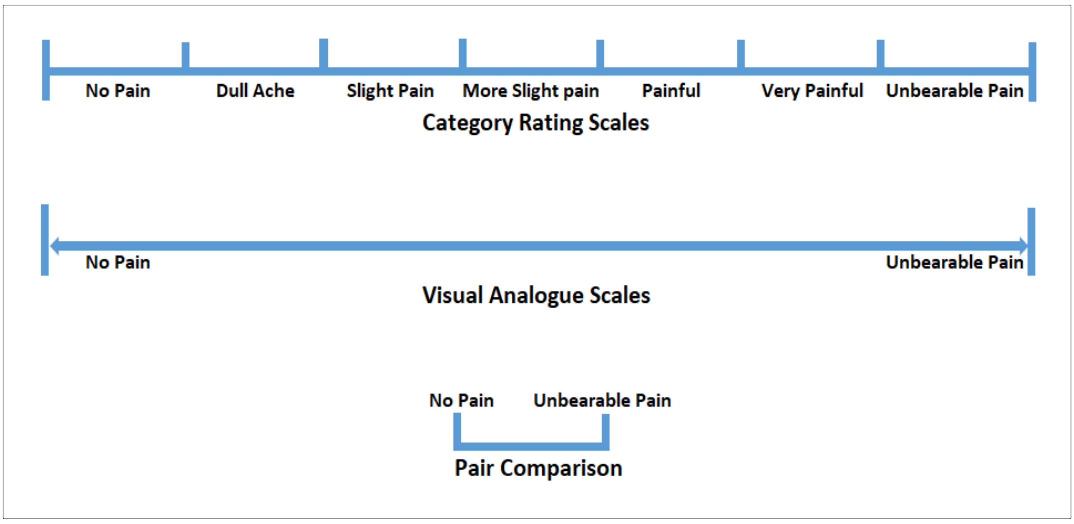


Figure 2.4.1 Example of category scales, visual analogue scale and PC (reproduce from (Mattacola et al., 1997)

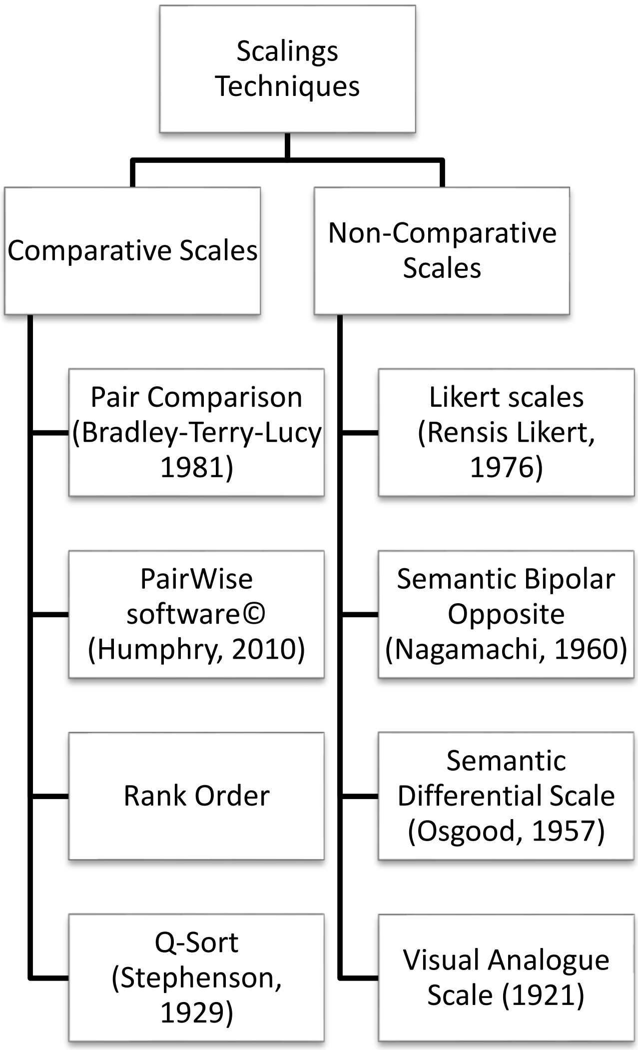


Figure 2.4.2 Scaling Techniques

Figure 2.4.2 illustrates that the scaling technique is divided into two categories. The Comparative scales and Non-comparative scales. In social science research, there are many methods associated with these scaling technique. However, there are few methods most commonly used in measuring subjective response, and affective response often used rating methods (Schutte, 2005) which were classified as Non-comparative scales. LS and SDS were the most popular scales. Comparative scales or ranking method is also widely used as an alternative to rating scales. Most commonly used is PC.

These three scales are among the favourite technique in measuring subjective responses. However, it is often associated with bias and error. To investigate further, these three scaling were demonstrated.

2.4.2 Rasch Model Theory

In classical test method, the simple calculation can define how many items correct from the person. However, this approach lack of fairness as the score cannot be compared meaningfully if the next person got the same score (Engelhard, George, 1997). Simple example if person A got eight items correct out of ten possible marks in his Sudoku level one test, then person B got the same score of eight in level two Sudoku test which more difficult, then how to distinguish which one was more clever as they reached the same score.

This is the reason why in the 1950s and 1960s, the Danish Mathematician, George Rasch proposed a new method called RM to education measurement that calculating the score participants ability (Rasch, 1980; Engelhard, George, 1997). Underlying this theory RM, construct a linear scale along with items are located according to difficulty level and participants abilities.

Rasch analysis in simple explanation is the formal testing of an outcome scales against a mathematical measurement model and was used to examine to what extent the response from a scale satisfy measurement structure (Tennant et al., 2007).

In this study have applied a probabilistic model called RM to measure participants' affective responses to products and to detect any anomalies from misfit data which normally associate with bias. Data will consider unbiased if the data is correct with a true value over repeated trials of the parameter as expected in RM (Engelhard, George, 1997).

Since the 1960s it was widely used in modern statistical models in measuring psychology, sociology and education although the application of Rasch analysis published in the health sciences (Andrich, 1988; Tennant et al., 2007). Andrich (1988) states that the RM is commonly understood and used procedures for social science measurement which the model derived from the work of Thurstone (Thurstone, 1927).

In this approach, rather than constructing a statistical model of the response data, it is to determine whether the response data fit the RM and, if it does, some measurement properties demonstrated. The RM, in the context of product evaluation, calculates the probability that someone will endorse a product as a mathematical function of the participants' ability or willingness to endorse and the difficulty of endorsing the particular product. The result derived a linear scale of how easy it is to endorse a product along the affective dimension the instrument is designed to measure.

The RM used widely in education and medicine such as measuring patients' health condition and pain assessment. In our previous research, the literature was reported and observed many instances in which the probability of participants endorsing each category on a response scale not sequentially ordered when we would expect it to be (Camargo and Henson, 2011). Thus, while categorical response scales are intuitive, participants are not able to easily discriminate the categories of the response scale.

In the recent survey, a part of observing the product properties using SDS and LS scales, the questionnaire was designed to assess and categorise the participants' knowledge and abilities to achieve the valid and reliable statistical outcome. The RM was the simplified method to estimate the invariance comparison and sufficiency to a scales. However, the outcome of the study reported the disadvantages of considering category scales which promote bias in statistics (Camargo, 2013).

AE required statistical model like RM for adding the value to the model. The model fit is a fundamental requirement govern in the Rasch Model and acts as a quality-control mechanism (Bond and Fox, 2015) and an indicator to measure the validity that portrait feature values to the study.

2.4.3 Rationale the choice of Rasch Model in AE study

In AE studies often used multi-variate statistical analysis, a subdivision of statistics encompassing the simultaneous data observation more than one analysis of results and how these analyses are correlated with each other. Previous AE studies used a multi-variate statistical available model such as PCA and factor analysis (FA) to observe which variables stem higher or lower loading to determine dimensionality and how the properties are correlated. However, in this study, AE used RM to extend the outcome which to determine probability distribution which often associates with linearity, linear regression and multiple regression.

In this study, the rationale to carry out RM analysis is to assess whether linear measurement of affective responses can be derived from PC using the Rasch Model. This approach, rather than constructing a statistical model of the response data, is to determine whether the response data fits into the Rasch model.

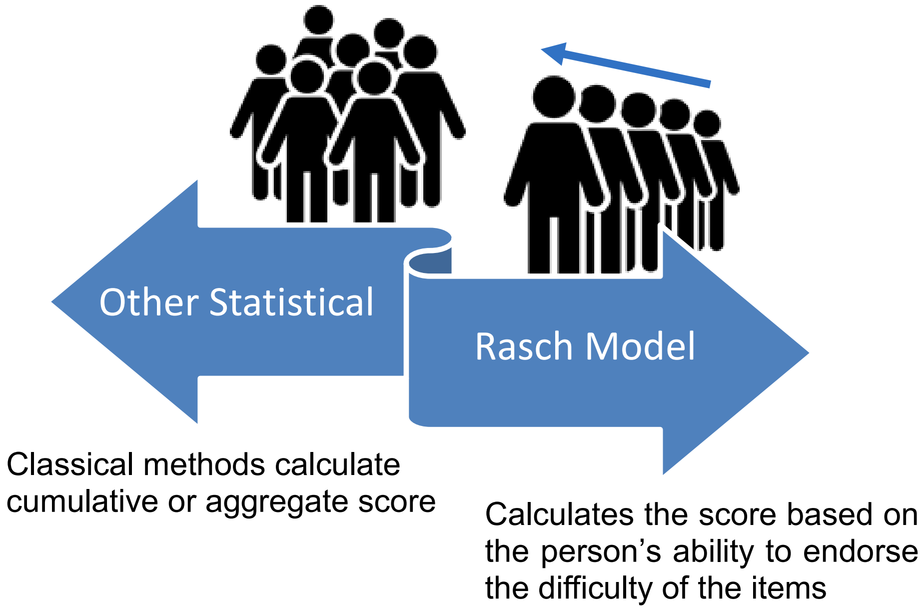


Figure 2.4.3 Statistical approach between classical methods and Rasch Model

2.5 PAIR COMPARISONS

The basis of the PC model is to determine a degree of preference between two objects (Glickman, 1999). PC uses binary scales or dual-scales to assess product properties.

One of the aims of this work is to establish whether participants might find it easier to evaluate products if the evaluations made as paired comparisons. When making a paired comparison, the participant merely has to indicate which of two products they endorse more readily, rather than thinking about which category of response one would elicit separately. The challenge is then to derive a linear scale of affective response from such comparisons.

There is a body of work from a discrete choice theory which based on making paired comparisons of products (Train, 2009). The aim in discrete choice theory, however, is to determine the relative importance of properties of the choices, which are often assumed to vary linearly, rather than to derive measurement. Thurstone's law of comparative judgement' (Thurstone, 1927) can be used to establish measurements from pairwise comparisons. It has been widely used in psychophysics to determine the relationship between perception and intensity of the stimuli. The Bradley-Terry-Luce (BTL) model (Bradley and Terry, 1952) is derived from Thurstone's law of comparative judgement but uses a slightly different statistical basis. In the context of education, it can be used to derive measurement from whether answers to questions are right or wrong, yes or no, true or false and agree or disagree (Cortes, 2013).

BTL can be shown to be equivalent to one of the forms of the RM. The model is not directly applicable to the evaluation of products because, in the educational condition, there is a response associated with each item for each participant while in a different context, the items and responses for each participant for each product.

In other words, the product forms an extra independent factor or facet for which the BTL model and Thurstone's law cannot account. There are, however, forms of the RM that might be able to account for the extra facet (Linacre, 1989).

The term PC in this study should not be confused with the use of PC for statistical comparison to compare different reader results. In this study, PC refers to a method whereby participants need to compare and choose between two stimuli side by side (Figure 2.5.1)

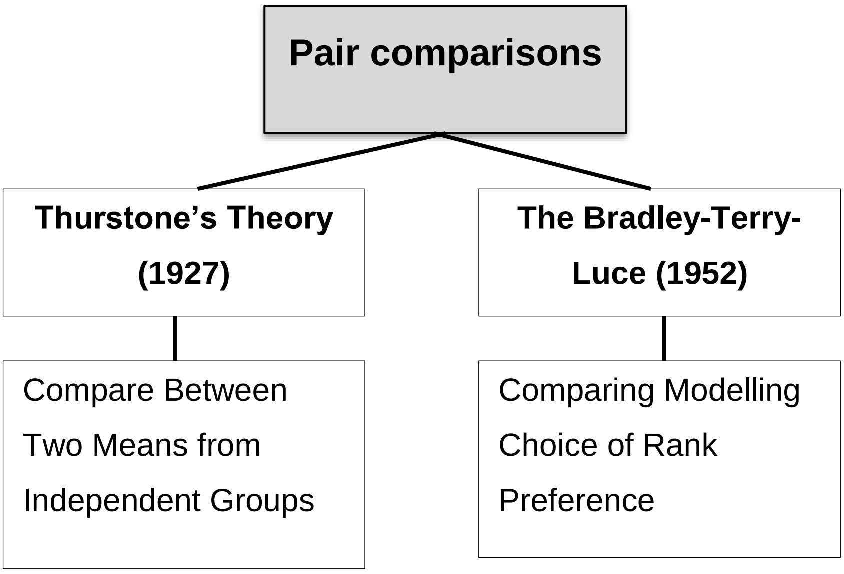


Figure 2.5.1 Two methods in assessing Pair Comparisons

2.5.1 Thurstone’ PC

In modern psychometric theory, PC using the Rasch Model approach is inspired based on Thurstone’s law of comparative judgment and fifth comparative matrix (Thurstone, 1927). Thurstone used the normal distribution in applications of the model. The final form of Thurstone’s Law of comparative judgement can be written as given in equation (1);

$$\boldsymbol{X_{AB}} = \frac{\boldsymbol{S_A - S_B}}{\boldsymbol{\sigma_{AB}}} \quad \text{Equation 1}$$

Where, X_{AB} is the deviate of the normal distribution corresponding to the proportion of judgements ‘A beats B’ and $S_A - S_B$ and σ_{AB} are as defined above. In other words, the scale separation between two objects on the psychological continuum is measured in units of the standard deviation (SD) of the difference between the distribution of their discriminable process (Bramley and Black, 2008).

Give a comparative example of previous studies using the first type of PC to analyse teachers’ reliable assessments on writing performances (Heldsinger and Humphry, 2010) and to compare expert judgement which used cumulative normal distribution using BTL model (Pollitt, 2012). This type of PC was made to compare different reader results using PairWise software© (Humphry, 2010) and in statistical analysis using analysis of variance (ANOVA) or Paired *t*-Test to compare between the means value of independent groups or conditions.

2.5.2 Bradley-Terry-Luce Model

Between Thurstone model, BTL is widely used to establish ranking based on criteria, to assign quantitative weight so that the qualitative ranking is satisfied, rate the importance and to prioritise users' needs based on comparable items

However, in some situation, the direct ranking may not be feasible for example when involving a large number of object or items to rank such as a sport tournament, estimating team strength using the BTL (Bradley and Terry, 1952; Luce, 1959). The model requires data on each team's performance against a set of opponents with each game treated as a PC having a dichotomous item for example win-lose, right-wrong and agree-disagree.

In the original model of BTL, the theory PC were the rank of an object in the pair is given in equation (2);

$$\boldsymbol{P_{ij}} = \frac{\boldsymbol{\delta_i}}{\boldsymbol{\delta_i + \delta_j}} \quad \text{Equation 2}$$

where P_{ij} denotes the probability that objects *i* is ranked higher that object *j* (or that *i* wins over *j*), or one win the other loses and δ is the scale location parameter for object *i* and *j* (Alvin and Adams, 2015). In context of win or lose, which widely used in sports and games, (Glickman, 1999) states, the probability team winning can be estimate as given equation (3)

$$\boldsymbol{Prob\ (a\ beat\ b \big| (v_{a,} v_b)} = \frac{\mathbf{exp}\ (\boldsymbol{v_a - v_b})}{\mathbf{1{+}exp}(\boldsymbol{v_{a,} - v_b})} \quad \text{Equation 3}$$

Where two objects *a* and *b* are compared according to some criteria for "Goodness". The true quality of each is given by their parameter $v_a and\ v_b$ then above equation will calculate the probability that objects *A* will be chosen as the better of two equality (Pollitt, 2012).

On the other hand, Pollitt (2012) has illustrated the simplest way in equating PC based on BTL model, and he claims that equation shows better form compare to Thurstone's original normal distribution form. The difference between two quality is equal to the log of the odds that *a* will be judged better than *b* if these comparisons are compared several times by different judges or participants, the resulting data allow us to estimate the gap between the two parameters.

The gap in this particular context will demonstrate estimate standard error or SD and maximum likelihood. PC can be written in terms of log odds (4);

$$\boldsymbol{P_{ij}} + \boldsymbol{P_{ji}} = \frac{\boldsymbol{\delta_i}}{\boldsymbol{\delta_i + \delta_j}} + \frac{\boldsymbol{\delta_j}}{\boldsymbol{\delta_j + \delta_i}} \quad \text{Equation 4}$$

$$P_{ij} + P_{ji} = \frac{\delta_i + \delta_j}{\delta_i + \delta_j}$$

$$P_{ij} + P_{ji} = 1$$

In the context of measuring affective response using confectioneries or example, PC formulation can be written as given in equation (5);

$$\boldsymbol{P_{fr\,mw}} + \boldsymbol{P_{mw\,fr}} = \frac{\boldsymbol{\delta_{fr}}}{\boldsymbol{\delta_{fr} + \delta_{mw}}} + \frac{\boldsymbol{\delta_{mw}}}{\boldsymbol{\delta_{mw} + \delta_{fr}}} \quad \text{Equation 5}$$

$$P_{fr\,mw} + P_{mw\,fr} = \frac{\delta_{fr} + \delta_{mw}}{\delta_{fr} + \delta_{mw}}$$

$$P_{fr\,mw} + P_{mw\,fr} = 1$$

$$\boldsymbol{P_{1,0}} + \boldsymbol{P_{0,1}} = \frac{\boldsymbol{\delta_1}}{\boldsymbol{\delta_1 + \delta_0}} + \frac{\boldsymbol{\delta_0}}{\boldsymbol{\delta_0 + \delta_1}} \quad \text{Equation 6}$$

$$P_{1,0} + P_{0,1} = \frac{\delta_1 + \delta_0}{\delta_1 + \delta_0}$$

$$P_{1,0} + P_{0,1} = 1$$

where $P_{fr\,mw}$ denotes the probability that Ferrero Rocher *(fr)* chocolate is ranked higher that Milky Way chocolate (code *mw*) or (code *fr*) wins over (code *mw*) 1 = win and 0 = lose as shown in (5). The proportion win or lose pairwise can written as (6). Pollitt (2012) also demonstrated, the similar equation can be write based on Rasch model (1960) maximum-likelihood (ML) where score 1 will be obtained for A if beats B, while if B wins, the score will be 0 for B that as given in equation (7);

$$X = 1\ if\ A\ wins\ , 0\ if\ B\ wins \qquad \textit{Equation 7}$$

While the second type of PC was used in this study to compare between two alternative stimuli, the works associated with this technique include studies to estimate the probability which team win the Australian Football League amongst sixteen teams (Alvin and Adams, 2015). Compare ten biomedical images of chest radiograph to determine which image appear sharpness (Phelps et al., 2015). To compare the individual loudness among the forty-five tone pairs (Schneider, 1980). PC also were used to establish ranking based on criteria, to assign quantitative weight so that the qualitative ranking is satisfied, rate the importance and to prioritise users' needs based on comparable items for example in comparing customer want list in quality function deployment (QFD) (Fernandes et al., 2008).

Studies by Glickman (1999) has differentiated between static and dynamic PC. Dynamic PC was usually used to estimate scores in sport or games where the performance change over time for example chess tournaments. While static PC used in comparing modelling choice of rank preference, for example, soft drink brand A versus soft drink brand B (Glickman, 1999) which it has much similar in this study to estimate rating preferences based on interior vehicle stimuli using BTL methods (Bradley and Terry, 1952).

Thurstone's Equal-Appearing interval or of PC which was not new to the research domain especially in psychophysics where Louis Thurstone, an American psychologist working way back in 1920s – 1950s in designing methods in measuring participants attitude such as perceived of seriousness of crime and perceived quality of handwriting (Thurstone, 1927; Bramley and Oates, 2011). The PC method was not only successful in education and health science, but the method has been practised in measurement judgement and choice in social science, market research and also sports (Bond and Fox, 2015).

Recently the method of PC were used to study friend choosing among adolescents (Özmercan and Kumandaş, 2016), psychology sentiments (Dalitz et al., 2018), colour differences (Guan and Luo, 1999) and recently to fast moving consumer products, where PC was used to estimate the specialness of the confectioneries (Ahmad et al., 2018).

In an automotive application, PC has been used to measure sound quality from diesel engines (Champagne and Shiau, 1997) and vehicle armrest to obtain overall touch and feel. This study, however, PC was used in slightly difference which to rank-order the stimuli but using dual scaling system of five-points scales and opposite bipolar adjectives to obtain items responses (Bhise et al., 2007).

2.5.3 Rank order using paired comparison

This study has used the work from BTL equation to estimate mean logits from all possible pairwise. The PC technique allows the participants to observe the pairwise stimuli each time. For example to respond on the specialness of four pieces of chocolates with some criteria "*This chocolate is premium''.* Participants were asked to response the item by comparing between two stimuli directly based on criteria given rather than judging separately in SDS and LS. Each stimulus for example *(A, B, C & D)* will be assigned in all possible pairwise combinations fairly beginning with stimuli *A vs B, A vs C, A vs D, B vs C, B vs D and C vs D;* this rank sequence eventually shaped six possible comparisons.

The concept of rank order using PC derived from the dichotomous probabilistic model. Guliksen and Tucker (1958) illustrated the concept of a compromise between direct ranking and complete paired comparisons where all the pairs are small subsets that can be directly compared. The rank order is sometimes known as Youden Square (Dunn-Rankin et al., 2004).

Table 2.5.1 Possible rank order sample

Rank order (S = sample)			Probabilities order	Consistency
1st	2nd	3rd		
S_a	S_b	S_c	➡ $P_{ab}\ x\ P_{ac}\ x\ P_{bc}$	Consistence
S_a	S_c	S_b	➡ $P_{ac}\ x\ P_{ab}\ x\ P_{cb}$	Consistence
S_b	S_a	S_c	➡ $P_{ba}\ x\ P_{bc}\ x\ P_{ac}$	Consistence
S_b	S_c	S_a	➡ $P_{bc}\ x\ P_{ba}\ x\ P_{ca}$	Consistence
S_c	S_a	S_b	➡ $P_{ca}\ x\ P_{cb}\ x\ P_{ab}$	Consistence
S_c	S_b	S_a	➡ $P_{cb}\ x\ P_{ca}\ x\ P_{ba}$	Consistence

This model will assign paired stimuli in possible rank order that denotes to possible probabilities. Table 2.5.1, demonstrates the simple probability can be modelled for the dichotomous model where participants *n* will endorse stimuli *A* is greater or special than *B* as simple elucidation. In order to run the Rasch model software called RuMM 2030, each probability must have consistence probabilities. However, sometimes the rank order have inconsistence order. This inconsistency will introduce some misfit during the Rasch model analysis.

In PC, the probabilities equation of sample *A* (S_a) is greater than sample *B* (S_b) and sample *B* (S_b) is greater than sample *C* (S_c) are the logic rank order derived similarly as 1st, 2nd and 3rd rank magnitude. The similar concept of the equation will construct possible probability of $P_{ab}\ x\ P_{ac}\ x\ P_{bc}$ that interpret as probability sample a > b, the probability a > c and probability b > c.

In the Rasch model, the logistic model denoted in Rasch Model Theory as an item characteristic curve (ICC). The degree of endorsement (logit) denotes to mean location as shown in Figure 2.6.1.

2.6 The Rasch Dichotomous model

An expected score is calculated for each object, using the current estimate parameters to predict the outcome of every comparison (Pollitt, 2012). The Rasch model articulates the probability that a participants will response an item with two-category response scale for example Yes or no, true of false, agree or disagree as a logistic function of the difference between the person's location (β) and the item's location (δ) on a linear continuum (Rasch, 1980; Andrich, 1988; Camargo, 2013) can be written as (8);

$$\mathbf{Pr}\{\boldsymbol{X_{ni}} = \boldsymbol{x} \mid \boldsymbol{\beta}, \boldsymbol{\delta}\} = \frac{\mathbf{exp}[\boldsymbol{x}\ (\boldsymbol{\beta_n} - \boldsymbol{\delta_i})]}{\boldsymbol{\gamma_{ni}}} \text{ Equation 8}$$

Given that $\gamma_{ni} = 1 + \exp[x\ (\beta_n - \ \delta_i)]$, where $x \in \{0,1\}$, taking *1* as a choice of response and *0* otherwise. $\Pr\{X_{ni} = x \mid \beta, \delta\}$ is the probability that a person *n* will endorse an item *i*, such that $0 \leq 1 \Pr\{X_{ni} = 1\}$ and - $\infty \ \leq \ (\beta_n - \ \delta_i) \ \leq = \infty$**.** The relationship between the difference in person location on the continuum and the probability of a positive response, denoted in the Rasch Model Theory (RMT) as item characteristic curve (ICC) Figure 2.6.1

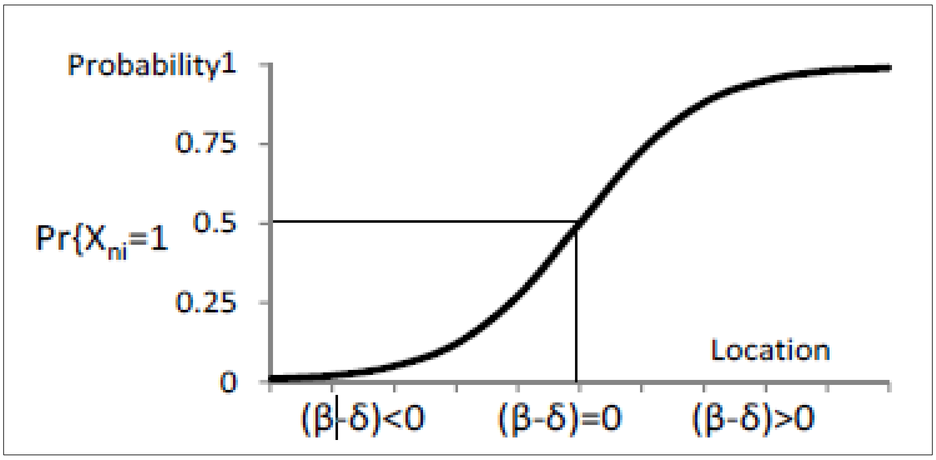


Figure 2.6.1 Probability of a positive response associated with persons' location on the continuum (Camargo, 2013)

Chapter 3

Applying the Rasch Model for Affective Response to Vehicle Interior Textures

In Chapter 3 two studies are reported. The aim is to determine how well the data measured participants' affective responses to vehicle interior textures using the Rasch model. Two studies were designed to elicit the affective responses from participants evaluating the PQ of vehicle interior textures using a seven-point SDS and a five-point LS namely SDS1 and LS1 respectively. A pool of items was developed using the AE approach. The observed data from both studies fitted the Rasch model reasonably well, but some misfit was identified. One of the violations indicates that the person-items are exposed to biases because of weaknesses in the rating scales. This study is important as it will examine the extent to which the biases affect quantification in measuring participants' affective responses.

3.1 INTRODUCTION

The literature review (Chapter 2) reported affective engineering research which, when commonly used to elicit affective responses, may present biases and errors if the measurement principles are violated (Camargo, 2013). Some studies in psychophysics and psychology, when measuring subjective responses, have reported that biases would potentially affect the goodness of fit in the statistical outcome (Kingdom and Prins, 2010; Kubinger et al., 2012; Camargo, 2013; Dalitz et al., 2018).

Studies in affective engineering have established the importance of the test instrument to meet statistical fit to achieve valid and reliable outcomes (Suhonen et al., 2013; Anselmi et al., 2015; Christensen et al., 2017). However, in most cases, the outcome from these studies is that it is very unlikely they will produce data that perfectly fits into RM (Camargo and Henson, 2014). This is one of the reasons why measuring a human's latent traits is a considerable challenge.

Two studies measured participants' affective responses to the material used in vehicle interiors. In these studies, the aims were to determine the difficulty of the task by completing SDS and LS scales. Participants were asked to endorse vehicle interior texture specimens against LS statements and SDS items to assess the PQ of vehicle interiors.

3.2 HYPOTHESIS

The research formulates the difficulties of rating scales for affective responses to a product using SDS and LS scales would eliminate biases and error to meet a statistical fit to achieve a valid and reliable outcome as expected in the Rasch model.

3.3 OBJECTIVES

This study aims to identify the sources of bias that affect the difficulty of the task of endorsing the items. The study observed the misfit outcome from bias and the extent to which it violated the statistical outcome.

The aims of the research are therefore to establish

- Whether linear measurement of affective response can be derived from SDS and LS scales
- How well the items fit into the model using an SDS and LS scale
- How much bias is evident and how likely that bias is
- How successful the calibrations process is
- Which rating scales fit best to achieve invariance stability

3.4 METHODS

The affective engineering method were used to design the test instruments, while the RM was used to calibrate and validate the test instrument and results. SPSS was used to perform PCA.

3.4.1 Affective Engineering method – Focus Group

Three focus groups were recruited to establish how participants' feel about the quality of interior vehicle trims, especially related to tactile texture. This included how participants perceived high quality or low quality, comfort and discomfort, as well as other subjective perceptions that affect the driving experience. Texture plaques of the interior vehicle trims were made for the focus groups. The aim of the focus groups was to generate good adjectives and statements with greater scope based on the research context and for analysis before developing a pool of items.

The focus groups were held in the Affective Engineering Laboratory of the Faculty of Mechanical Engineering, University of Leeds. Eighteen participants (fifteen males and three females) were involved in this study with ages ranging between twenty-three and forty-two (SD 4.33). All the participants are regularly a driver or a passenger in a car. The first focus group involved nine male participants between the ages of twenty-three and thirty-two years old. The second focus group involved six male participants ranging from twenty-six to forty-two years old, and the third focus group discussion involved solely female participants ranging from twenty-three to twenty-five years old. All were undergraduates or postdoctoral staff within the university faculty, and all participants received £5 as compensation for their time taking part in the study.

3.4.2 Generating Affective words

3.4.2.1 User experience theory

To tailor questionnaires on the research context, the focus group was designed based on the elements of four pleasures (Jordan, 2000), a hierarchy of needs to reflect a hierarchy of human needs (Maslow, 1971) and understand consumer behaviour through internal psychology of the product experience (Desmet and Hekkert, 2007). All these theories are connected to the Mindspace approach (Dolan et al., 2012), which was outlined in the specific context where there are better ways to understand consumers' behaviour naturally.

The structure of focus group was used the original AE method known as Kansei Engineering Type 1 – Category Classification (Mitsuo Nagamachi, 2011) to facilitate focus groups in generating adjectives for a questionnaire. This technique was the most feasible to find relational design attributes between the affective and the design specifications (Nagamachi, 2008; Lokman et al., 2013).

3.4.2.2 Affective Engineering Approach

The focus groups began with a warming-up session watching two short video clips to aid participants with a greater understanding of this research context. The session started with a general discussion about interior quality of vehicles, the full space of product experience and the driving experience.

During the focus group, participants shared their experiences of using vehicles for daily tasks, such as fetching children from school, going to the office, supermarket, post office, bank and hospital or to a hardware store. The experience was extended to casual occasions such as holidays and visiting friends and family over the weekend. Some of the travel experiences were memorable, most enjoyable, fun and sometimes somewhat boring, dangerous or unpredictable. All these experiences were documented based on the questions posed by the assessor beforehand. The focus groups were carried out according to the AE framework.

Participants were asked to evaluate the visual texture and to touch and feel the physical texture stimuli. A questionnaire asked how they felt about the visual and physical textures if used to design vehicle interiors. These items were developed based on AE theory in assessing texture features such as aesthetic, physical, operation and sensation traits (Lokman and Nagamachi, 2010).

The affective word was recorded, then the word was adapted into SDS and LS scale questionnaires. These transcripts were examined in order to distinguish the adjective statements' quality. Vague statements such as assumptions, predicting or non-related context were deleted to avoid ambiguity.

Figure 3.4.1 Participants were asked to give their feedback based on the scenario given.

3.4.2.3 Affinity Diagram

SDS items were developed based on the reduction of adjectives using the Affinity diagram method (Lokman, 2010a). Two Affinity diagram focus groups were recruited for this study involving six native speakers who were research students from the Institute of Design Robotic and Optimisation (IDRO), Faculty of Mechanical Engineering, University of Leeds. The objective of having two groups as cross-validation strategy was to observe the consistency when comparing different groups.

One hundred thirty-seven adjectives were clustered, where participants were required to sort those adjectives into several groups. The adjectives were written on 3M stick-on yellow paper. Participants were required to sort the adjective cards into several groups. Then the header of the card was chosen for each group that represented a similar meaning from the list within the group.

Some of the adjectives that did not fit into any groups or did not have a clear meaning or were vague were left as stand-alone cards. These words will temporarily be parked into the small ‘parking lot’. After completing all the sorting, participants went back to the ‘parking lot’ and repeated the sorting process until the card could not fit into any of the groups. Participants were also allowed to duplicate the words and place one copy into each group when they could not come to any consensus or agreement. For the last task, participants were asked to develop the super header card as the ‘parent’ of the header card.

Table 3.4.1 Twenty-four adjectives for item pool

Characteristics			
Positive Sensory	Low-Quality	Emotional	Visual
Negative Sensory	Fundamental	Temporal	Touch & Feel
High-Quality	Functional	Futuristic	Textured
Perceptions			
Personal	Extreme	Indifferent	Stylish
Problematic	Positive	Extraordinary	Visual
Safe	Memorable	Normal	Functional

3.4.3 Data Collections

There were one hundred eighty-two participants, (45.16 percent female and 54.84 percent male), between nineteen and fifty-nine years of age (SD 13.1). All the participants consisted of staff, researchers, postgraduates and undergraduate students at the University of Leeds that were recruited into two groups. For the first group of seventy-five participants participated in the SDS1 study. Participants were asked to evaluate the quality of interior vehicle trims design with twenty SDS bipolar adjectives against seven texture stimuli, eventually creating one hundred and forty items in total.

The second group of one hundred and seven participants participated in the LS1 study. Participants were asked to evaluate the quality of interior vehicle trims design with twenty LS adjective statements against seven texture stimuli, eventually creating one hundred and forty items in total.

Each of the first and second studies were conducted in separate sessions. Both studies were held in the Affective Engineering Laboratory in the Mechanical Engineering Faculty, University of Leeds. All participants were compensated for their time with £5. Each session was no longer than forty-five minutes.

3.5 ETHICAL APPROVAL, CONSENT FORM AND RISK ASSESSMENT

Ethical approval was granted from The University of Leeds (Approval Reference No. MEEC 14-025) conforming to participants’ health and safety procedures. The ethical approval, consent forms and risk assessment are included Appendix B.

The protocol was briefly explained to the participants. Information about instructions and demonstration of touching texture samples and a demonstration of ticking the SDS box in the questionnaire was also briefly explained. The protocol and questionnaire are included in Appendix C.

3.6 APPARATUS AND EXPERIMENT SETTING

The studies used seven stimuli in SDS and LS scale data collection. The stimuli were specially made using standard automotive polypropylene plastic (PP) with standard dimensions of 225mm x 160mm x 3mm and were engraved with sandblasting in the organic to geometric patterns that are mostly seen in interior passenger cars. Each stimulus had been injected with a semi-gloss level of between 1.5GU to 2.2GU (gloss unit) in various colours such as black, matte black, brown and beige, labelled one to seven (Figure 3.6.1).

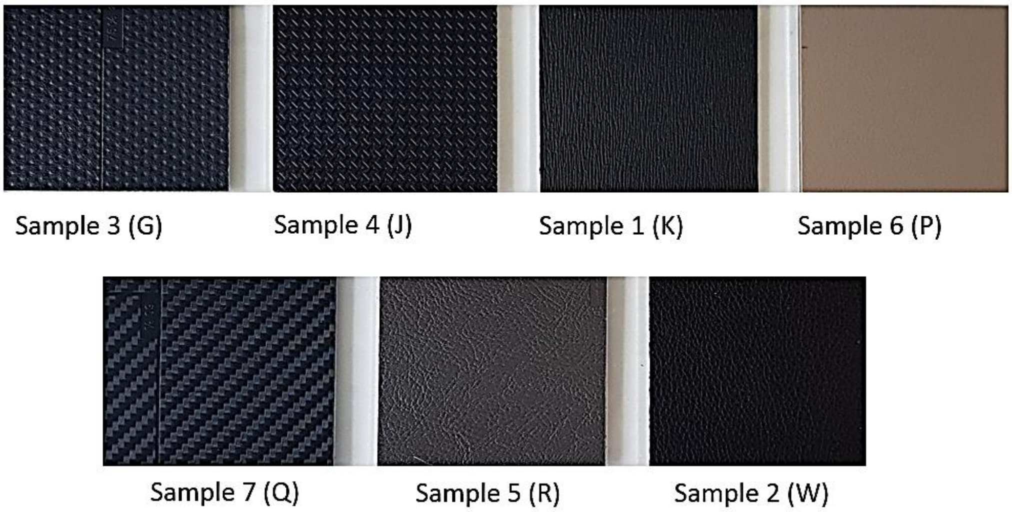


Figure 3.6.1 Seven stimuli of vehicle interior texture

3.6.1 The rationale in choosing automotive textures as stimuli

The automaker has acknowledged the importance of PQ with the aim to improve the vehicle interior design craftsmanship quality. The quality of craftsmanship often associates with interior trim components such as plastics textures, fabrics, leather, decorating finishing and other materials. However, in this study, the texture was chosen because this component is most noticeable and easy to be acknowledged by the users either high or low quality. The stimuli went through several stages of selection before being used as stimuli for data collection.

The automotive textures were chosen in this study because of the stimuli as the following reason;

- People often associate with interior plastic trim and materials in particular surface textures in measuring vehicle quality perceptions or PQ For example, high quality of product often associate with good properties of texture and materials finish.
- Texture stimulates the aesthetic values of the product; the way it looks, the feel of the material, the tactile or haptic response of controls and good ergonomics that give a certain amount of pleasure and satisfaction for customers.
- Vehicle interior texture provides physical properties that able to quantify subjectively and objectively through static and dynamic surface evaluation such as visual aesthetics, geometrics, fit and finish and haptic control, for example touch and feel.

3.6.2 Relationship of stimuli to the focus group study

Texture plaques of the interior vehicle trims were made for the focus groups. The aim of the focus groups was to generate good adjectives and statements with greater scope based on the research context and for analysis before developing a pool of items.

The stimuli were used to induce interest in the research context where the stimuli were used to mimic the actual interior vehicle trims. The stimulus was used to complement and aid the participants' understanding of the research context. Where some of the activities within the focus group asked the participants to discuss, demonstrates and estimates the quality magnitude on a scales and criteria given. Each item response may require some logical reason or feedback to justify their opinion, which normally associates with users' driving experience. The stimuli were used to generating an initial perception of high and low-quality attributes of products for example how the surface tactile, colour and glossiness properties affect their quality judgement. Prior to the activity, all participants were required to pay attention to a few short videos about the driving experience. Three standing panels were used to evoke the study context, which illustrates the vehicle exterior and interior.

3.6.3 Data administration

3.6.3.1 Building the SDS items

The questionnaire was made up of paired adjectives of opposite meaning in random order. There were twenty pairs of adjectives repeated across seven pages, equivalent to seven stimuli or texture samples.

The SDS scales were constructed along a seven-point scale in which positions from smaller to greater on both opposite affective words direction which splitting by a neutral value in the middle (Table 3.6.2).

The adjectives derived from the affinity diagram and some of the adjectives were modified slightly when adapted into the SDS and LS scale questionnaires, where the format was developed based on theory in assessing texture features as aesthetic, physical, operational and sensation traits (Lokman and Nagamachi, 2010).

For each texture sample, the participant was asked to decide which of the two words best describes their feelings about the texture sample if it were used in the design of the vehicle interiors. The items were printed out in random order.

The SDS scales were structured using a pair of opposite adjectives which had been formulated based on AE (Mitsuo Nagamachi, 2011), which was initially developed by Charles E. Osgood (Osgood et al., 1957) for measuring the magnitude of connotative meaning of adjective pairs.

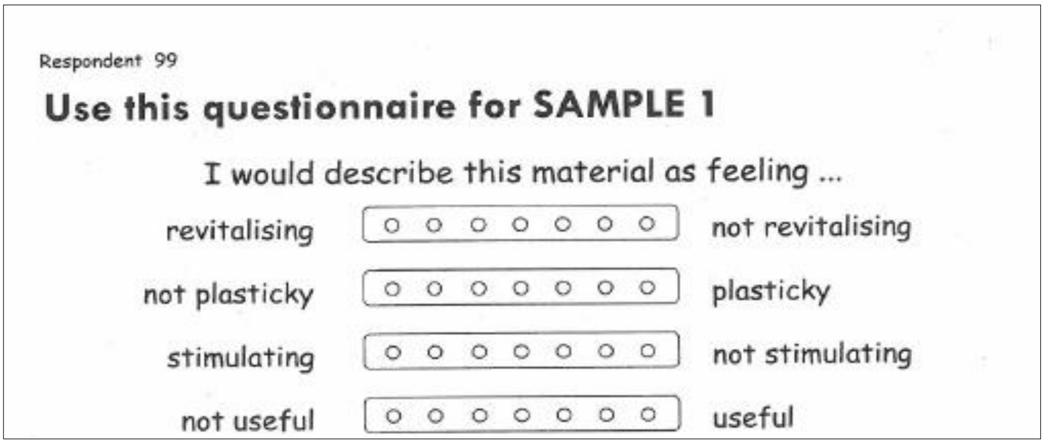
Twenty items were administrated to evaluate the texture samples for SDS1 study, as shown in Table 3.6.1.

Table 3.6.1 Original pool of items for SDS1 study

Code	Descriptions	Code	Descriptions
I01	Abrasive	I11	Plasticky
I02	Bumpy	I12	Revitalising
I03	Comfortable	I13	Rugged
I04	Easy-Handling	I14	Safe
I05	Edgy	I15	Sporty
I06	Effortless	I16	Stimulating
I07	Enjoyable	I17	Stunning
I08	Eye-catching	I18	Timeless Design
I09	Grippy	I19	Useful
I10	Nice-Quality	I20	Velvety

Table 3.6.2: Coding for seven-point SDS1 scales

Strongly Agree	Agree	Partly Agree	Neutral	Partly Agree	Agree	Strongly Agree
0	1	2	3	4	5	6


Respondent 99

Use this questionnaire for SAMPLE 1

I would describe this material as feeling ...

revitalising	o o o o o o o	not revitalising
not plasticky	o o o o o o o	plasticky
stimulating	o o o o o o o	not stimulating
not useful	o o o o o o o	useful

Figure 3.6.2 Seven-point bipolar SDS1 items

3.6.3.2 Building the LS1 Scale Items

The LS1 scale questionnaire was made up of statements on a five-point scale that was originally developed by Rensis Likert (Likert and Likert, 1976). An LS scale using five levels of evaluation value can be treated as a numerical value at least on the ordinal scales. Twenty statements were replicated randomly across seven pages equal to seven stimuli or texture samples. Some of the statements were modified slightly to better suit the method of LS scales and the research context.

As in the similar SDS scale study, for each texture sample, participants were asked to decide which linguistic quantifier was most suitable to describe how they felt about the texture sample if it was used to design the vehicle interiors.

The LS scale typically established by convention consists of a five-point scale that represents five response categories within two degrees disagreement to the left and two degrees agreement to the right, split by a neutral point. This is distinguished by a linguistic quantifier described as strongly disagree, disagree, neutral, agree and strongly agree (Table 3.6.4).

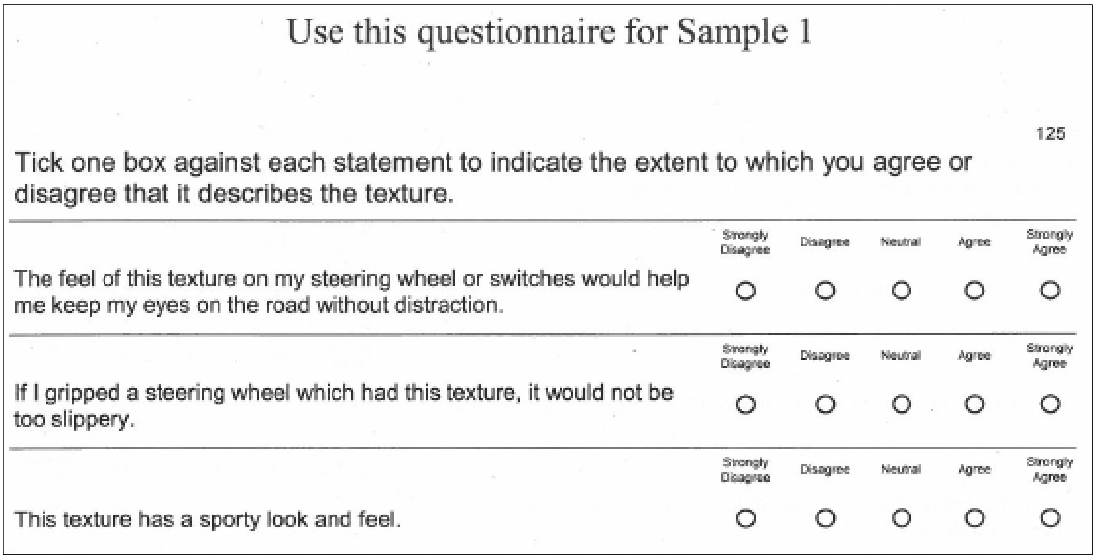
The LS1 was constructed with equal distance threshold values to represent magnitude estimation on a continuum endorsing particular items. In RUMM2030®, the response categories were therefore coded where the greater point would represent a higher degree of agreement while the smaller point indicates a higher degree of disagreement.

Table 3.6.3 Original pool of items for LS1 study

Code	Descriptions
01	Before I touch this texture, I can see that it would feel **grippy.**
02	I have the impression that this texture would make my car feel spacious and **neat**.
03	I have the impression this texture is modern and contemporary-**looking**.
04	I would expect to see this texture with a good touch and feel in a reasonable **price car**.
05	If I gripped a steering wheel which had this texture, it would feel very **safe**.
06	If I gripped a steering wheel which had this texture, it would not be too **slippery**.
07	The feel of this texture on my steering wheel or switches would help me to keep my eyes on the road without **distraction**.
08	The feel of this texture would help me feel confident with my **driving**.
09	The look of this texture makes me want to touch it straight away.
10	This texture has a sporty look and **feel**.
11	This texture looks nice **quality**.
12	This texture does not look overly cheap and **plasticky**.
13	Touching this texture feels **pleasant.**
14	Touching this texture is **relaxing**.
15	Touching this texture makes me feel **warm**.
16	Touching this texture would make me feel connected when operating the switches in the **vehicle**.
17	Vehicle controls with this texture would give good feedback when shifting, pulling, turning and **rotating**.
18	When I touch this surface, I get a sensation of **luxury.**
19	With this texture, I would be able to operate the controls without needing to **look**.
20	With this texture, I would feel comfortable inside **the car**.

Table 3.6.4 Coding for five-point LS1 scales

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
0	1	2	3	4


Use this questionnaire for Sample 1

125

Tick one box against each statement to indicate the extent to which you agree or disagree that it describes the texture.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
The feel of this texture on my steering wheel or switches would help me keep my eyes on the road without distraction.	O	O	O	O	O
If I gripped a steering wheel which had this texture, it would not be too slippery.	O	O	O	O	O
This texture has a sporty look and feel.	O	O	O	O	O

Figure 3.6.3 Five-point LS1 scales items

3.6.3.3 Data analysis

The data were analysed using the Rasch model software Professional RUMM2030® software from RuMM Laboratory Pty Ltd (Andrich, Sheridan and Luo, 2012). RUMM2030® is a software which was developed similarly like Winsteps – Copyright Linacre@2011 to evaluate the person and item of dichotomous, polytomous items, multiple choice, multiple rating scales, partial credit and pair comparisons in according to Rasch Model Theory.

The observed and calibrated data for the SDS1 and LS1 scales study were analysed using Rasch-facet design. Original data collection was administrated using an Excel spreadsheet (Microsoft Corporation), and PCA analysis was performed using an Excel spreadsheet and analysed using SPSS version 21 (IBM® SPSS, IBM Corp, Armonk, NY, USA).

3.7 RESULTS OF SDS1 STUDY

3.7.1 PCA Analysis

A PCA analysis was used in this study for dimensionality reduction in SDS space. This is a multivariate statistical procedure used to reduced dimensions of the adjectives that account for the maximum amount of variance underlying a dataset while maintaining the most important variance.

Table 3.7.1 shows the rotated component matrix in PCA for SDS1 study using Varimax rotation with Kaiser normalisation. Based on the PCA analysis, a maximum variance of orthogonal rotation was adopted. Three factors were obtained with an eigenvalue greater than one with a total loading value of the observed variables.

The significant Kaiser-Meyer-Olkin (KMO) value was >0.8, and for individual items it was considered meritorious, this value indicating the most important variance of material attributes have a significant impact on participants' overall perception (Claudia Newton, 2017). In AE structure, the remaining attributes with much less impact in probability can be removed (Lokman, 2010b).

A preliminary analysis was performed in order to obtain the eigenvalues for each factor in the data that were obtained from the scree plot.

In AE study, PCA analysis is often associate with semantic space to strategies the dimension reduction because qualitative studies in AE often have some redundancy issues of variables due to subjective perception (Lokman, 2010a). The redundancy means some of the attributes are correlated or carry similar meaning with other attributes. Thus, PCA was used as a dimension reduction technique to provide a smaller set of a variable. The important variable was calculated based on higher factor loading that accounts in PCA1, PCA2 and PCA3 according to evaluative, potency and activity (EPA) dimensions.

The factor loadings in a rotated component matrix, among the twenty-item variables, eight items or factor were identified as highlighted and loaded into PCA1, denoted as evaluative such as *abrasive, bumpy, edgy, eye-catching, grippy. rugged, sporty and velvety.* While six items were loaded into PCA2 and denoted as potency include *comfortable, easy-handling, effortless, plasticky, safe and useful.* Four items were loaded into PCA3 and denoted as an activity such as *enjoyable, nice-quality, revitalising and stunning* which represented the dimension of rating responses in semantic space (Osgood, 1971).

Table 3.7.1 Rotated component matrix for SDS1 study according to EPA dimensions

Rotated component matrix[a]

	Component		
	PCA1	PCA2	PCA3
Dimension	Evaluative	Potency	Activity
Abrasive	.966	-.137	-.101
Bumpy	.872	.396	-.129
Comfortable	-.514	.736	.349
Easy-Handling	.276	.847	.384
Edgy	.813	-.573	.035
Effortless	-.280	.814	.362
Enjoyable	.083	.506	.809
Eye-catching	.726	-.210	.610
Grippy	.836	.501	.192
Nice-Quality	-.247	.207	.946
Plasticky	.166	-.968	-.052
Revitalising	.278	.245	.914
Rugged	.962	.164	.116
Safe	.345	.893	.201
Stimulating	.687	.241	.585
Sporty	.895	-.124	.318
Stunning	.109	.254	.940
Timeless Design	-.570	.444	.666
Useful	.288	.856	.372
Velvety	-.954	.024	.157

Extraction method: Principal components analysis.
Rotation method: Varimax with Kaiser Normalisation.

Figure 3.7.1 illustrates scatter plot loading based on factors PCA1, PCA2 and PCA3 vector showed the location of texture sample two was chosen as the highest component score amongst the seven texture stimuli, while texture sample one was chosen as the lowest component score.

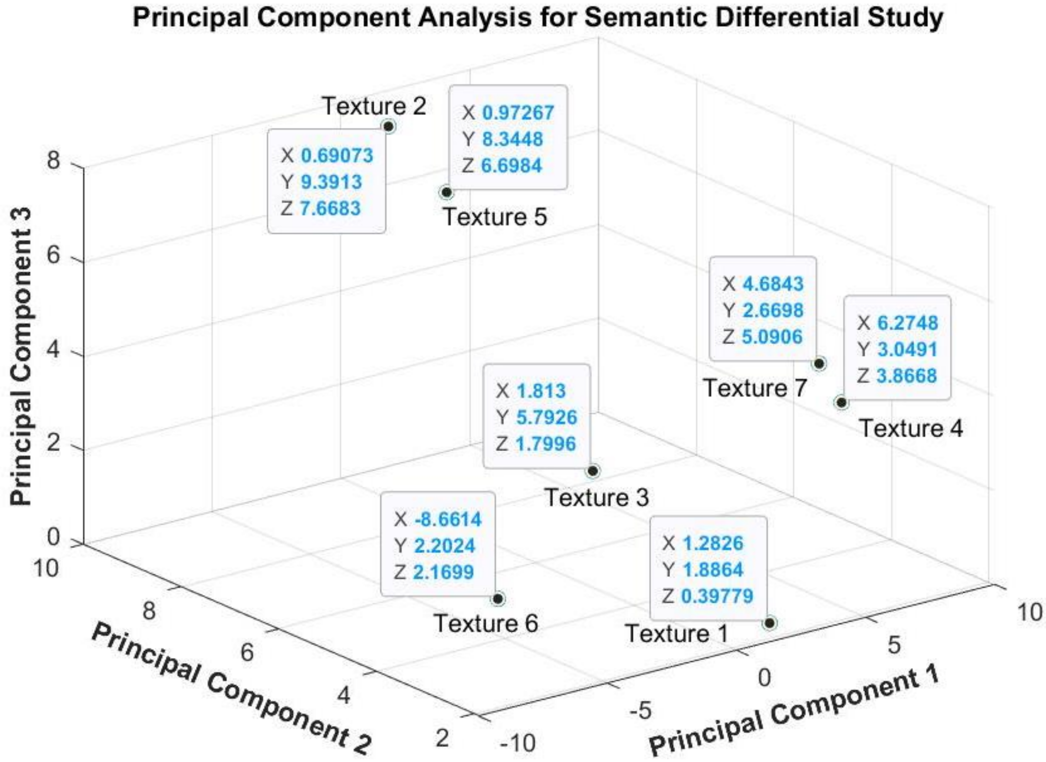


Figure 3.7.1 Scatter plot illustrating PCA composition in SDS1 study

The scatter plot is plotted based on product matrix which was calculated using binary operation though *MMULT function* in Excel to calculate the product matrix by multiplying two elements from the PCA rotated component matrix and the observed data matrix in each texture sample to produce combine element as shown in Table 3.7.2.

Table 3.7.2 PCA were plotted based on combined product matrix

	X-Axis	Y-Axis	Z-Axis
Texture	PCA 1	PCA 2	PCA 3
Texture 1	1.28262257	1.886409	0.3977859
Texture 2	0.69073007	9.391289	7.6683153
Texture 3	1.8129814	5.792587	1.7995995
Texture 4	6.27482507	3.049115	3.8668381
Texture 5	0.9726696	8.344766	6.6984469
Texture 6	-8.6613656	2.202416	2.1699278
Texture 7	4.68434243	2.669784	5.090553

3.7.2 Preliminary result - Summary of statistics for SDS1 study

Summary of fit-statistics for the preliminary analysis examined the preliminary item pool of twenty and sample of seventy-five participants. The likelihood-ratio test indicates that the outcomes for all stimuli were significant *(p* <0.01).

Table 3.7.3 Preliminary facet design by location order

Item Facet				**Stimuli Facet**				**Metric**			
Item	**Locn**	**SE**	**FitRes**	**Stimulus**	**Locn**	**SE**	**Fit Res**	**X^2**	**df**	**p**	**PSI**
14	-0.333	0.08	0.446	Texture 2	-0.118	0.08	0.515	429	280	0.01	0.84
4	-0.238	0.08	0.475	Texture 5	-0.066	0.07	0.394				
19	-0.220	0.08	0.409	Texture 7	-0.053	0.06	0.408				
3	-0.158	0.07	0.409	Texture 4	-0.049	0.07	0.500				
10	-0.117	0.07	0.335	Texture 3	0.016	0.07	0.752				
7	-0.104	0.07	0.418	Texture 1	0.118	0.07	0.761				
9	-0.096	0.07	0.264	Texture 6	0.153	0.06	0.964				
12	-0.014	0.08	0.437								
8	-0.113	0.06	0.379								
15	-0.007	0.07	0.362								
17	-0.006	0.07	0.383								
16	0.038	0.06	0.629								
6	0.040	0.07	0.868								
18	0.040	0.06	0.760								
11	0.064	0.06	1.503								
13	0.113	0.06	0.651								
20	0.187	0.06	0.768								
5	0.190	0.06	0.993								
1	0.268	0.07	0.932								
2	0.368	0.07	1.096								

The summary of the preliminary facet locations for SDS1 study is presented in Table 3.7.3. The item facet exhibits twenty items in location order, while stimuli facet exhibits seven stimuli in location order.

Table 3.7.4 Summary of fit-statistics for the preliminary analysis for SDS1 study

Item Interaction				Person Interaction				Item-trait interaction		
Item location		Item-fit residual		Persons location		Persons-fit residual				
Mean	SD	Mean	SD	Mean	SD	Mean	SD	(df)	χ2	*p*
0.00	0.26	0.64	0.49	0.05	0.12	-0.00	3.93	280	429.12	<0.01
N=75								PSI index		0.84
								Alpha α		0.83

3.7.3 Power of fit statistics for SDS1 study

The overall model fit given by the Chi-Square probability showed a value of less than 0.01. By applying the Bonferroni adjusted alpha value of $p \leq 0.000357$ (0.05/140), the value indicates the degree of fit between the data and the model.

An item-trait interaction statistic, reported regarding Chi-Square, reflects the property of invariance across the trait. In this study, the Chi-Square value of $p \leq 0.01$ indicates a lack of the desired scale invariance where some of the items were not working as expected at group levels denoted to class intervals (Table 3.7.4 column *p*).

Chi-Square calculated item fit statistics based on person-item deviations and deviation by the ability level within the same group known as a class interval. The equal class interval of (26 + 25 + 24 = 75) indicates good interval distribution.

3.7.3.1 Reliability – Internal consistency

Initial fit statistics, as shown in (Table 3.7.4), demonstrate that the exploration of person fit which the power of test fit and the indicator of reliability as represented by the Cronbach's alpha and person separation index (PSI). The preliminary analysis of the person separation index was reported to be 0.84, indicating the internal consistency which the ability to statistically differentiate at least three ability groups. The Cronbach's alpha value was reported to be 0.83, which indicates a good level of reliability.

3.7.3.2 Item fit location and fit residuals

The results of this study show the item location mean was zero, and the SD was 0.26 (Table 3.7.4). These indicated that there was a reasonable fit of the items in the data as these values fall into standardised item-fit between approximately zero and an SD of approximately one.

The results of this study show the fit residual mean was 0.64 and the SD was 0.49. These indicated that there was a reasonable fit of the items in the data as these values fall into standardised item-fit within the conventionally accepted range of +/- 2.50.

3.7.3.3 Person fit location and fit residuals

The results of this study show the person location mean was 0.05, and the SD was 0.12 (Table 3.7.4). These indicated that there was a reasonable fit; however, for the exploration of person fit residual, the mean value was recorded as -0.00 and the SD was 3.93. This value associated some misfit where persons indicated a lack of the expected probabilistic relationship among the items within a scale.

Table 3.7.5 Summary of fit-statistics for the calibrated analysis SDS1 study

Item Interaction				Person Interaction				Item-trait interaction		
Item location		Item-fit residual		Persons location		Persons-fit residual				
Mean	SD	Mean	SD	Mean	SD	Mean	SD	(df)	$\chi 2$	*p*
0.00	0.48	0.58	0.45	0.05	0.15	-0.18	4.26	280	448.62	<0.01
*n*75								PSI index	0.84	
								Alpha α	0.83	

Calibrated fit statistics in Table 3.7.5 show the improvement results after the several adjustments to the source of misfit, the fit statistics resulting in the improvement in terms of item and person fit residual. The calibrated result has generally repositioned the difficult items to be easier to endorse when the item fit residual has resulting in a better mean location of 0.58 logit. The improvement of the person fit also resulted in the better mean location of person ability when endorsing the difficult items that have moved to -0.18 logit. However, for the exploration of the person fit residual, the SD was 4.26, indicating this value remains a misfit.

3.8 BIAS IN SDS1 STUDY

The SDS1 study demonstrates five biases;

1. Lack of invariance
2. Lack of the expected probabilistic targeting
3. High fit residual of person-items
4. Large number of items holds disordered thresholds
5. Items remain disordered after calibrations

3.8.1 Lack of invariance

The item-trait interaction statistics reported a Chi-Square value of $p \leq 0.01$, which indicates a lack of the desired scale invariance, where some of the items are not working as expected at group levels denoted to class intervals.

3.8.2 Lack of expected probability targeting

Figure 3.8.1 shows the person-item distributions illustrate person correlation in association with the difficulty of the items. The items were equally distributed to all level of difficulty that were significant in Rasch analysis. The initial person-item distribution graph in the upper part showed that the majority of the participants are well targeted in the set of calibrated items. Although the graph indicates the thresholds hold a good spread, however, approximately half of the items were difficult to answer for a large number of participants. The calibrated results exhibit a very small improvement on the distribution thresholds bar where items below the floor increased compared with the initial results. The calibrated results showed a mean location of 0.05 and a SD of 0.15.

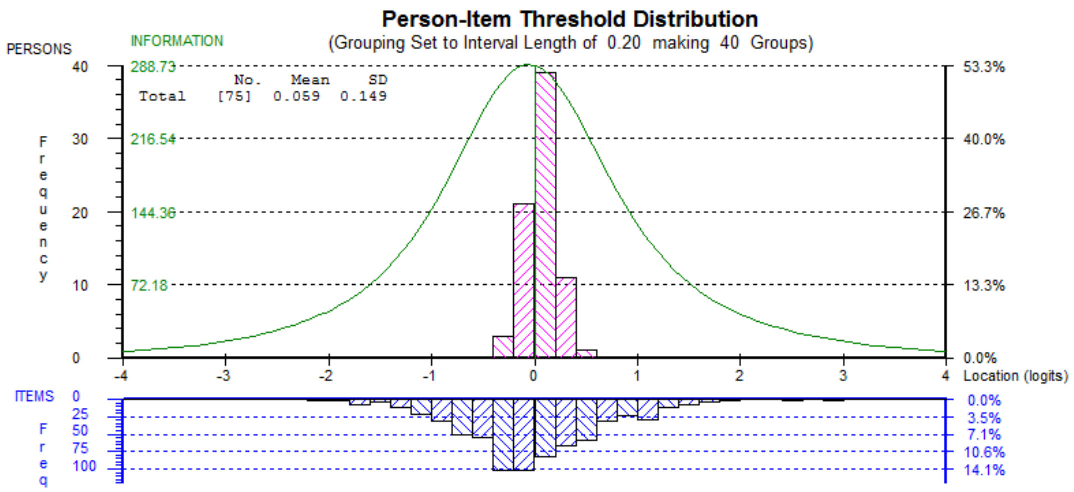


Figure 3.8.1: Calibrated person-item threshold distribution for SDS1 study

3.8.3 The high fit residual of person-items

A person holding greater fit residuals can misfit the model, in which the individual response patterns may exhibit as unfit if they are unexpected or contain too much dependence. As a reference, values fall into standardised item-fit within the conventionally accepted range of +/- 2.50.

In the preliminary analysis in the SDS1 study, approximately thirty-five persons were identified with higher fit residuals. Approximately eighteen participants were recorded before calibrations as unfit, consuming a higher positive fit residual range between 3.180 and 9.767 logit (Table 3.8.1) while seventeen participants were recorded before calibrations as consuming higher negative fit residual range between -2.615 and -12.342 (Table 3.8.2). These values affect the overall person fit residual aggregate of -0.18 and SD of 4.26 logit (Table 3.7.4).

Table 3.8.1 Persons with positive high fit residual

No	Person ID	Location	SE	FitResid	
1	7	0.087	0.049	9.767	Upper
2	9	0.106	0.049	8.563	
3	41	0.048	0.048	7.963	
17	69	0.007	0.048	3.277	
18	58	-0.043	0.047	3.180	Lower

Table 3.8.2 Persons with negative high fit residual

No	Person ID	Location	SE	FitResid	
1	73	0.041	0.048	-12.342	Upper
2	29	0.039	0.048	-7.757	
3	20	-0.083	0.047	-7.35	
16	15	0.379	0.056	-2.731	
17	64	-0.071	0.047	-2.615	Lower

3.8.4 A large number of items holds disordered thresholds

The investigation of disordered thresholds for SDS1 study was carried out. Overall initial analysis before calibrations showed 94.28 percent or one hundred and thirty-two out of one hundred and forty items hold disordered thresholds (Table 3.8.3).

Table 3.8.3 Items hold disordered thresholds after adjustment in SDS1 study

Stimuli	An item with ordered thresholds	An item with disordered thresholds	Items remain disordered after adjustments
Texture 1	12 and 14	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 13, 15, 16,17,18,19 and 20	1, 2, 8, 10, 13, 16, 17 and 20
Texture 2	12	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 13, 14, 15, 16, 17, 18, 19 and 20	3, 4, 5, 7, 8, 11, 13, 14, 16, 18 and 19
Texture 3	4 and 19	1, 2, 3, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18 and 20	2, 6, 7, 9, 11 and 20
Texture 4	12	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 13, 14, 15, 16, 17, 18, 19 and 20	1, 2, 3, 4, 5, 6, 8, 9, 10, 14, 16, 17, 18 and 19
Texture 5	Nil	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19 and 20	1, 3, 7, 8, 9, 10, 11, 14, 15, 16, 18, 19 and 20
Texture 6	19	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18 and 20	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 13, 15, 18 and 20
Texture 7	12	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 13, 14, 15, 16, 17, 18, 19 and 20	5, 6, 7, 8, 10, 11, 15, 16, 18 and 20
Total	8 Items	132 Items	77 Items

The double asterisks (**) in thresholds map indicated the items hold disordered thresholds (Figure 3.8.2). The explanation of these item descriptions refers to the original pool of items for SDS1 study (Table 3.6.1)

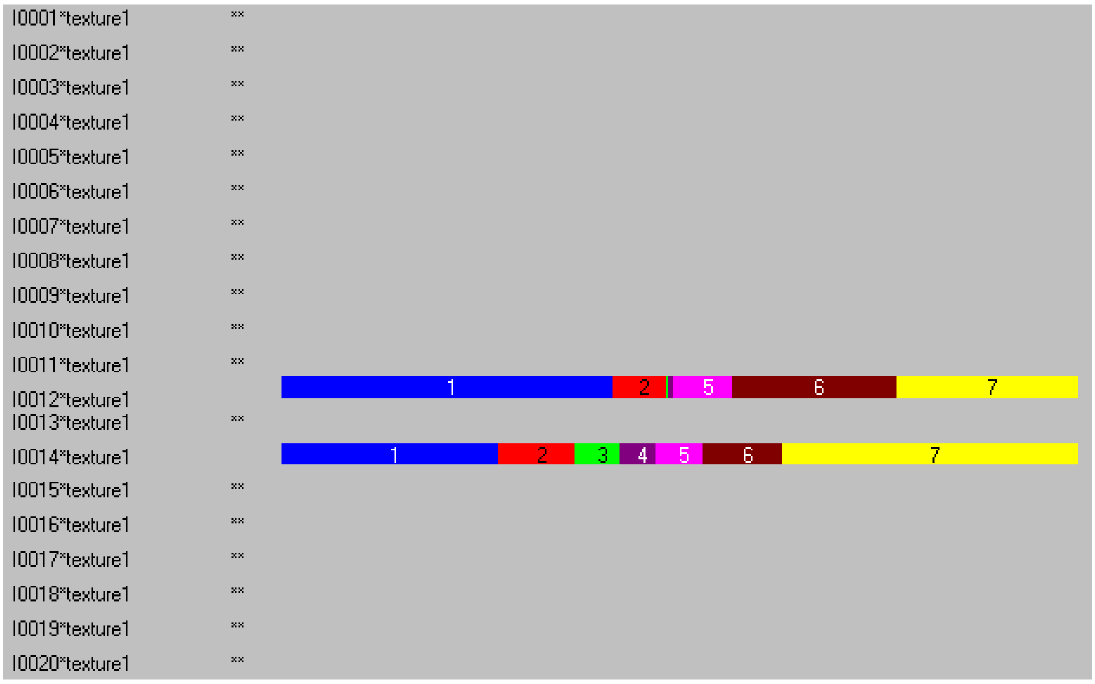


Figure 3.8.2 Initial analysis of disordered thresholds before calibrations for SDS1 study

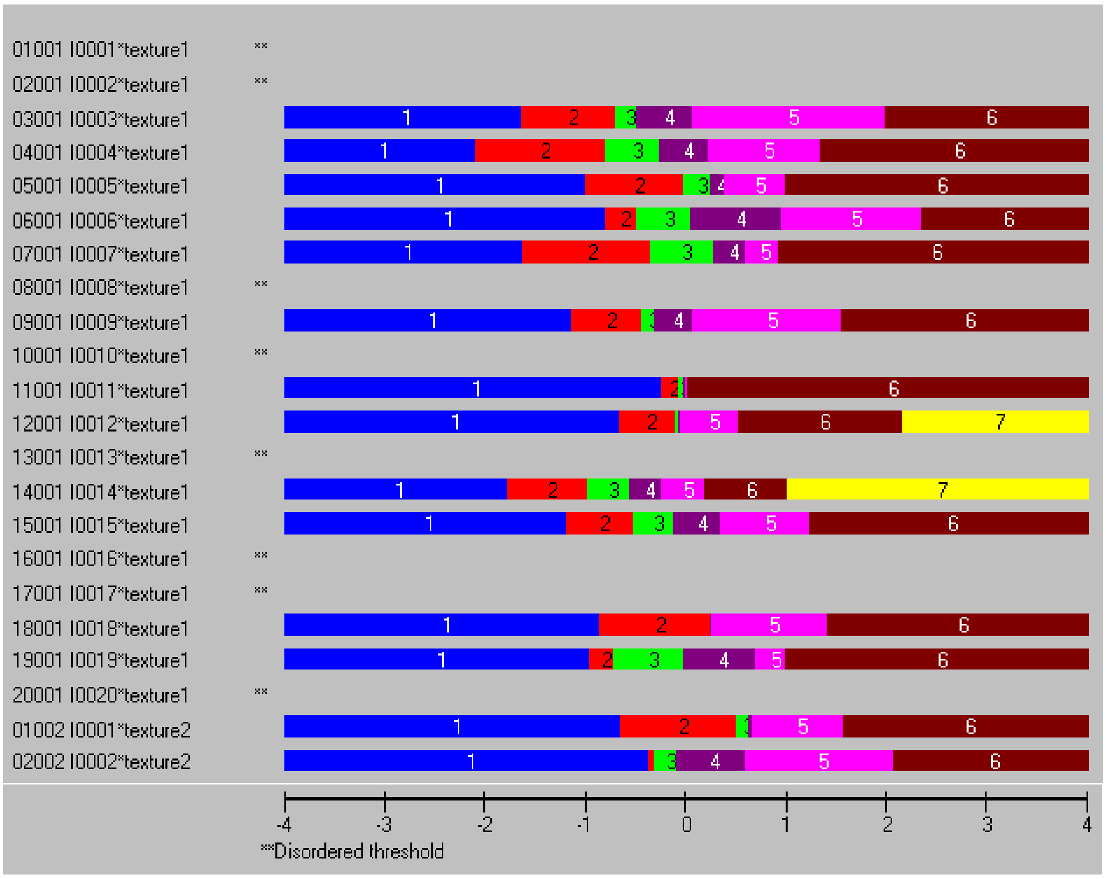


Figure 3.8.3: Calibrated analysis of ordered thresholds map after calibrations for SDS1 study

3.8.5 Items remain disordered after adjustments

Disordered thresholds on category scales indicate the response options do not work in the way that was initially designed, meaning some of the category scales were over or under discriminated between adjacent category scales. It may be speculated that the overlapping thresholds are associated with some of difficulty items that are not working using SDS scales.

RM statistics allowed the disordered thresholds to be rescored for improving the response pattern. However, in SDS analysis, the calibrations process did not perform well during the rescoring works.

The fourth column in (Table 3.8.3) demonstrates one hundred and thirty-two items holding disordered thresholds were able to be rescored into ordered thresholds after a few adjustments. Although several calibrations were successfully made, large numbers of disordered thresholds do not degrade where seventy-seven items remain disordered.

This indicates that the category structure does not work in the way that it was designed, especially when the structure is dealing with a certain level of difficulty items.

3.9 RESULT OF LS1 STUDY

3.9.1 PCA analysis

A PCA analysis was used in this study for dimensionality reduction in semantic space. This is a multivariate statistical procedure used to reduce the dimensions of the statements that account for the maximum amount of variance underlying a dataset while maintaining the most important variance.

Table 3.9.1 shows the rotated component matrix for twenty items for LS study using Varimax rotation with Kaiser normalisation. The variables items have simplified code from the last words, which have been highlighted in bold as in the original pool items (Table 3.6.3). For example, item 1 was coded as ‘grippy’ derived from item one in the LS statement in the original pool items.

Based on the PCA analysis, a maximum variance of orthogonal rotation was adopted. Three factors were obtained with an eigenvalue greater than one with a total loading value of the observed variables.

The significant KMO was >0.8, and for individual items was considered meritorious, this value indicating the most important variance of material attributes have a significant impact on participants’ overall perception (Claudia Newton, 2017). In AE structure, the remaining attributes which have much less impact can probably be removed (Lokman, 2010b).

In AE study, PCA analysis is often associate with semantic space to strategies the dimension reduction because qualitative studies in AE often have some redundancy issues of variables due to subjective perception. The redundancy means some of the attributes are correlated or carry similar meaning with other attributes (Lokman, 2010b). Thus, PCA was used as a dimension reduction technique to provide a smaller set of a variable. The important variable was calculated based on higher factor loading that accounts in PCA1, PCA2 and PCA3.

The factor loadings in a rotated component matrix, among the twenty-item variables, eight items or factor were identified as highlighted and loaded into PCA1, denoted as evaluative such as *grippy, safe, slippery, distraction, driving, vehicle, rotating and look*. While nine items were loaded into PCA2 and denoted as potency include *neat, price car, quality, plasticky, pleasant, relaxing, warm, luxury and the car*. Two items were loaded into PCA3 and denoted as an activity, are *looking and away* which represented the dimension of rating responses in semantic space (Osgood, 1971).

Table 3.9.1 Rotated component matrix according to EPA dimensions

Rotated component matrix[a]

	Component		
	PCA1	PCA2	PCA3
Dimension	Evaluative	Potency	Activity
Q1 Grippy	.963	-.107	.226
Q2 Neat	-.375	.884	.231
Q3 Looking	.128	.164	.973
Q4 Price car	.536	.731	.375
Q5 Safe	.963	.233	.127
Q6 Slippery	.981	.083	.120
Q7 Distraction	.959	.258	.019
Q8 Driving	.912	.318	.188
Q9 Away	.518	.269	.787
Q10 Feel	.629	-.203	.686
Q11 Quality	.417	.807	.371
Q12 Plasticky	.326	.908	.178
Q13 Pleasant	.341	.828	.315
Q14 Relaxing	-.090	.940	-.226
Q15 Warm	.297	.919	-.173
Q16 Vehicle	.959	.137	.212
Q17 Rotating	.968	.190	.158
Q18 Luxury	-.062	.945	.292
Q19 Look	.970	.121	.206
Q20 The car	.303	.936	-.172

Extraction method: Principal components analysis
Rotation method: Varimax with Kaiser Normalisation

Figure 3.9.1 illustrates scatter plot loading based on factors PCA1, PCA2 and PCA3 vector showed the location of texture sample two was chosen as the highest component score amongst the seven texture stimuli, while texture sample one was chosen as the lowest component score.

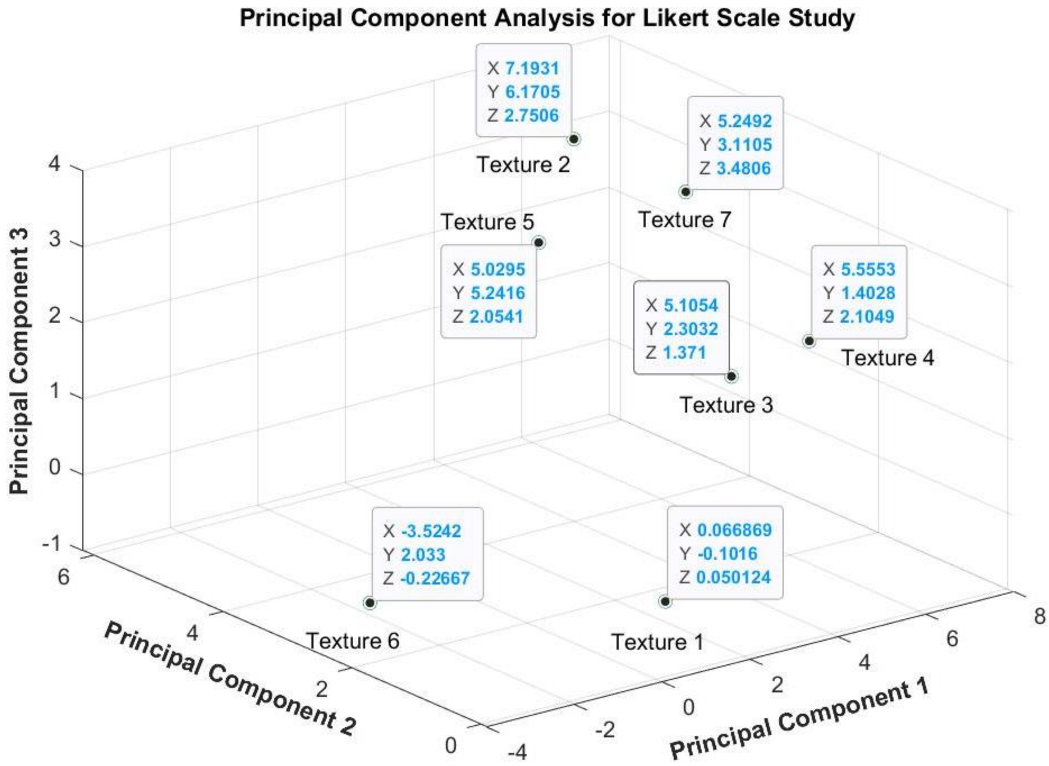


Figure 3.9.1 Scatter plot illustrating PCA composition in LS1 study

The scatter plot is plotted based on product matrix which was calculated using binary operation though *MMULT function* in Excel to calculate the product matrix by multiplying two elements from the PCA rotated component matrix and the observed data matrix in each texture sample to produce combine element as shown in Table 3.9.2.

Table 3.9.2 PCA were plotted based on combined product matrix

	X-Axis	Y-Axis	Z-Axis
Texture	PCA 1	PCA 2	PCA 3
Texture 1	0.066868809	-0.101597664	0.0501238
Texture 2	7.19305409	6.170509066	2.7505757
Texture 3	5.105355247	2.303237139	1.3710121
Texture 4	5.555320925	1.402838511	2.1049481
Texture 5	5.029517228	5.241626581	2.0541454
Texture 6	-3.524155684	2.03295331	-0.226669
Texture 7	5.249188664	3.110496981	3.4806065

3.9.2 Preliminary result - Summary of statistics for LS study

Summary of fit-statistics for the preliminary analysis examined from the preliminary item pool of twenty items and sample of one hundred and seven persons. The likelihood-ratio test indicates that the outcomes for all stimuli were significant *(p* <0.01*).*

Table 3.9.3 Preliminary facet design by location order

Item Facet				**Stimuli Facet**				**Metric**			
Item	**Locn**	**SE**	**FitRes**	**Stimulus**	**Locn**	**SE**	**Fit Res**	**X^2**	**df**	**p**	**PSI**
20	-0.165	0.10	0.162	Texture 2	-0.274	0.10	0.374	429	280	0.01	0.83
17	-0.152	0.10	0.243	Texture 5	-0.165	0.10	0.320				
11	-0.113	0.09	0.479	Texture 7	-0.159	0.09	0.452				
13	-0.109	0.10	0.279	Texture 3	-0.058	0.10	0.035				
1	-0.086	0.09	0.747	Texture 4	-0.031	0.09	0.655				
4	-0.085	0.10	0.513	Texture 6	0.340	0.09	1.000				
8	-0.037	0.10	0.107	Texture 1	0.348	0.09	0.326				
3	-0.026	0.09	0.719								
2	-0.023	0.10	0.343								
7	-0.003	0.10	0.457								
9	0.003	0.09	0.523								
16	0.028	0.10	0.385								
19	0.028	0.10	0.219								
6	0.039	0.09	0.552								
12	0.044	0.09	0.872								
5	0.051	0.10	0.272								
14	0.067	0.10	0.131								
10	0.113	0.09	0.724								
15	0.182	0.10	0.376								
13	0.243	0.08	0.943								

The summary of the preliminary facet locations for LS study is presented in Table 3.9.3. The item facet exhibits twenty items in location order, while stimuli facet exhibits seven stimuli in location order.

Table 3.9.4 Summary of fit-statistics for the preliminary analysis

Item Interaction				Person Interaction				Item-trait interaction		
Item location		Item-fit residual		Persons location		Persons-fit residual				
Mean	SD	Mean	SD	Mean	SD	Mean	SD	(df)	$\chi 2$	*p*
0.00	0.32	0.45	0.53	0.22	0.33	-0.30	3.60	280	336.68	0.01
N = 107								PSI index		0.94
								Alpha α		0.93

3.9.3 Power of fit statistics

The overall model fit given by the Chi-Square probability showed a value of less than 0.01. By applying the Bonferroni adjusted alpha value of $p \leq 0.0003571$ (0.05/140), the value indicated the degree of fit between the data and the model.

An item-trait interaction statistic, reported as Chi-Square, reflects the property of invariance across the trait.

In this study, the Chi-Square value of $p \leq 0.01$ indicates a lack of the desired scale invariance, where some of the items are not working as expected at group levels denoted to class intervals.

Chi-Square calculated item fit statistics based on person-item deviations and deviation by the ability level within the same group are known as a class interval. The equal class interval of (36 + 36 + 35 = 107) indicates good interval distribution.

3.9.3.1 Reliability – internal consistency

The initial fit statistics as shown in Table 3.9.4 demonstrated the exploration of person fit which the power of test fit was the indicator of reliability and was represented by the Cronbach's alpha and person separation index. The preliminary analysis of the person separation index was reported as 0.94 and illustrated the internal consistency of the ability to statistically differentiate at least three ability groups. The Cronbach's alpha value was reported to be 0.93, which indicates a good level of reliability.

3.9.3.2 Item fit and fit residuals

The results of this study show the item location mean was zero and the SD was 0.32 (Table 3.9.4). These indicate that there was a reasonable fit of the items in the data, as these values fall into standardised item-fit between approximately zero and a SD of approximately one.

While the result of this study shows the fit residual mean was 0.45 and the SD was 0.53. These indicate that there was a goodness of fit of the items in the data as these values fall into a standardised item-fit within the conventionally accepted range of +/- 2.50.

3.9.3.3 Person fit and fit residuals

The results of this study show the person location mean was 0.22, and the SD was 0.33 (Table 3.9.4). These indicated that there was a reasonable fit. However, for the exploration of person fit residual, the mean value was recorded was -0.30 and the SD was 3.60. This value is associated with some misfit where persons indicate a lack of the expected probabilistic relationship among the items within a scale.

Table 3.9.5 Summary of fit-statistics for the calibrated analysis

Item Interaction				Person Interaction				Item-trait interaction		
Item location		Item-fit residual		Persons location		Persons-fit residual				
Mean	SD	Mean	SD	Mean	SD	Mean	SD	(df)	χ2	*p*
0.00	0.41	0.30	0.50	0.27	0.43	-0.55	3.77	280	352.10	<0.01
N = 107								PSI Index		0.95
								Alpha α		0.93

Calibrated fit statistics in Table 3.9.5 show the improvement in results after several adjustments to the source of the misfit, and the fit statistics have resulted from the improvement in terms of the item and person fit residual. The calibrated result has generally repositioned the difficult items to become easier to endorse with the item fit residual achieving a better result mean location of -0.30 logit. The improvement of the person fit has also resulted in a better mean location of person ability when endorsing the difficult items that have moved to -0.55 logit. However, for the exploration of person fit residual, the SD was 3.77, indicating this value remain a misfit.

3.10 BIAS IN LS STUDY

The LS study has demonstrated five bias outcomes that affect the biases in rating scales, as evidenced by the following evidence:

1. Lack of invariance
2. Lack of the expected probabilistic targeting
3. The high fit residual of person-items
4. A large number of disordered thresholds
5. Items remain disordered after adjustments

3.10.1 Lack of invariance

The item-trait interaction statistics reported the Chi-Square value of $p \leq 0.01$ indicates a lack of the desired scale invariance, where some of the items are not working as expected at group levels denoted to class intervals.

3.10.2 Lack of expected probability targeting

Figure 3.10.1 shows the person-item distributions illustrate person correlations in association with the difficulty of the items. The items were equally distributed to all level of difficulties that were significant in Rasch analysis.

The initial person-item distribution graph in the upper part showed that the majority of the participants are well targeted in the set of calibrated items. Although the graph indicates the thresholds hold a good spread, however, approximately half of the items were difficult to answer for a large number of participants.

The calibrated results exhibit a minimal improvement in the distribution thresholds bar where items below the floor increased compared with the initial result. The calibrated results showed a mean location of 0.06 and a SD of 0.43.

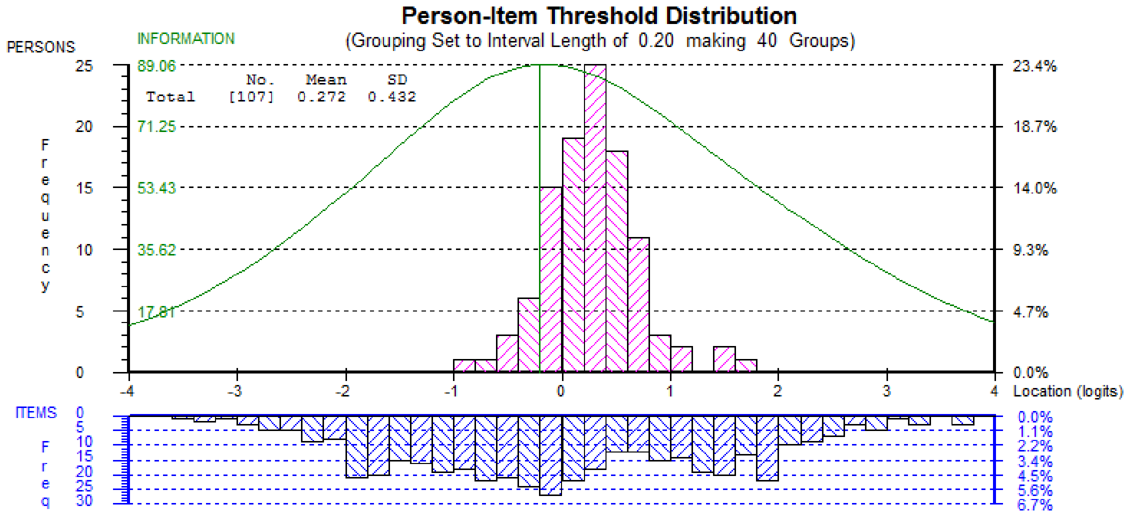


Figure 3.10.1: Calibrated person-item thresholds distribution

3.10.3 The high fit residual of person-items

A person holding greater fit residuals can misfit the model, in which the individual response patterns may exhibit as unfit if they are unexpected or contain too much dependence. As a reference, values fall into standardised item-fit within the conventionally accepted range of +/- 2.50.

Table 3.10.1 Persons with positive high fit residual

No	Person ID	Location	SE	FitResid	
1	22	0.862	0.097	9.098	Upper
2	40	-0.027	0.079	9.019	
3	32	0.246	0.081	7.564	
19	28	0.116	0.080	2.730	
20	38	0.686	0.091	2.603	Lower

Table 3.10.2 Person with negative high fit residual

No	Person ID	Location	SE	FitResid	
1	35	-0.003	0.079	-9.020	Upper
2	88	0.340	0.083	-7.920	
3	59	0.104	0.080	-7.334	
23	58	0.246	0.081	-2.813	
24	37	0.135	0.08	-2.736	Lower

Preliminary analysis of the LS study shows approximately forty-four persons were identified with higher fit residual. Twenty participants were recorded before calibrations consuming a higher positive fit residual range between 2.603 to

9.098 logit (Table 3.10.1) while twenty-four participants were recorded before calibrations consume higher negative fit residual range between -2.736 to -9.020 logits (Table 3.10.2). These values affect the overall person fit residual aggregate of -0.30 and SD of 3.60 logits (Table 3.9.4).

3.10.4 A large number of items holds disordered thresholds

The investigation of disordered thresholds for LS study was carried out. Overall initial analyses before calibrations showed that 80 percent or one hundred and twelve out of one hundred and forty items had a disordered threshold (Table 3.10.3).

Table 3.10.3 Items holding disordered thresholds after adjustment in LS1 study

Stimulus	An item with Ordered Thresholds	An item with Disordered Thresholds	Items Remain Disordered After Adjustments
Texture 1	7, 8, 9 and 16	1, 2, 3, 4, 5, 6, 10, 11, 12, 13, 14, 15, 17, 18, 19 and 20	All items hold ordered thresholds
Texture 2	2, 7, 10, 15, 16 and 18	1, 3, 4, 5, 6, 8, 9, 11, 12, 13, 14, 17, 19 and 20	All items hold ordered thresholds
Texture 3	2, 5, 7 and 16	1, 3, 4, 6, 8, 9, 10, 11, 12, 13, 14, 15, 17, 18, 19 and 20	All items hold ordered thresholds
Texture 4	2, 7, 11, 13, 16 and 17	1, 3, 4, 5, 6, 8, 9, 10, 12, 14, 15, 18, 19 and 20	18
Texture 5	7, 10 and 16	1, 2, 3, 4, 5, 6, 8, 9, 11, 12, 13, 14, 15, 17, 18, 19 and 20	All items hold ordered thresholds
Texture 6	2	1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19 and 20	All items hold ordered thresholds
Texture 7	2, 7, 8 and 15	1, 3, 4, 5, 6, 9, 10, 11, 12, 13, 14, 16, 17, 18, 19 and 20	12
Total	28 Items	112 Items	2 Items

The double asterisks (**) in the thresholds map indicate that the items hold disordered thresholds (Figure 3.10.2). The explanation of these items descriptions refer to the original pool of items for LS study in Table 3.6.3.

3.10.5 Items remain disordered after adjustments

Disordered thresholds on category scales indicate the response options do not work in the way that was initially designed. Meaning some of the category scales were over or under discriminated between adjacent category scales. It may be speculated that the overlapping thresholds are associated with some of difficulty items which are not working using the LS scales.

RM statistics allowed the disordered thresholds to be rescored in order to improve the response pattern. The fourth column in Table 3.10.3 shows one hundred and eighteen items hold disordered thresholds that could be rescored after adjustments. Despite the LS calibrations process performing quite well during the rescoring works, there are a small number of disordered thresholds that do not completely degrade where two items remain disordered.

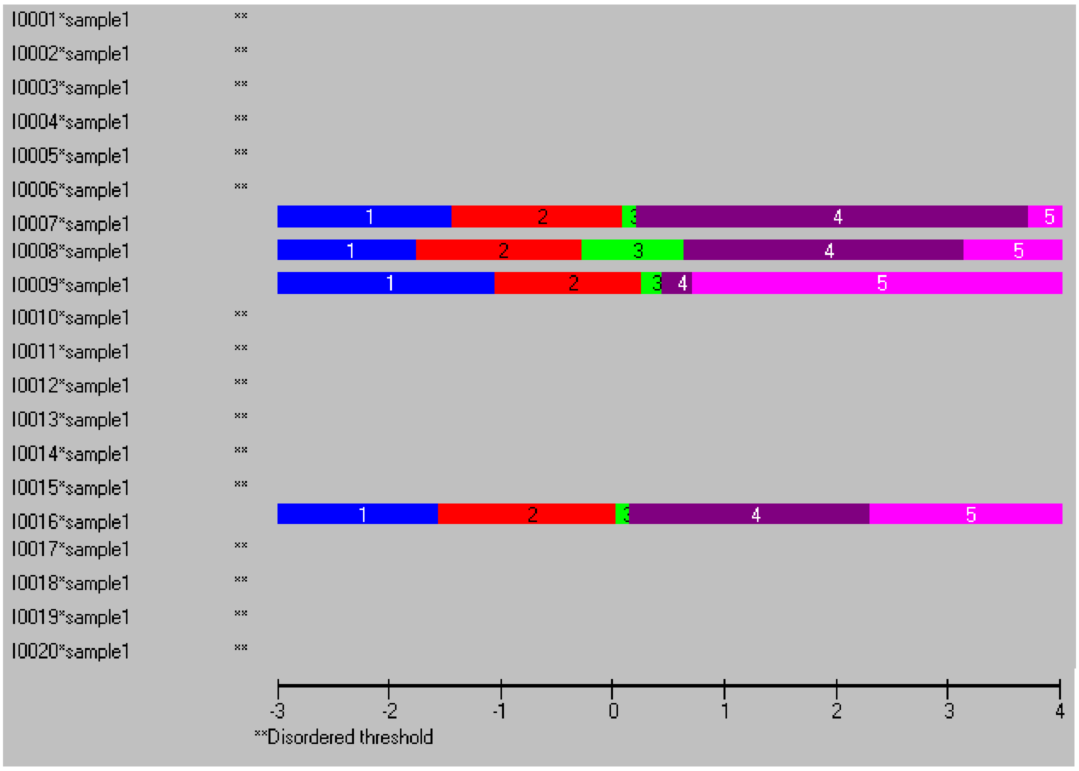


Figure 3.10.2: Initial analysis of disordered thresholds before adjustment for LS1 study

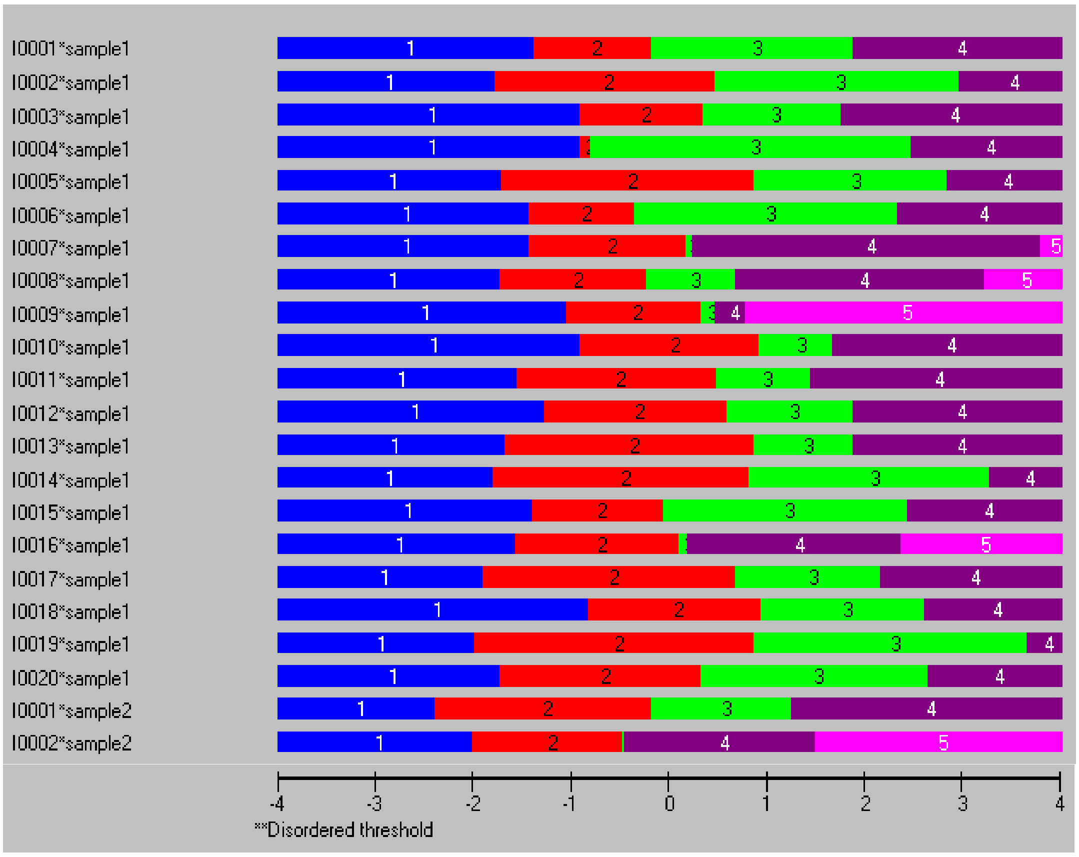


Figure 3.10.3: Calibrated analysis of ordered thresholds map after adjustment for LS1 study

3.11 DISCUSSIONS

3.11.1 Delimitation of the studies

The sample sizes for both studies were insufficient to carry out the Rasch analysis. The ideal sample size, as suggested by published guidelines, is approximately N=250 for ninety-nine per cent confidence that the item calibrations will be stable to within half a logit, even under a condition of poor-targeting. A sample over N=500 offers robust results (Horton, 2017).

The redundancy item can be tested using an independent T-test. This test will determine whether the item holds multidimensionality or unidimensionality. However, the unidimensionality test requires specific analysis through which a binomial test should validate it.

3.11.2 The absence of a unidimensionality test

In this chapter, the unidimensionality construct was not performed for the following reasons:

- One of the objectives in Chapter 3 is to determine the relative importance of properties associated with biases and errors to a greater degree, which often violate the statistical outcome. If the unidimensionality test was applied, the evidence could be restricted.
- The unidimensionality test does not affect the invariance with regards to biases and error, specifically in determining invariance, targeting, disordered threshold, item fit residuals and calibrations procedures.
- Nevertheless, the unidimensionality test is crucial to derive linearity in affective responses. For this reason, the unidimensional test was performed separately in Chapter 5.

3.11.3 Discussion of result of SDS1 and LS1 studies

The outcome of this study has tested and validated some of the literature evidence. This study summaries four evidences of bias and the likelihood of misfits suffering from bias. This evidence has reflected the objectives of this study as follows.

3.11.3.1 Lack of invariance

This considers how well the items fit into the model using a rating scale (SDS1 and LS1). The overall model fit given by the Chi-Square probability showed both studies are reasonable fits between data and the model. Despite having good linear measurement and reliability value indicating good internal consistency with the ability to statistically differentiate at least three ability groups, the item-trait interaction statistics reported the Chi-Square value of $p \leq 0.01$ indicates a lack of the desired scale invariance where some of the items are not working as expected at group levels denoted to class intervals.

3.11.3.2 Lack of the expected probabilistic targeting

The exploration of fit residuals has reported the mean and SD values associated with some misfits, where persons indicate a lack of the expected probabilistic relationship among the items within a scale known as the person item distribution.

Targeting is the terminology given to describe the relationship between a person and item distribution. Ideally, a good result requires the normal distribution or approximately equal distribution to achieve well-targeted performance (Horton, 2017). Poor targeting can happen when the distribution is skewed to the left or right degree. This indicates that the item is either too difficult to be endorsed by the person with too low ability or vice-versa.

3.11.3.3 Higher fit residuals

In the SDS1 and LS1 studies, a total of seventy-nine or 43.4 percent of participants from one hundred and eighty-two have reported higher fit residual or unfit extreme persons: thirty-five participants were recorded in the SDS1 study while forty-four participants were recorded in the LS1 study. High residual values indicate the participant's most likely response to the item was an unexpected pattern response. The statistical point of view suggests the group needs to be removed as it will affect the overall score aggregate in RUMM2030®.

In the SDS1 study, approximately eighteen participants were recorded before calibrations consuming a higher positive fit residual range between 3.180 and

9.767 logit, while seventeen participants were recorded before calibrations consuming a higher negative fit residual range between -2.615 and -12.342

While in the LS1 study approximately twenty participants were recorded before calibrations consuming the higher positive fit residual range of 2.603 to 9.098 logit, twenty-four participants were recorded before calibrations consuming the higher negative fit residual range between -2.736 and -9.020.

The statistical point of view suggests the group needs to be removed as it will affect the overall score aggregate in RUMM2030®.

3.11.3.4 Large number items hold disordered thresholds

In the SDS1 study, the preliminary results before calibrations showed 94.28 percent or one hundred and thirty-two out of one hundred and forty items hold disordered thresholds. While 80 percent or one hundred and twelve out of one hundred and forty items hold disordered thresholds in the LS1 study, this indicates a large number of adjacent category intervals will significantly impact on the weakened ability of the participants to understand, distinguish or discriminate between the adjacent category scales. Disordered thresholds occur when participants have difficulty consistently discriminating between response categories. This may usually happen when participants have too many response options or category scales that make them confusing or too complicated to choose between them (Hendriks et al., 2012). In the Rasch model, disordered thresholds, also known as reverse ordered, are effected by poor targeting into response options (Petrillo et al., 2015). The response of the structure appears to no longer function as intended (Horton and Perry, 2016).

3.11.4 Investigating person-item misfit

In the SDS1 and LS1 studies, there were some extreme persons identified based on the fit residual analysis. RUMM2030® software suggested the person-item fit residual greater than +/- 2.50 needs to be removed as these groups affect the overall aggregate in statistical analysis (Pallant and Tennant, 2007; Camargo, 2013). RM theory states that if the person fit residual value has fallen within the range of +/- 2.50, this indicates that the individual person is behaving in the same way as another person does when endorsing the survey.

This is important in AE research in that the harmonised value indicates a homogeneity factor among the participants that will result in higher precision measurement.

A similar analysis detected that some items have negative fit residuals which are generally associated with redundancy or over-discrimination of an item. This means some of the items are obtained through the same interpretation or meaning by the participants.

3.11.4.1 Removing extreme persons

In total, seventy-nine or 43.4 percent of participants out of one hundred and eighty-two reported a higher fit residual or unfit extreme person with thirty-five participants recorded in the SDS1 study and forty-four recorded in the LS1 study. High residual values indicate the participant’s most likely response to the item was an unexpected pattern response. The statistical point of view suggests the group needs to be removed as it will affect the overall score aggregate in RUMM2030®.

3.11.4.2 Extreme items

The analysis of item fit was reported. Both studies indicate none of the preliminary analysis holds higher item fit residuals. This indicates the items work well within the designated scales.

3.11.5 Problematic calibrating process

3.11.5.1 Suffering from the calibrations process

The calibrations process shows that the seven-point SDS and five-point LS hold large numbers of under-discriminated category scales indicating overlapping thresholds between the adjacent categories. Although a few rescoring items are successfully made in the first attempt, most of the items do not degrade even though multiple attempts with calibrations were made to collapse the category thresholds with differences in the rescore style. A large number of disordered thresholds was recorded as problematic, where seventy-seven items were still suffering from misfits in SDS calibrations and two items in LS1 calibrations. This indicates that the category structure does not work in the way that it was designed to, especially when the structure is dealing with a certain level of difficult items.

3.11.5.2 Item calibrations – Rescoring

The objective of rescoring the category scale is to reduce the biases and errors which redefine the scale in order to allow the response pattern to work very well, according to the item characteristics as expected by the model.

The rescoring can be calibrated using RUMM2030® software and is legal by a Rasch convention. Horton (2017) states that the rescoring is needed when the category response does not work as expected when aiming over the entire item set. Rasch analysis is a statistical technique that enables items or scales to be modified to fit the data as expected by the model (Hendriks et al., 2012). This technique will allow the original data to be rescored and removed for a better statistical structure and outcome.

RUMM2030® suggests that this item should be removed as it will affect the overall aggregate as this item was unfit regardless of any rescoring formatting that occurred (Pallant and Tennant, 2007; Camargo, 2013). Rescoring format refers to merging the categorical scales which are under-discriminated.

A category structure does not always work in the way that it was designed, especially when the structure deals with a certain level of difficulty items. Some items work very well within the structure, and some not. Therefore, a generic rescore is needed to fix the category scales. Several attempts were made to fix the items with a disordered threshold. The rescoring focus on the category is overlapped or under-discriminated by other category scales. The reason would probably be the participants feel confused or the item is probably too complicated for them to respond to it. The notion of this calibrations is to help establish the internal consistency and reliability of a set of the items so that it will establish the invariance of the items and equal the interval scale scores.

3.12 SUMMARY

The outcome of these studies demonstrates the evidence that measuring affective responses to vehicle interior textures using SDS and LS has introduced biases and error.

The studies have demonstrated a lack of invariance to the data using rating scales. The rating scales give non-linear measurement and introduce some instrument bias because participants have difficulties understanding and using a rating scale. The impact of invariance has affected the estimate of the reliability of a scale.

As a comparison, the LS scales demonstrate a better scaling structure compared with the SDS and mostly fit in targeting, disordered thresholds and calibrations successes in achieving invariance on an interval scale. A study using LS scales provides greater discrimination compared with SDS, which indicates participants find it easy to discriminate the product along the affective dimensions of interest and response style, using the LS scale in a much shorter time than SDS.

Five statistical pieces of evidence have been reported on how bias and error can be detected in RM. There is no clear evidence bias and error affected by sampling bias. However, in this study, bias in the category scales has resulted in difficulties for the participants endorsing the items using SDS and LS. The biasing effect on the scales derived from the unfair decision for every option choice offered in SDS and LS, which affects distortion of judgement, known as targeting. This evidence corresponded in line with the academic theories that the greater number of alternative choices in category scales introduced the risk of biases (Ekman and Lennart, 1964; Cliff, 1973; Field, 2013; Joubert et al., 2015; Brown and Maydeu-Olivares, 2017; Sung and Wu, 2018).

The difficulties, on the other hand, have resulted in participants hardly able to think or respond spontaneously, and the most likely effect of this confusion leads to multiple interpretations and vagueness (Floyd J. Fowler, 1995). This has introduced risk to participants in endorsing the items and may decrease their speed and accuracy (Henson and Camargo, 2014).

The average data collection speed in SDS and LS1 is approximately twenty-five to thirty-five minutes with average speed per items around 12.42 seconds and 12.64 seconds.

Bond and Fox (2015) outlined the most plausible reason that impact from inaccuracies and biases probably inflates the problem of guessing in scaling, which complements Stevens' theory of biases that the participants made their subjective units to judge and discriminate (Poulton, 1989).

The outcome of this study is composed of valuable information on how biases and error can be remedied using PC technique. The PC is likely to offer a possible alternative to SDS1 and LS1. PC involves a dichotomous scale that works in reducing the difficulty of the items and offers greater discrimination between two possible alternatives.

In Chapter 4, the concept of PC is demonstrated.

Chapter 4

Measuring Affective Response of Confectionaries Using PC Method and Rasch Model

Previous chapter findings demonstrate statistical evidence that the use of category scales on self-report questionnaires in affective engineering can be subject to biases and errors. In this chapter, the Pair comparison study 1 (PC1) model was designed to determine whether it can provide a better alternative to SDS and LS in minimising the effect and size of biases and errors. The use of Rasch analysis of PC on products to derive a linear measurement of affective responses is tested. Four pieces of confectionery and twelve evaluative statements measuring the dimensions of specialness, which was validated in previous research, were used. A computer-based self-report system presented one hundred and fifty-seven participants with pictures of pairs of confectionery and the evaluative statements in all pairwise combinations; the participants were asked to indicate which confectionery matched the statement best. The analysis demonstrates the viability of using Rasch analysis to obtain measures of affective response from paired comparisons, that participants find it easier to make paired comparisons compared with evaluating products separately against Likert statements. The study also demonstrated a strong correlation value at R^2 0.9487 indicate the PC1 data from affective responses does not vary within the same context in previous LS study (LS2011) (Camargo and Henson, 2011) but that in this case, the fit of the data to the Rasch model is very poor.

4.1 INTRODUCTION

Normative decision theories assume that participants have stable and consistence preferences for any type of response questionnaire presented to them; however, an increasing amount of literature evidence suggests otherwise (Hsee, 1996). Each response technique exhibits greater variance of negative and positive outcomes in statistical analysis.

The previous chapter has demonstrated this. Bias and error are contributing to sources of error in a statistical outcome where both SDS1 and LS1 studies have demonstrated an inconsistent affective response.

While the categorical response scales are intuitive, participants have difficulties understanding and using a rating scale to make a fair decision for category scales offered in SDS and LS. This results in a distortion of judgement known as targeting, lack of invariance and the reliability of a scale. The structure of scales can be a source of disturbance in data when the items contain misinterpretation, ambiguity and unfamiliarity (F. Camargo and Henson, 2015).

The greater the number of category scales, the greater the difficulty in making judgements, and the more difficult the items, the more troublesome it is for the participants to endorse the item correctly (Wright and Stone, 1999). The complexity in measuring attitudes is often associated with inconsistent judgement that collapses the interval measure scales, especially when using SDS and LS. Participants are often unable to clearly discriminate the category scales during product evaluation (Camargo and Henson, 2011).

Another reason why category responses might be troublesome is that with discrete stimuli, participants are asked to evaluate products separately without reference to a benchmark product. Discrete observation exposes poor true linguistic contrasts for estimation because the contrast is not the same distance: for example, Hard – Soft - how hard is hard and how soft is soft? Some of the products' attributes are hardly evaluated independently, especially involving continuous variables: for example, "*clarity*" requires dependent evaluation (Hsee, 1996; Ahmad et al., 2018).

Christopher (1996) demonstrates that evaluation attributes not only require post-diction information to the attributes which were evaluated discretely but it also requires predictions (Hsee, 1996). Thus, discrete observation loses one dimension of attributes compared with direct comparison, which can be examined as both prediction and post-diction simultaneously.

To minimise the impact size of bias and error, the concept of PC is established in this chapter. The body of work influences the use of PC in Thurstone's discriminable process (Thurstone, 1987). This theory was associated with a discrete choice theory which is based on making a PC of products (Train, 2009) to determine the relative importance of properties in the choices, which are often assumed to vary linearly, rather than to derive measurement.

One of the factors contributing to bias and error is the difficulty of the task of endorsing the product. Therefore, PC is likely to offer a better targeting technique to aid participants endorsing the items more easily using PC. PC works with dichotomous scales to reduce the difficulty of the items and offer greater discrimination between two possible alternatives.

In this chapter, the PC1 was tested and validated using the Rasch model (RM). In these studies, participants were asked to endorse the specialness of confectioneries using the PC approach.

4.2 HYPOTHESIS OF THE STUDY APPROACH

This study aims to determine whether the use of PC improves the target judgement where the binary scales allow the items to be endorsed with greater discrimination.

PC1 formulates that the observed data from affective responses using PC does not vary within the same context when using SDS and LS, but it also fits as per the expectation in the RM-structure.

The study formulates that PC1 likely has a better scaling structure for minimising bias and error when measuring participants' affective responses compared with SDS and LS.

4.3 OBJECTIVES

When making a paired comparisons, the participant merely has to indicate which of two products they endorse more readily, rather than thinking about which category of response one would elicit separately. The challenge is then to derive a linear scale of affective responses from such comparisons.

The aims of the research are therefore to establish

1. Whether participants might find it easier and faster to evaluate products if the evaluations were made as paired comparisons
2. Whether the studies using PC1 satisfy the assumption in minimising the effect size of bias and error
3. To determine how much bias is evident in PC1 and required for the calibrations process to fit the statistics in RM theory
4. To determine whether observed data from affective responses using PC1 does not varies within the same context when using SDS and LS scales and offers resemblance logit as LS. If it does, then the linear correlation can be established.

4.4 METHODS

The approach taken is to use statements developed in LS2011 previous research intended to measure the specialness of confectionary (Camargo and Henson, 2011). The previous research established that the statements could be used as a unidimensional instrument for measuring affective responses. In the research reported here, the statements are used again to evaluate the same confectionary, but instead of SDS and LS assessing each confectionary separately against Likert statements, the user is presented with all the confectionary in all pair combinations, and the participant indicates which of each pair satisfies each evaluative statement best. The responses were then analysed to determine their fit to the RM. The analyses of the paired comparisons data are compared with those of the original research, which used Likert statements.

Thus, the PC1 approach to the confectioneries was used as a pilot stage to examine and validate the measurement structure before applying it to a different application, such as the vehicle interior texture.

4.5 ETHICAL APPROVAL, CONSENT FORM AND RISK ASSESSMENT

Ethical approval was granted from The University of Leeds with approval reference No. MEEC 15-027, conforming to participants' health and safety procedures. The ethical approval, consent forms and risk assessment are included in Appendix B.

The protocol was briefly explained to the participants. Information about instructions and demonstrations of assessing the confectionaries using computer-based PC1 was also briefly explained. The protocol and questionnaire are included in Appendix C.

4.5.1 Data collections

In previous LS2011 research, three hundred and six participants were asked to rate four pieces of confectionery against twenty-four Likert statements on a five-point scale to measure the specialness of confectioneries (Camargo and Henson, 2011). Four items of confectionery that are readily available in the market were used for these experiments: namely Ferrero Rocher®, Lindor®, and Caramel® and Milky Way® from the Mars Celebrations® assortment (Figure 4.6.1)

The confectionery products were chosen because they are likely to elicit different responses to statements about their specialness. The statements used in the experiment were determined through UK-based consumer research by a large confectionery company. The responses to the four pieces of confectionery were analysed using the multi-facet RM to establish a unified scale, for which twelve statements fitted the model for all four pieces of confectionery as shown in Table 4.6.2. The experiment established a linear scale for the measurement of the specialness of confectionery.

In PC1 study, participants were asked to endorse the confectionery against the statements for specialness using PC. One hundred and fifty-seven participants (eighty-three males, seventy-four females) were recruited to take part in this study with an age range of seventeen to fifty-seven years old (SD 7.538). Participants received £5 as compensation for taking part in the study.

4.6 APPARATUS AND EXPERIMENT SETTING

The confectioneries were selected based on their popularity and specialness positioning. None of the participants were required to eat or taste the confectioneries during the study, on health and safety grounds. The confectioneries may have contained nut, gluten, lactose, soy, milk, alcohol and other potential allergens which result in difficulty digesting certain foods and having an unpleasant reaction to them or introducing a possible health risk.

Figure 4.6.1 Four items of confectionery

Table 4.6.1 Pool of Pairwise stimuli

Code	**Pairwise combinations**
Pair 1	Milky Way® vs Caramel®
Pair 2	Milky Way® vs Lindor®
Pair 3	Milky Way® vs Ferrero Rocher®
Pair 4	Caramel® vs Lindor®
Pair 5	Caramel® vs Ferrero Rocher®
Pair 6	Lindor® vs Ferrero Rocher®

4.6.1 Data administration

4.6.1.1 Building the item pool

Twelve statements for measuring the dimension of the specialness of confectioneries, which had been validated in previous research, were used in this experiment (Camargo and Henson, 2011), all of them from a unidimensional construct.

Table 4.6.2 Pool of items in PC1 study

Code	Statement
01	A box of these chocolates would be an appropriate 'thank you' gift.
02	A box of these chocolates would make a thoughtful gift.
03	This is premium chocolate.
04	This chocolate does not need to shout about how good it is.
05	This chocolate would show that someone took the time to choose just the right chocolate for the occasion.
06	I would keep chocolates like this one for myself.
07	The chocolate in this wrapper is likely to exceed people's expectations.
08	This chocolate is like a little present for me.
09	With this chocolate, you feel like you are getting more than just chocolate.
10	This chocolate is stylish.
11	This chocolate would be nice at the end of a dinner party.
12	This chocolate would be good to enjoy with my loved-one on a quiet night.

4.6.1.2 Computer-based survey

Data from participants' affective responses were collected using a bespoke, computer-based, self-report system written by Justine Gallagher (Gallagher, 2016) using Microsoft Visual Basic 2016 (Figure 4.6.2). The system presented each participant with each pair of confectionery in all combinations against each of the twelve statements concerning specialness, and the participant was asked to indicate which of each pair best matched the statement.

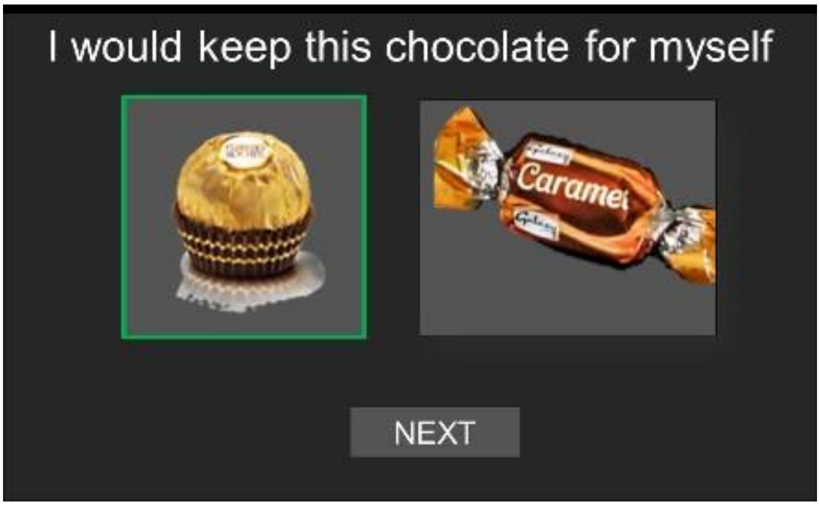


Figure 4.6.2 Example of the self-report interface using a computer

The pair combinations, the order of statements and the order of each pair on the screen were randomised. Thus, for each participant, there were seventy-two statements in total. Participants were encouraged to evaluate the physical confectionery which was placed close to the computer terminal when responding to the item.

4.6.1.3 Data coding and analysis process

The PC1 data were administrated and stored in an online PC program database (Figure 4.6.3) created by Justine Gallagher (Gallagher, 2016). All the data collection were exported to a Microsoft Excel spreadsheet from Microsoft Corporation for cleaning, then imported to the Rasch Model software called RUMM2030® Professional edition by RuMM Laboratory Pty Ltd (Andrich, Sheridan and Luo, 2012) for analysing and calibrating the PC1 study, which was analysed using the Rasch-facet design approach.

The data were analysed using secondary software called PairWise© Software, version 1.5.4198 Copyright Active One Software 2007-2009 (Humphry, 2010; Humphry et al., 2017) to obtained the mean location logit for each stimulus. Original data collection was administrated using an Excel spreadsheet (Microsoft Corporation), and the unidimensionality test was performed and validated using statistical test methods of T-Test and binomial test that were performed using RUMM2030® software and an Excel spreadsheet (Microsoft Corporation).

Sentence	Choice	Order	Date
This chocolate would show that someone took the time to choose just the right chocolate for the occasion.	B	A-B	2016-03-15 14:10:12
This chocolate is like a little present for me.	B	A-B	2016-03-15 14:10:10
This chocolate is stylish.	B	A-B	2016-03-15 14:10:08
With this chocolate, you feel like you are getting more than just chocolate.	B	B-A	2016-03-15 14:10:04
A box of these chocolates would make a thoughtful gift.	B	A-B	2016-03-15 14:09:59
This chocolate would be good to enjoy with my loved-one on a quiet night in.	B	B-A	2016-03-15 14:09:56

Figure 4.6.3 Real-time database

4.7 RESULTS

In PC1 study, two results presented. The first results consist summary statistics of all pairwise confectioneries were treated as separate facet and second results present summary statistics of individual confectioneries treated as separate facets.

4.7.1 Summary of statistics – All paired confectioneries

In PC1 study, RM was used to fit statistics and graphical inspection to assess how well data fitted into RM for the discrete groups known as class intervals along the scale.

In PC1 study, the Chi-Square probability value of $p \leq 0.01$ indicates a lack of the desired scale invariance, indicating inconsistent intervals between class intervals affected by the poor fit of the items to the model (Table 4.7.10 column *p* – page 120). The variance is associated with anomalies in the data which can interpret that the endorsing task was difficult and does not work towards a better fit to the model as expected.

Chi-Square calculated item fit statistics are based on person-item deviations and the deviation by the ability level within the same group known as a class interval. The equal class interval excludes extreme persons of (51 + 52 + 51 = 154), indicating good interval distribution.

The power of analysis of the fit indicated that there was a reasonable degree of fit with the items in the dataset. Initial tests of goodness of fit summarises the overall model fit including the individual person fit and item fit, targeting, the exploration of thresholds and calibrations statistics.

4.7.2 Reliability and internal consistency

The power of test-of-fit is a visual representation of the PSI, which indicates how well the PC1 can distinguish or discriminate between the participants' latent trait locations.

The initial fit statistics (Table 4.7.10 – page 120) exhibit the preliminary analysis of PSI, which was reported as 0.83, and illustrated the PC1 construct meeting the acceptance level to statistically differentiate and equally distribute between three groups of participants across the number of a sample size of a one-and hundred-fifty participants.

The Cronbach's alpha statistic is often used to assess the internal reliability of a scale. The value was reported in initial fit statistics (Table 4.7.10 – page 120), exhibiting the Cronbach's alpha at 0.86 and indicating a good level of reliability.

4.7.3 Item fit location and fit residuals

By default, the RUMM2030® software will automatically assign all the items location mean centralised at zero (Table 4.7.10 – page 120). The item logarithm or log on the same continuum along the x-axis uses a standard unit of termed logits.

The overall fit residual statistics of the model were examined using the mean item log residual test. Item fit residuals refer to how easily the items can be endorsed by participants. RM analysis estimates the degree of divergence known as the residual between the data observed by participants and the expected data from the model. The data will consider that it fits into the model when the values have mean of zero and SD of one. However, Rasch conventionally accepted a residual log range of < +/- 2.50 logits.

The overall fit residual statistics were examined using the mean item log residual test (Table 4.7.10 – page 120). The mean value was recorded at -0.238 logits and the SD was 1.237 logits. These indicated that the items were a reasonable fit and easy to endorse; however, the values SD assigns some misfit to the model.

The negative residuals are usually associated with high item-total correlation in classical test theory. This would usually be interpreted to indicate the redundancy or over-discrimination of an item. RM suggests removing these items because it will significantly influence the overall statistical results.

Items with a higher degree of difficulty are more likely to have fewer chances to be endorsed correctly. While the item with a lower degree of difficulty is more likely to have a higher chance of being endorsed correctly.

Table 4.7.1 indicates three items that do not fit the model well, including high fit residuals, and suggests they should be removed from the dataset. Item three has a log residual of 3.216 logits (*This is premium chocolate),* while item twelve has a log residual of 3.150 logits *(This chocolate would be good to enjoy with my loved-one on a quiet night)* and item eleven has a log residual of 3.084 logits *(This chocolate would be nice at the end of a dinner party*). All these items are associated on Pairwise six *(Lindor® vs Ferrero Rocher®).*

Table 4.7.1: Items with high fit residuals

Item	PairWise	Location	SE	**Fit Resid**	DF	Chi-Sq	DF	Prob
3	6	0.947	0.187	**3.216**	150.87	17.553	2	<0.01
12	6	1.297	0.180	**3.150**	149.89	8.287	2	0.015
11	6	0.845	0.190	**3.084**	150.87	12.595	2	<0.01

4.7.4 Person fit location and fit residuals

Person fit location refers to participants' ability to endorse the items. In RM the participants with higher ability level are likely to have more positive endorsements towards the item difficulty on a scale. The participant's logarithm or log is on the same continuum along the x-axis using a standard unit of termed logits.

The overall person location fit statistics exhibit the person location mean value at 1.964 and the SD value at 1.031 (Table 4.7.10 – page 120). This indicates the participants have a higher ability level in responding to difficult items. Despite having sufficient fit location, person fit that contains some misfit indicates a lack of the expected probabilistic relationship among the items within a scale.

The overall fit residual statistics were examined using a mean person log residual test (Table 4.7.10 – page 120). The mean value was recorded as -0.260 and the SD was 1.005. This value considers a reasonable fit to indicate the participants are behaving in the same way as the others when responding to the test. Participants holding extreme fit residuals can misfit the model, in which the individual response patterns may exhibit as unfit if they are unexpected or contain too much dependence.

Further investigation was carried out to determine the source of misfit. In the individual person analysis, 1.91 percent or three participants out of one hundred and fifty-seven participants were affected by higher positive fit-residuals. Person IDs 98, 63 and 100 had log residuals greater than >+2.50 logits where the residual value holds at of 2.601 to 3.837 logits as shown in Table 4.7.2. However, there are no participants identified before adjustment with higher negative fit residuals than > -2.50 logits.

Table 4.7.2 Participants holding high fit-residuals

No	PersonID	Location	SE	Total/Exp Sc	**Fit Resid**	Sex	Age
1	98	-0.701	0.275	25 / 72	**3.837**	1	3
2	63	-1.655	0.329	14 / 71	**3.562**	0	3
3	100	-0.072	2.601	34 /72	**2.601**	1	0

Another source of misfit, the outlier was identified in this study. Table 4.7.3 exhibits 4.45 percent or seven participants out of one hundred and fifty-seven participants were identified as an outlier or extreme person with higher location logits. This group are often associated with poor targeting sampling because the survey was carried either in an inappropriate way or handled negligently or excessively consistency tin making a judgement.

This extreme group was skewed too far to the right in the continuum which is quite a distance from the rest of participants (Figure 4.7.2, page 109). In common practice, the Rasch analysis suggested these groups need to be removed as they will interrupt and affect the overall aggregate in the statistical analysis.

Table 4.7.3: Participants identified as outliers hold higher location logits

No	PersonID	**Location**	SE	**Total/Exp Sc**	Fit Resid	Sex	Age
1	91	**5.041**	1.223	**71.807 / 71**	**-**	0	2
2	95	**5.041**	1.223	**71.807 / 72**	**-**	0	1
3	12	**5.041**	1.223	**71.807 / 72**	**-**	1	1
4	112	**4.251**	0.838	**71 / 72**	-0.581	0	1
5	80	**4.251**	0.838	**71 / 72**	-0.042	0	3
6	70	**4.251**	0.838	**71 / 72**	-0.599	1	1
7	99	**4.251**	0.838	**71 / 72**	-0.477	1	1

Gender demographic profile of person factors was coded as zero for male and one for female, while age demographic profile was coded as following Table 4.7.4.

Table 4.7.4 Age demographic profiles of person factor

Age	Code
16 to 20 years old	0
21 to 30 years old	1
31 to 40 years old	2
41 to 50 years old	3
Above 51	4

4.7.5 Fit statistics for facet analysis – All paired confectioneries

The PC1 study involved two facet analyses using RM. The first facet indicated twelve items of questionnaires that corresponded to six Pairwise stimuli, as shown on the second facet. Table 4.7.5 shows preliminary analysis of two facet designs before alteration by location order. Facet one showed item location discriminated by the level of difficulty. Facet two showed stimuli location discriminated by the ability of participants corresponding to the stimuli. The facet level illustrates items and Pairwise in hierarchical difficulties to endorse.

Table 4.7.5: Preliminary facet analysis of all paired confectioneries

Level	Item	Locn	SE	FitRes
1	10	-0.888	0.45	-0.481
2	2	-0.814	0.44	-0.576
3	3	-0.560	0.34	0.375
4	1	-0.463	0.35	-0.093
5	5	-0.236	0.38	-0.129
6	9	-0.052	0.27	-0.444
7	11	0.186	0.24	0.153
8	12	0.386	0.22	-0.394
9	7	0.473	0.22	-0.092
10	8	0.500	0.21	-0.569
11	4	0.727	0.20	-0.793
12	6	0.741	0.20	-0.766

Level	Pairwise	Locn	SE	Fit Res
3	Pair 3	-1.006	0.42	-0.527
5	Pair 5	-1.003	0.41	-0.224
2	Pair 2	-0.157	0.26	-1.298
4	Pair 4	0.361	0.22	-0.392
1	Pair 1	0.839	0.20	-0.326
6	Pair 6	0.965	0.19	1.469

Table 4.7.6 Preliminary facet analysis (Individual confectionery)

Seq	Item	Locn	SE	FitRes
1	40	-2.105	0.38	-0.182
2	8	-2.096	0.36	-0.228
3	4	-1.836	0.31	-0.276
4	12	-1.771	0.31	-0.154
5	36	-1.495	0.25	-0.348
6	20	-1.477	0.26	-0.306
7	44	-1.352	0.22	-0.212
8	48	-1.084	0.20	-0.355
9	32	-1.016	0.19	-0.322
10	28	0.992	0.19	-0.390
11	16	0.908	0.18	-0.328
12	24	0.896	0.18	-0.475
13	11	-0.810	0.22	-0.342
14	7	-0.791	0.21	-0.436
15	19	-0.746	0.20	-0.297
16	03	-0.725	0.20	-0.416
17	39	-0.711	0.21	-0.485
18	47	-0.537	0.18	-0.440
19	43	-0.519	0.19	-0.323
20	27	-0.506	0.18	-0.396
21	31	-0.475	0.18	-0.408
22	35	-0.391	0.18	-0.434
23	23	-0.356	0.17	-0.357
24	15	-0.172	0.17	-0.335
25	14	0.453	0.18	-0.369
26	46	0.475	0.19	-0.470
27	22	0.483	0.17	0.428
28	26	0.571	0.19	0.368
29	30	0.573	0.18	-0.413
30	42	0.645	0.19	-0.392
31	34	0.703	0.22	-0.395
32	21	0.730	0.18	-0.699
33	13	0.736	0.17	-0.612
34	18	0.814	0.23	-0.500
35	29	0.838	0.18	-0.606
36	10	0.862	0.26	-0.398
37	25	0.903	0.19	-0.495
38	38	1.025	0.29	-0.330
39	2	1.044	0.26	-0.404
40	45	1.101	0.20	-0.580
41	6	1.140	0.32	-0.335
42	41	1.219	0.22	-0.427
43	33	1.233	0.21	-0.525
44	17	1.392	0.24	-0.447
45	1	1.602	0.26	-0.441
46	5	1.689	0.28	-0.407
47	9	1.750	0.29	-0.326
48	37	1.786	0.33	-0.352

Level	Pairwise	Locn	SE	Fit Res
1	MilkyWay®	-1.231	0.23	-0.390
2	Caramel®	-0.714	0.23	-0.344
3	Lindor®	0.552	0.19	-0.454
4	Ferrero®	1.393	0.26	-0.320

4.7.6 Fit statistics for facet analysis – Individual confectionery

Table 4.7.6 illustrates facet analysis of individual confectionery were treated as separate facets by location order. The first facet indicated forty-eight items of questionnaires that correspond to four stimuli, as shown on the second facet.

The facet locations appear to be ordered according to the ease with which the confectionery in each pair can be discriminated. In other words, those confectionery pairs that are most similar and are difficult to discriminate have higher positive values, whereas those that are very different have more negative values. For example, the confectionery pair three of Milky Way® and Ferrero Rocher® has a large negative location relative to the other pairs, whereas pair six of Ferrero Rocher® and Lindor®, which are more similar in term of specialness, has the most positive value of the pairs.

This result cause speculation that it might be possible to interpret the location of the pairs as an indication of how easy it is for the participants to discriminate between those pairs, in which case the overall person-item distribution would be an indication of the ease of overall discrimination through paired comparisons. If this is the case, then the person-item distributions demonstrate that participants find it much easier to carry out paired comparisons as opposed to Likert statements.

4.7.7 Expected probability targeting

Targeting refers to how well participants endorse the difficulties of the items; in RM the targeting uses person-item distribution to plot the logarithm or log on the same continuum along the x-axis using a standard unit of termed logits.

The person-item distribution graph (Figure 4.7.2) indicate the analysis using all pairwise confectioneries treated as a separate facets, found participants very easy to endorsed the items and that the matching of the difficulty of the items which plot the normal distribution kurtosis.

The summary of fit statistics for preliminary analysis are presented in graphical representation to indicate how well the person and items are distributed is divided into two parts, the lower part, is a graphical representation of item difficulties where the positive logits indicate the difficulty of the item within the difficulty region, while negative logits indicate the item within easiness region.

The upper part is a graphical representation of participants' abilities, where the positive logits indicate the higher ability of participants endorsing the item while negative logits indicate the lower ability of participants endorsing the item. The targeting would be expected to follow the probability relationship of the person-item in a normal distribution.

In comparison with the previous LS2011 study (Figure 4.7.1) illustrates the person-item distribution for the previous research in which participants were asked to rate the confectionary on a five-point category scale against Likert statements (Camargo and Henson, 2011).

The items, in this case, are a combination of each Likert statement with each product. It can be seen that, overall, the participants found it easy to endorse the items but that the targeting of the difficulty of the items to participants could be improved; a large number of the problematic items held in difficult item regions indicate that high-ability participants are having difficulties responding to the most difficult items.

Meanwhile, in the PC1 study, the person-item distribution for the analysis of the current data of individual pieces of confectioneries were treated as separate facets (Figure 4.7.3) shows that the participants found it very easy to endorse the items and that the matching of the difficulty of the items with participants' willingness to endorse was quite poor.

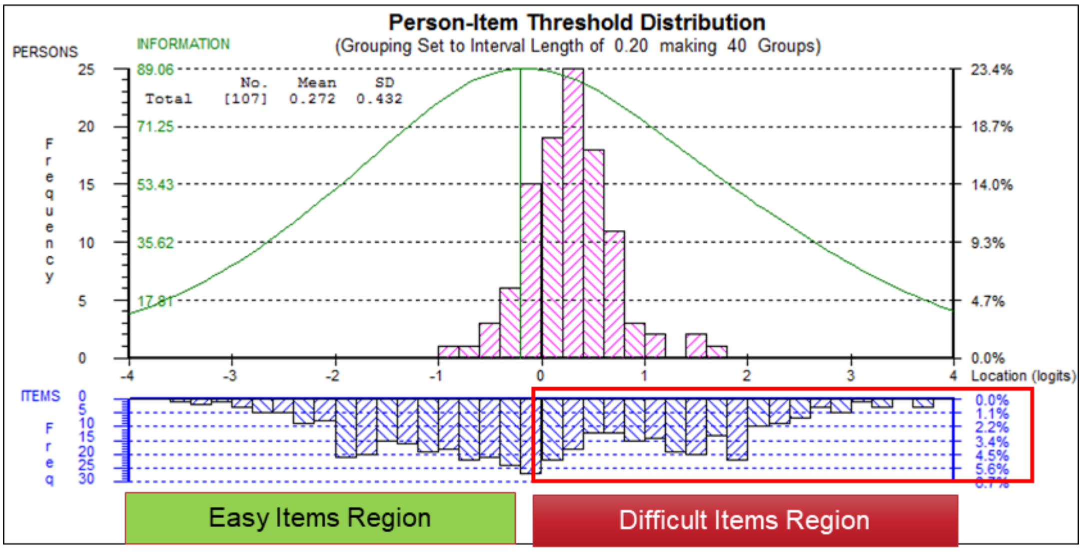


Figure 4.7.1 Person-item distribution in LS2011 study

The person-item distribution for the analysis in PC1 which individual pieces of confectionery were treated as a separate facet (Figure 4.7.3) shows that approximately half of the items were very easy to endorse and the other half were very difficult to endorse.

However, because of the way the data are coded for analysis, each comparison is represented by two data points, and consequently, each easy item has a mirror-image difficult item. In theory, therefore, the item distribution should be symmetrical. The coding of the same-confectionery pairs by random ones and zeros might account for the small amount of asymmetry in the distribution. The very narrow spread of the participants' willingness to endorse shows that there was not much variation in participants' affective assessments of the confectionery (Figure 4.7.3).

Together with the person-item distribution in (Figure 4.7.2) this can be interpreted and seen to demonstrate that participants find the paired comparisons task very easy. The widespread distribution of the items compared with the distribution of the participants indicate that these paired comparisons are perhaps too easy and that participants were too consistent in making the comparisons.

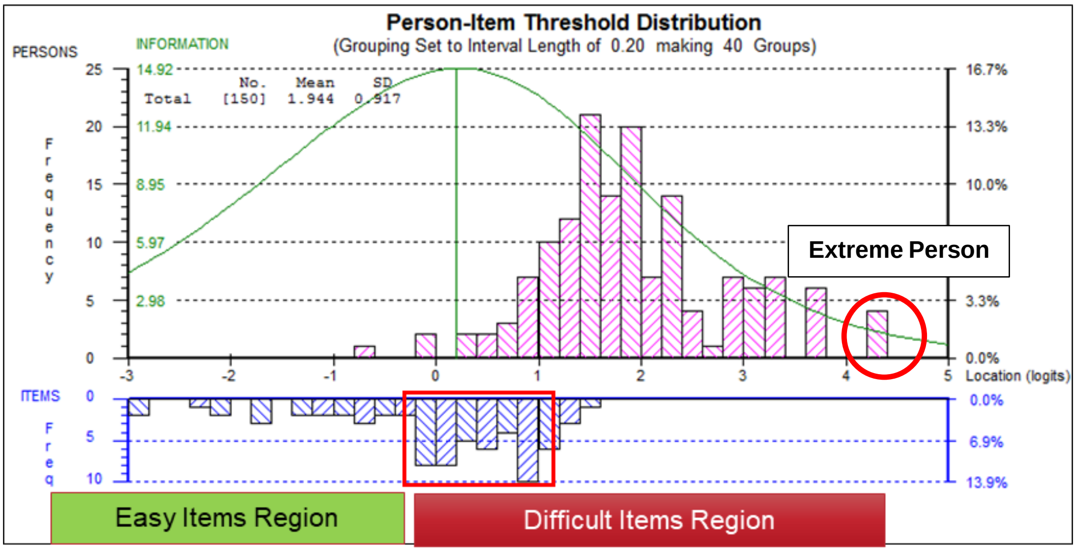


Figure 4.7.2 Person-item distribution in PC1 which all paired of confectioneries were treated as separate facets

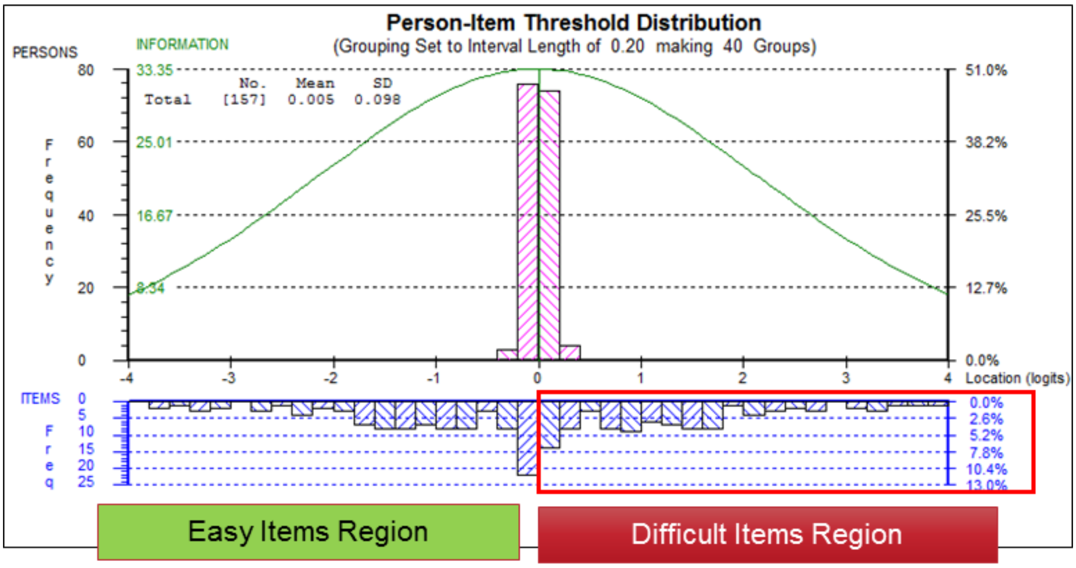


Figure 4.7.3 Person-item distribution in PC1 which individual pieces of confectionery were treated as separate facets

The graphical representation in (Figure 4.7.4), exhibits the comparison of person location logits in endorsing the specialness of confectioneries between the LS2011 and PC1. This illustrates the willingness for endorsing PC1 is much higher than LS2011. This indicates that measuring affective response using PC is much more effortless.

More Inclined to endorse

Likert Scales			Pair comparisons
	3.60	3.60	***
	3.50	3.50	
	3.40	3.40	
	3.30	3.30	
	3.20	3.20	****
	3.10	3.10	
	3.00	3.00	***
	2.90	2.90	
	2.80	2.80	****
	2.70	2.70	
	2.60	2.60	*
*	2.50	2.50	
*	2.40	2.40	**
	2.30	2.30	
***	2.20	2.20	*******
**	2.10	2.10	
**	**2.00**	**2.00**	****
****	1.90	1.90	
**	1.80	1.80	**********
*****	1.70	1.70	
***	1.60	1.60	*******
**	**1.50**	**1.50**	
***	1.40	1.40	*************
**********	1.30	1.30	
*****	1.20	1.20	******
************	1.10	1.10	
****************	**1.00**	**1.00**	*****
************	0.90	0.90	
******************	0.80	0.80	****
***************	0.70	0.70	
****************	0.60	0.60	**
************	**0.50**	**0.50**	
********************	0.40	0.40	*
*****************	0.30	0.30	
*************	0.20	0.20	*
*********	0.10	0.10	
***************	**0.00**	**0.00**	
************	-0.10	-0.10	
********	-0.20	-0.20	*
*******	-0.30	-0.30	
*****	-0.40	-0.40	
***	**-0.50**	**-0.50**	
***	-0.60	-0.60	
**	-0.70	-0.70	
****	-0.80	-0.80	*
*	-0.90	-0.90	
*	**-1.00**	**-1.00**	

Less Inclined to endorse

(* = 2 Persons)

Figure 4.7.4 Comparison of persons location logits between LS2011 and PC1

4.7.8 Disordered Thresholds

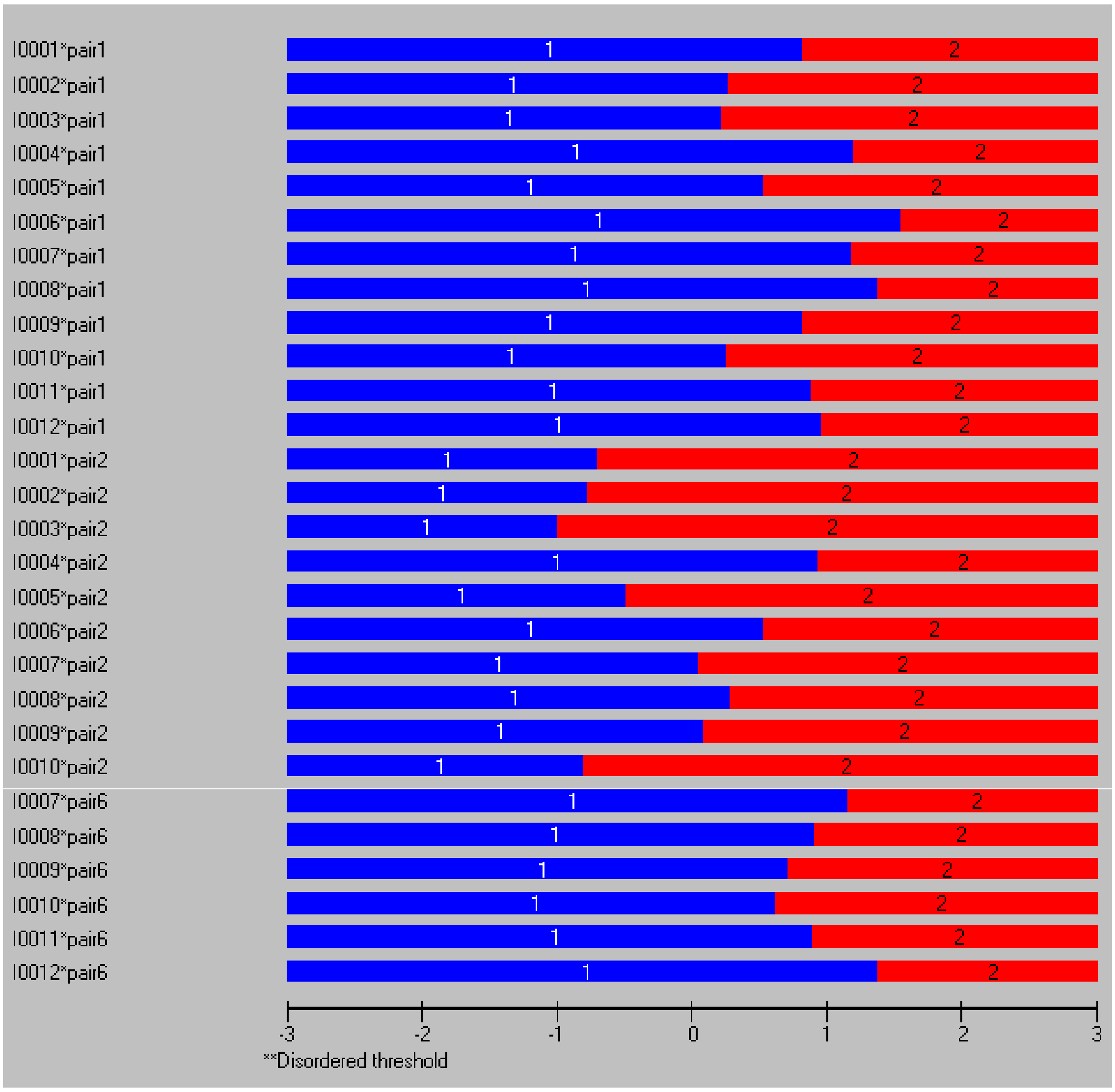


Figure 4.7.5: Thresholds map in PC1

Disordered thresholds occur when participants have difficulties consistently to endorsing or discriminating between response options categories on a scale.

The investigation of disordered thresholds for PC1 study (Figure 4.7.5) illustrates that there are no disordered thresholds identified in the PC1 analysis. No overlapping thresholds or under-discrimination were identified resulting in all the category scales being ordered.

Thus, this analysis was rejected the assumption that lack of invariance is affected by disordered thresholds. The graphical representation shows the participants are able to use PC to discriminate between two possible answers.

4.7.9 Calibrating PC1 processes

Preliminary analysis of PC1 analysis required minor calibrations to the process to fit the statistics in RM theory. Unlike SDS and LS, which deal with problems as the SDS and LS hold a large number of disordered thresholds for both preliminary and after calibrations, PC required minor calibrations to remove unfit participants who hold missing data, high fit residuals and extreme ability participants.

High residual values indicate the participant’s most likely response to the item was an unexpected pattern response. The statistical point of view suggests the group needs to be removed as it will affect the overall score aggregate. Table 4.7.10 – page 120 (row calibrated data one to calibrated data four) has illustrated the calibrations activities of PC1 study. The key features in PC1 do not require rescoring the adjacent sections in the category scales. This is respectable for maintaining the originality of dataset.

4.7.10 Differential item functioning (DIF) analysis

The DIF analysis was performed in this chapter. DIF refers to another possible source of a misfit in the data, which could be affected by demographic factors such as gender and age affecting the statistical results.

The DIF analysis showed there is no gender bias Table 4.7.7 below indicates there is no statistical difference in ability between the male and female subgroups (p=0.112891). This indicates that participants have the same level of probability in responding to Pairwise items without being influenced by demographic factors of gender and age.

Mean location value of 2.129 logits indicates males are displaying better ability than females at 1.871 logits in responding to most of the tasks. Male was coded with the blue colour bar, while female was coded in red (Figure 4.7.6)

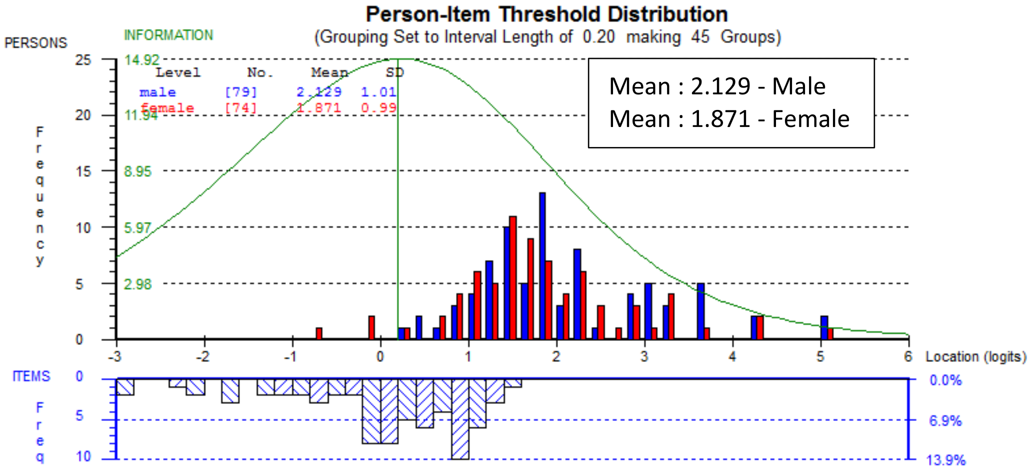


Figure 4.7.6: DIF by Gender

Table 4.7.7 No statistical difference between genders

One-way Analysis of Variance

Source	SumSquare	DegF	MeanSS	F-Stat	Prob
Between	2.54	1	2.54	2.542760	0.112891
Within	150.89	151	1.00		
Total	153.43	152			

4.7.10.1 Examining gender bias of DIF by ICC

The analysis characteristics curve (ICC) of item ten Pairwise three, as shown in Figure 4.7.7 from this plot, shows that over the upper three class intervals, males display greater ability in their responding to the easiest item correctly than females. However, for endorsing difficult items to endorse, such as item six pairwise one, this trend is reversed. Females are displaying a greater ability to respond where two class intervals are plotted above the slope, as shown in Figure 4.7.8.

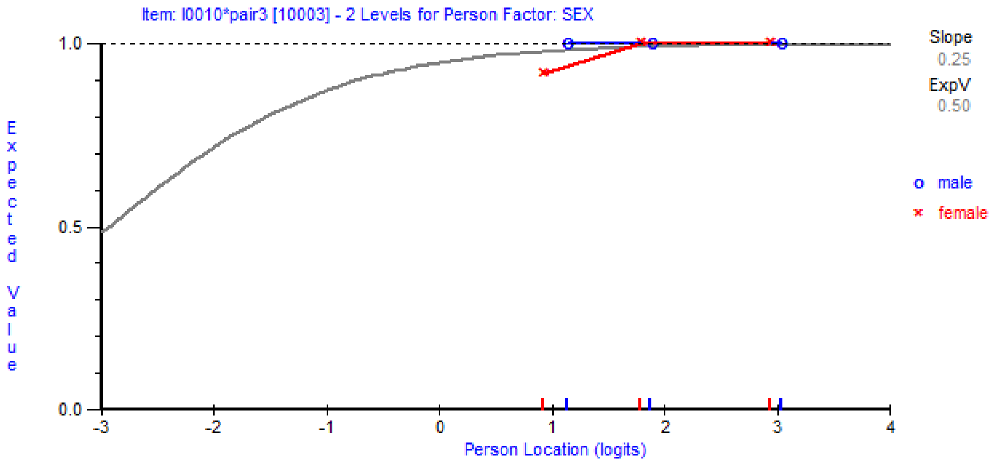


Figure 4.7.7: ICC showing plot by gender for the easiest item to endorse

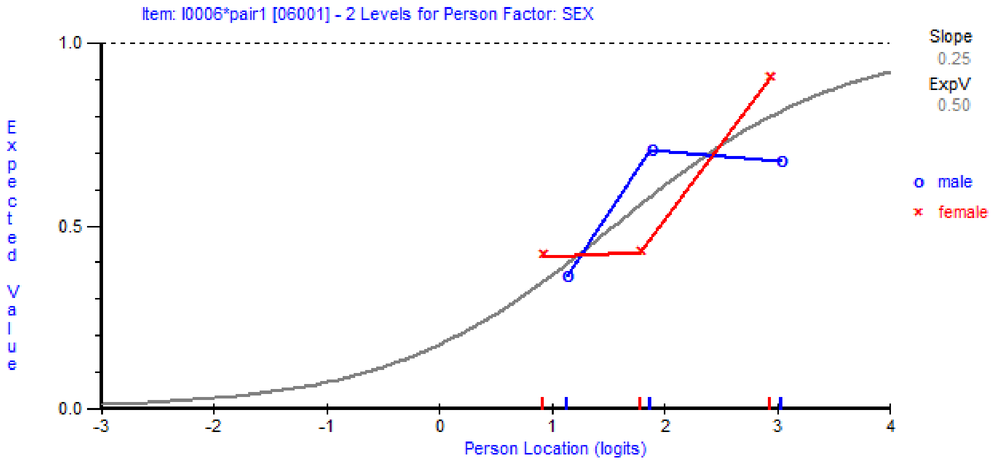


Figure 4.7.8: ICC showing plot by gender for the most difficult item to endorse

4.7.10.2 DIF by age group

Another source of the misfits in the data could be due to the differential item functioning of certain items. However, the DIF analysis showed there is no age bias Table 4.7.8 below indicates there is no statistical difference in ability between the age groups (p=0.229400). The age of the group is divided by five classes from sixteen to sixty years old.

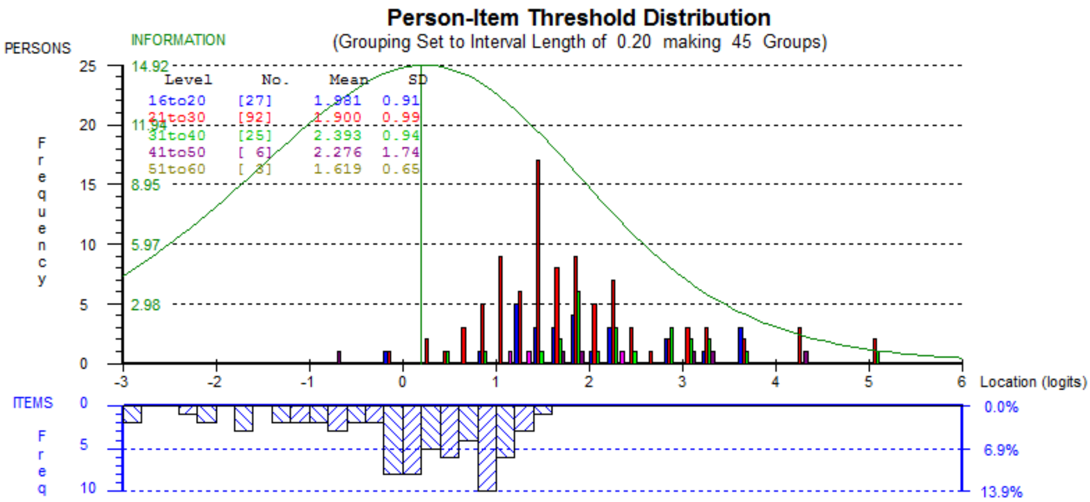


Figure 4.7.9: DIF by age group

Table 4.7.8 No statistical difference by age group

One-way Analysis of Variance

Source	SumSquare	DegF	MeanSS	F-Stat	Prob
Between	5.68	4	1.42	1.422226	0.229400
Within	147.75	148	1.00		
Total	153.43	152			

4.7.10.3 Examining age bias of DIF by ICC

The ICC item ten pairwise three, as shown in Figure 4.7.10 from this plot, shows that over the upper level, all three class intervals display a greater ability to endorse the easiest item correctly. However, for endorsing difficult items, this trend is slightly different. Two age class intervals (age group sixteen to twenty, and thirty-one to forty years old) display a greater ability to respond to the difficult items correctly, where these groups are plotted above the slope, as shown in Figure 4.7.11. The rest of the class intervals (age groups twenty-one to thirty, forty-one to fifty, and fifty-one to sixty years old) display a lesser ability to endorse difficult items.

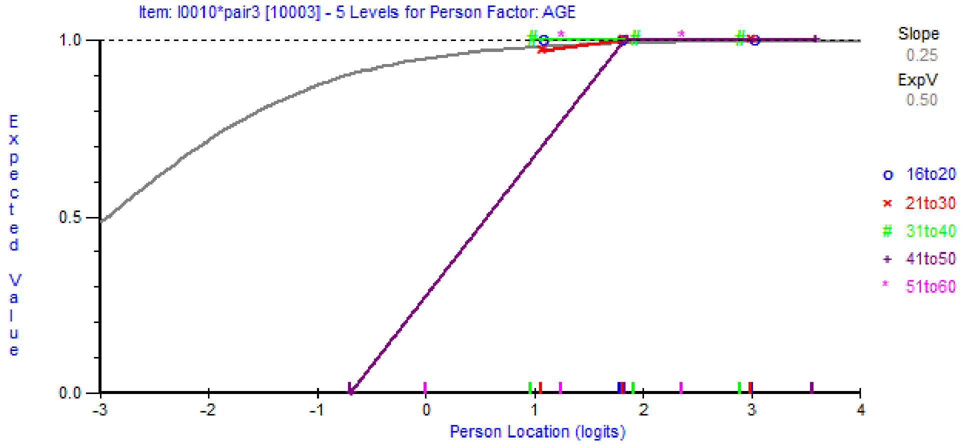


Figure 4.7.10: ICC showing plot by age group for the easiest item to endorse

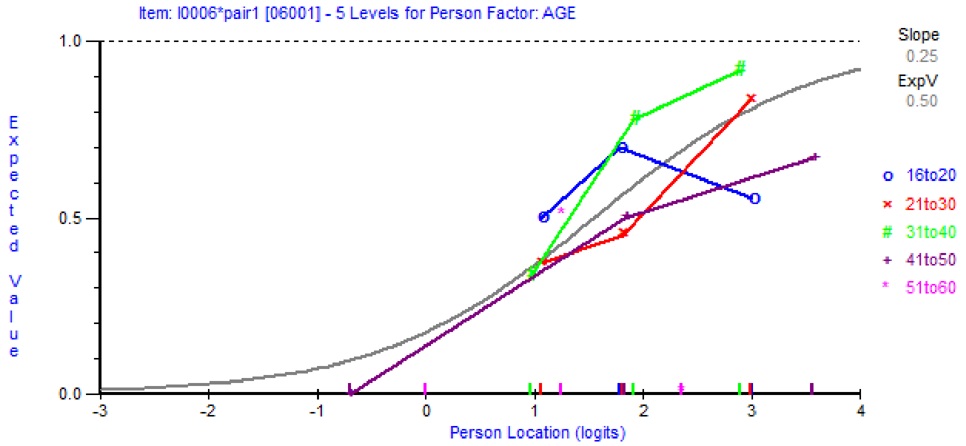


Figure 4.7.11: ICC showing plot by age group for the most difficult item to endorse

4.7.11 Data collection speed

The duration of each study took approximately 4.25 to 6.17 minutes to complete the seventy-two pairs of comparison items for each participant. The individual response rate for each comparison item took an average of 3.7 to 6.1 seconds per comparison.

4.7.12 Calibrated statistics

Theory in RM was rejected if the data did not fit into the model. However, in many cases, further investigation was required to examine any anomalies or discrepancies between the observed data and the model that affected the poor statistical signal.

The RM is a statistical technique that enables unfit items and scales to be modified and therefore known as calibrations. These allow some alterations to the preliminary dataset to improve the statistical model which results in good Chi-Square value.

Underlying this notion, the RM plots normal distribution if the items and stimuli are within the expected fit-residuals range. The accepted Chi-Square value of item-trait interaction indicates a reasonable degree of fit between the data and the model, which reflects the property of invariance across the trait.

Thus in this calibrations process, the preliminary dataset was calibrated as per the following approach:

1) Removing unfit pairwise stimuli
2) Removing unfit stimuli
3) Register dataset in Pairwise matrix representation
4) Register and validate using PairWise© Analysis

Another calibrations approach to determine whether in a way dataset being administrated in difference representation may reflect the properties of statistics.

4.7.13 Calibrations Results

The third calibrations provide significant results of summary statistics compare with the first and second calibrations approach. While maintaining the stimuli, the calibrations process using variable pairwise matrix as shown in Table 4.7.9 achieve sufficient statistics as expected by the RM. However, the reliability result was poor because variable pairwise matrix assigns the dataset in a matrix representation where the comparison of each of the items is mirrored and, with those items with the same comparison, the value will automatically be omitted. In RM the grey space was counted as missing value.

Table 4.7.9 Variable Pairwise Matrix

	Milky Way®	Caramel®	Lindor®	Ferrero®
Milky Way®		0.284	0.143	0.090
Caramel®	0.716		0.192	0.092
Lindor®	0.857	0.808		0.302
Ferrero®	0.910	0.908	0.698	

4.7.13.1 Summary of statistics – Removing one Pairwise

Table 4.7.12 – page 121 shows the initial test of fit summarsing of overall model fits including the individual person fit and item fit; the exploration of thresholds and probability curves after removing pair one (Milky Way® – Caramel®) shows better chi-square value.

A Chi-Square value of item-trait interaction of $p < 0.076$ indicates the better interaction value of the invariance across the traits. The calibrated analysis exhibits a good item-trait interaction of $p < 0.186$ (Table 4.7.12 column *p* – page 121). This value indicates one of the stimuli (Milky Way® – Caramel®) was unfit and contained difficulties for participants to endorse readily.

4.7.13.2 Summary of statistics – Removing one stimulus

Table 4.7.14 – page 122 investigates the initial test of fit summarising overall model fit including the individual person fit and item fit; the exploration of thresholds and probability curves after removing one of stimulus (Milky Way®) significantly improved the Chi-Square value.

A Chi-Square value of item-trait interaction of *p*< 0.022 indicates the better interaction value of the invariance across the traits. The calibrated analysis shows a good item-trait interaction of *p*< 0.937 (Table 4.7.14 column *p*– page 122). This value indicates the Milky Way® stimulus was unfit and presented difficulties for participants to endorse readily.

4.7.13.3 Summary of statistics – Individual Confectioneries

Table 4.7.16 – page 123 investigates the initial test of fit summarising the overall model fit including the individual person fit and item fit, the exploration of thresholds and probability curves when considering the raw data administrated in pairwise matrix in five variable datasets. Under this calibrations process, five Rasch analyses were administrated in difference representation (RuMM 1 to RuMM 5) to observe any variation of the statistical outcomes. However, all of the analyses exhibit a small degree of statistical fit indicating consistency across the difference analysis. The Chi-Square interaction shows inadequate item-trait interaction of *p*< 0.05.

Table 4.7.17 – page 123 column average is *"in bold",* exhibiting the average value of mean location logit of each stimulus of PC1 study. This logit was used to overlay the value from the LS study to observe the linear correlation (Table 4.7.20 – page 125).

4.7.13.4 PairWise© software results

PairWise© analysis (Humphry, 2010; Humphry et al., 2017) was used to measure algorithms from the dataset to obtain the mean location logit for each stimulus. The program was written to conform with standard statistical measurement in item response theory (IRT), which is similar to RM. The result demonstrated Ferrero Rocher®'s rank as the most perceived specialness at 1.403 logits; second, Lindor® at 0.608 logits; third, Caramel® at -0.608 logits; and fourth, Milky Way® at -1.352 logits (Table 4.7.18 – page 124). The reason the PairWise© application was used in this study was to determine mean logits, which are absent in RM. Instead of giving a location estimate based on each stimulus, RM offers a PairWise© format which is difficult to compare and validate with the location estimate in LS study to determine linearity.

Table 4.7.10: Fit Statistics for all pairwise were treated as a separate facet

Analysis	Item Loc		Person Loc		Item Fit Res		Person Fit Res		Chi-Square Interaction				
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Value**	**dF**	***p***	**Psi Index**	**Alpha**
Initial Data (n=157)	0.000	1.010	1.964	1.031	-0.238	1.237	-0.260	1.005	227.5	144.0	**<0.001**	0.83	N/A
Calibrated Data 1 (n=150) Ref : DelMsing	0.000	1.048	1.944	0.917	-0.275	1.065	-0.257	0.984	214.8	144.0	**<0.001**	0.83	0.86
Calibrated Data 2 (n=150) Ref : Del3ExtrmPrsn	0.000	1.048	1.944	0.917	-0.276	1.065	-0.258	0.984	214.8	144.0	**<0.001**	0.83	0.86
Calibrated Data 3 (n=148) Ref : Del2HiFitResPrsn	0.000	1.105	2.003	0.886	-0.275	0.933	-0.259	0.913	196.8	144.0	**<0.002**	0.81	0.83
Calibrated Data* 4 (n=146) Ref : Del2HiFitResPrsn	0.000	1.048	1.880	0.844	-0.234	1.115	-0.252	0.994	212.9	144.0	**<0.001**	0.81	0.85

Table 4.7.11: Prelimenary facet design (All Pairwise)

Confectionery pair	Location (logit)	Standard error
Pair 1 Milky Way® - Caramel®	0.835	0.19
Pair 2 Milky Way® - Lindor®	-0.151	0.25
Pair 3 Milky Way® - Ferrero®	-0.961	0.38
Pair 4 Caramel® - Lindor®	0.346	0.22
Pair 5 Caramel® - Ferrero®	-0.959	0.39
Pair 6 Lindor® - Ferrero®	0.89	0.19

Table 4.7.12: Fit statistics for the initial and calibrated analysis – Removing pair one

Analysis	Item Loc		Person Loc		Item Fit Res		Person Fit Res		Chi-Square Interaction				
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Value	dF	*p*	Psi Index	Alpha
Initial Data (n=157)	0.000	1.006	2.035	1.021	-0.309	1.183	-0.294	1.007	142.8	120.0	**<0.076**	0.81	N/A
Calibrated Data 1 (n=143) Ref : DelMsing	0.000	1.054	2.078	0.997	-0.347	1.035	-0.294	0.987	141.3	120.0	**<0.089**	0.80	0.85
Calibrated Data 2 (n=140) Ref : DelExtrmPrsn	0.000	1.054	2.078	1.007	-0.416	1.001	-0.271	0.986	131.8	120.0	**<0.216**	0.80	0.85
Calibrated Data 3 (n=141) Ref : DelExtrmPrsn	0.000	1.117	2.146	0.967	-0.339	0.911	-0.299	0.913	133.8	120.0	**<0.186**	0.78	0.82

* Removed Pair 1 (Milky Way® – Caramel®)

Table 4.7.13: Preliminary facet design (Five Pairwise)

Confectionery pair	Location (logit)	Standard error
Pair 1 Milky Way® - Lindor®	0.039	0.25
Pair 2 Milky Way® - Ferrero®	-0.781	0.37
Pair 3 Caramel® - Lindor®	0.505	0.22
Pair 4 Caramel® - Ferrero®	-0.808	0.38
Pair 5 Lindor® - Ferrero®	1.044	0.19

Table 4.7.14: Fit statistics for the initial and calibrated analysis – Removing one stimulus

Analysis	Item Loc		Person Loc		Item Fit Res		Person Fit Res		Chi-Square Interaction				
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Value	dF	*p*	Psi Index	Alpha
Initial Data	0.000	1.013	1.658	0.944	-0.281	0.836	-0.210	0.820	97.9	72.0	**<0.0225**	0.74	N/a
Calibrated Data N137	0.000	1.095	1.702	0.935	-0.299	0.763	-0.199	0.752	118.8	144.0	**<0.9377**	0.74	0.81

* Removed Milky-Way

Table 4.7.15: Preliminary facet design (Three Pairwise)

Confectionery pair	Location (logit)	Standard error
Pair 1 Caramel® - Lindor®	0.258	0.22
Pair 2 Caramel® - Ferrero®	-1.142	0.42
Pair 3 Lindor® - Ferrero®	0.884	0.20

Table 4.7.16: Fit Statistics for the preliminary analysis of individual confectioneries were treated as a separate facet

Analysis	Item Loc		Person Loc		Item Fit Res		Person Fit Res		Chi-Square Interaction				
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Value	dF	*p*	Psi Index	Alpha
RuMM 1 Initial Data	0.000	1.564	0.005	0.098	-0.208	0.511	-1.042	2.787	448.8	384.0	**0.0135**	-2.117	-2.028
RuMM 2 Initial Data	0.000	1.565	-0.010	0.101	-0.233	0.459	-1.016	2.792	424.2	384.0	**0.0764**	-1.900	-1.818
RuMM 3 Initial Data	0.000	1.567	0.024	0.094	-0.190	0.478	-1.008	2.781	464.7	384.0	**0.0029**	-2.407	-2.310
RuMM 4 Initial Data	0.000	1.565	-0.010	0.108	-0.261	0.431	-1.018	2.793	454.5	384.0	**0.0075**	-1.515	-1.445
RuMM 5 Initial Data	0.000	1.565	0.010	0.107	-0.270	0.469	-1.017	2.794	438.4	384.0	**0.0284**	-1.602	-1.527
Average Data*	0.000	1.565	0.004	0.102	-0.232	0.470	-1.020	2.789	446.1	384.0	**0.0257**	-1.908	-1.825

Table 4.7.17: Average individual logit RUMM 1 – RUMM 5

Confectionery	Average		RuMM 1		RuMM 2		RuMM 3		RuMM 4		RuMM 5	
	Location (logit)	Standard error	Location (logit)	Standard error	Location (logit)	Standard error	Location (logit)	Standard error	Location (logit)	Standard error	Location (logit)	Standard error
Milky Way®	**-1.239**	0.23	-1.231	0.23	-1.245	0.23	-1.248	0.23	-1.234	0.23	-1.238	0.23
Caramel®	**-0.720**	0.23	-0.714	0.23	-0.710	0.23	-0.720	0.23	-0.742	0.23	-0.714	0.23
Lindor®	**0.552**	0.19	0.552	0.19	0.566	0.19	0.541	0.19	0.551	0.19	0.549	0.19
Ferrero®	**1.409**	0.29	1.392	0.26	1.399	0.26	1.427	0.26	1.425	0.26	1.402	0.26

Table 4.7.18: PairWise© results

Confectionery	Preferred	Involved	Estimated	SE	Outfit	Chi-Sqr	DF	Class Interval
Milky Way®	973	5652	**-1.352**	0.037	1.177	57.123	2826.0	1
Caramel®	1883	5652	**-0.659**	0.033	0.891	37.679	2826.0	1
Lindor®	3707	5651	**0.608**	1.009	1.009	15.925	2825.0	2
Ferrero®	4740	5651	**1.403**	1.097	1.097	38.988	2825.0	3

Table 4.7.19: Summary statistics

Property	Value
Outer Loop Count	10
Mean Location	0.00
Variance	1.150
Mean Square Error	0.001
PSI	0.999
Sum Chi-Square X^2	74.857
DF	5651.5
DF / Element	1.00

4.7.14 Linear correlation between scales

The linear correlation in this study demonstrates strong relationship between PC1 and previous LS2011 study (Camargo and Henson, 2011).

The correlation R-square value exhibit R^2 0.9487 indicate strong relationship between individual confectionery were treated as separate facets in PC1 study and individual texture in previous LS study (Camargo and Henson, 2011) (Figure 4.7.12).

The result indicate the PC1 data from affective responses using confectioneries does not vary within the same context using LS2011 study (Camargo and Henson, 2011) and offers resemblance logits. This is indicate PC1 was sufficient to hold the stability across a difference samples.

Similar results were obtained when comparing the mean logits of PC1 using the PairWise© software, which exhibit almost identical logits of 0.9964 of R-square values indicating a strong correlation between them (Figure 4.7.13). The table below demonstrates the comparison logits between methods (Table 4.7.20).

Table 4.7.20: Comparison mean location of confectionery between methods

Confectionery	LS - Study 2011 Location (logit)	Standard error	PC1-RuMM2030® Location (logit)	Standard error	PC1-PairWise© Location (logit)	Standard error
Ferrero Rocher®	**1.080**	0.10	**1.409**	0.260	**1.403**	0.038
Lindor®	**0.780**	0.10	**0.552**	0.190	**0.608**	0.033
Caramel®	**-0.500**	0.10	**-0.720**	0.230	**-0.608**	0.033
Milky Way®	**-1.370**	0.10	**-1.239**	0.230	**-1.352**	0.037

The locations of the confectionery on a scale of specialness derived from the PC1 are consistent with those derived from the experiment using LS2011. Although the confectioneries exhibit identical ranks of individual stimulus, the standard error varies among the methods.

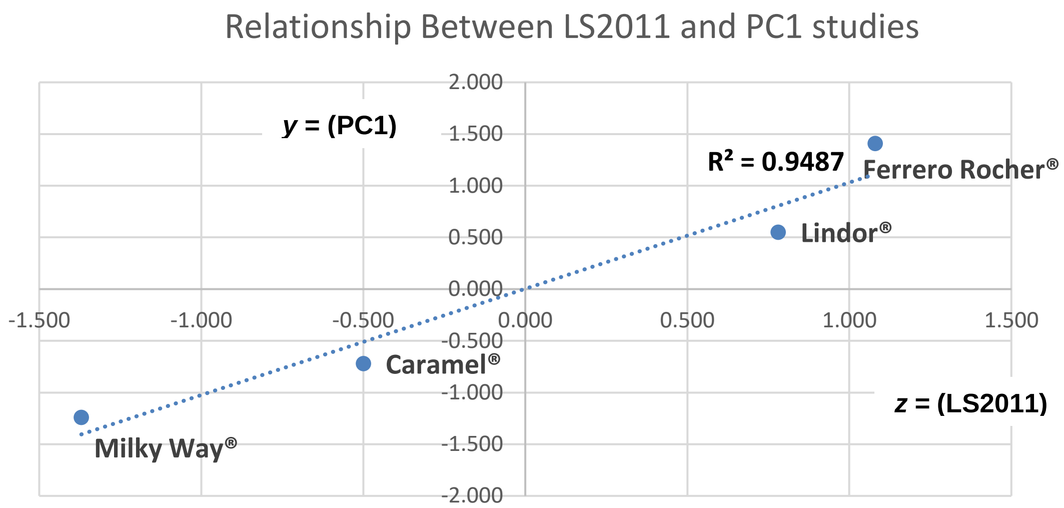


Figure 4.7.12: Relationship between PC1 and LS2011 study

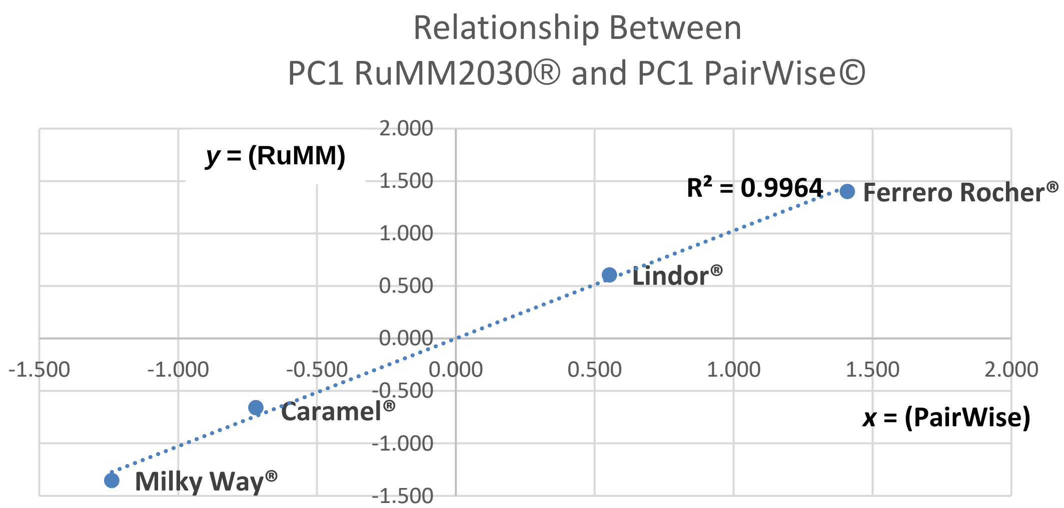


Figure 4.7.13: Relationship between RuMM2030® and PairWise©

The facets map, as shown in Figure 4.7.14, illustrates three graphs of stimuli facets mapped on the same logit scale. A logit is an expression of the probability that a particular item will be endorsed.

The graphical representation of this facet indicates the degree of how special the stimuli are from the overall perspective of the participants. The facet map represents the endorsement by hieratical rank. The value less than zero logits represents the lower region, denoting the lower degree of specialness, while values above zero logits represent the upper region, denoting a higher degree of specialness.

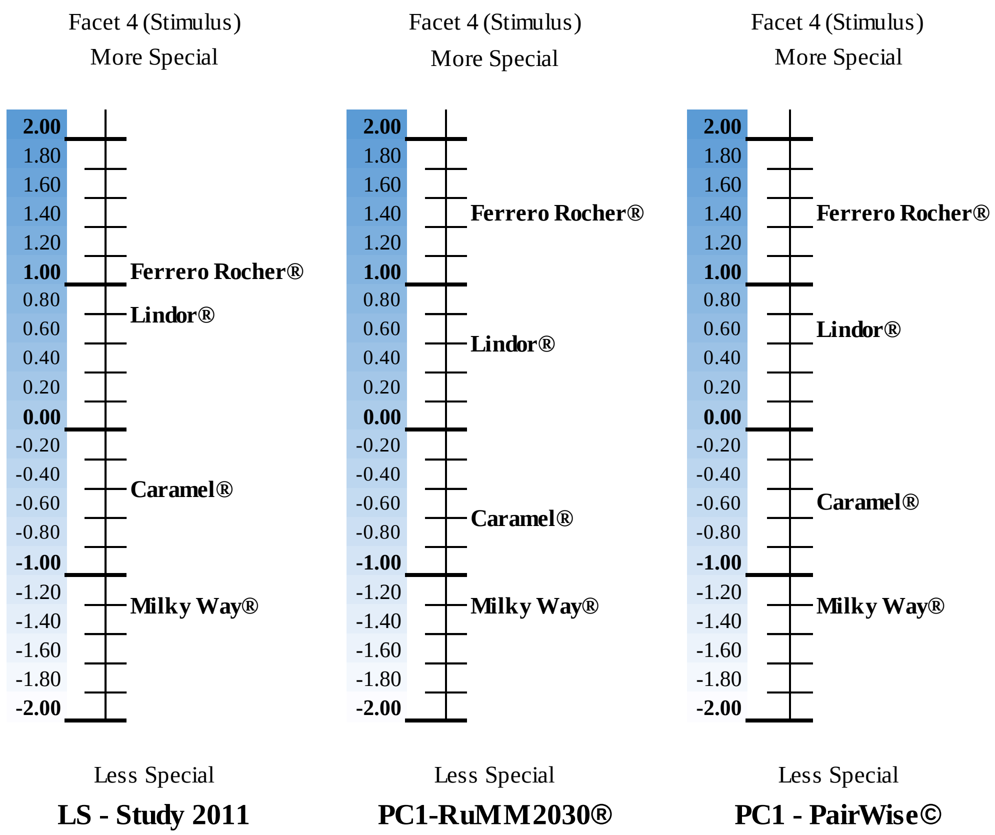


Figure 4.7.14: Comparison of locations of stimuli between methods

The facets map as shown in Figure 4.7.15 illustrates four graphs showing the locations of all facets on the same logit scale. A logit is an expression of the probability that a particular item will be endorsed. The first three columns represent the distribution along the continuum of person location, item location and pairwise location.

The first three columns illustrate a kind of yardstick symbolic of a distribution continuum which is divided into two regions separated by zero logits. The values less than zero logits represent the lower region, denoting the easiness of endorsement, while values above zero logits represent the upper region, denoting the difficulty of endorsement for the persons and items.

While the last column represents the specialness location of the particular stimulus. The values less than zero logits represent the lower region, denoting a lower degree of specialness, while values above zero logits represent the upper region, denoting a higher degree of specialness.

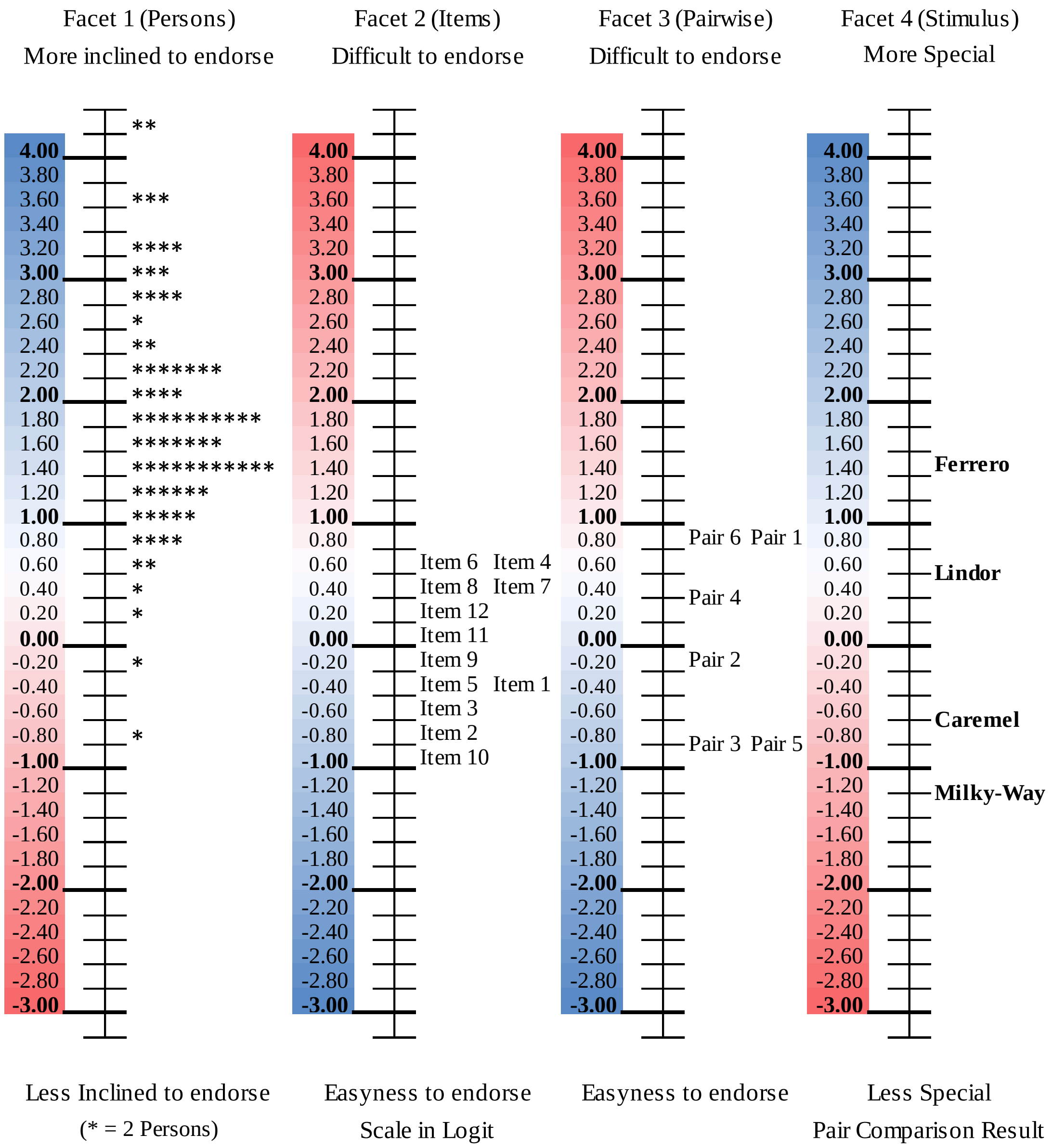


Figure 4.7.15: Facets map for the specialness of confectionaries

Person locations are plotted on the scale represented in the first column. Participants are plotted according to greater inclination to endorse the attribute specialness of the six pairwise confectioneries where the frequency distribution is located on the top of the scales. This graph illustrated that most of the participants are identified as higher-ability persons when endorsing the items easily. The second column illustrates the location of the items pool that is difficult to endorse, which is located at the top of the scales while easy to endorse is at the bottom of the scales.

The third and fourth columns illustrate the location of pairwise and individual confectioneries on the continuum, denoting a higher and lower endorsement of specialness. Pairwise stimuli on the continuum illustrate that pair six (Lindor® vs Ferrero Rocher®) and pair one (Milky-way® vs Caramel®) are the most difficult stimuli to endorse, whereas pair three (Milky-way® vs Lindor®) and pair five (Caramel® vs Ferrero Rocher®) are the most easily (Figure 4.7.15). The fourth facets illustrate the location of each stimulus on the continuum.

4.8 DISCUSSIONS

4.8.1 Stimuli bias

The discriminating factors that are likely reflect the brand and price positioning that influence the specialness attributes are explicitly recognised by the participants. Participants are able to clearly distinguish when comparing the opposing price and brand segment: for instance, in pair three (Milky-way® vs Lindor®) and pair five (Caramel® vs Ferrero Rocher®) because the pairwise combination holds greater contrast to discriminate clearly (facet three in Figure 4.7.15).

However, the participants find it quite difficult to distinguish when comparing within the same price segment. Pair six (Lindor® vs Ferrero Rocher®) and pair one (Milky-way® vs Caramel®) are the most difficult pairwise combinations (facet three in Figure 4.7.15).

This study speculates that the difficulties have arisen because of both pairwise options offer poor contrast to make distinctions, where the distance is quite close.

Procedures bias is another source of bias identified during data collection. The computer program does not indicate progress response features to notify how far participants have completed the questionnaire. This will make participants impatient to complete the PC1 questionnaires and at the same time introduces boredom perceptions and fatigue to some of the participants.

To minimise the effect size of bias, each item was seamlessly programmed to be in random order. However, this introduced negative perception among participants because they felt that the computer program attempted to check the consistency of their ability because every item displayed on the computer screen seems identical program appears identical items although it was not.

Stimuli bias has been identified in this survey. The quality of evaluating the specialness introduces some perception bias because the judgements are purely based on packaging aesthetics rather than taste experience.

Participants acknowledge that two of the four stimuli, Ferrero Rocher® and Lindor®, were classified as premium confectioneries, while Caramel® and Milky Way® were claimed to be regular chocolate bars. Some of the participants, especially international participants, acknowledged at least one of the four confectioneries that have been used in this study were not offered in their market region. This was probably because the particular brand was sold locally but not internationally. Some participants acknowledged that Lindor® was unattractively packaged compared with Ferrero Rocher®. However, the Lindor® tastes better than Ferrero Rocher® due to their purchasing experience of the brand.

Some of the participants suggest the experiment should allow them to taste the confectioneries for better judgement. However, because of the ethical issues of health and safety grounds, none of the participants could eat or taste the confectioneries during the study.

The result reported in identical fashion where participants could not tell whether or not these confectioneries appeared to be special when comparing between same price segment on the same scales: for example, between Lindor® and Ferrero Rocher®. A previous study using LS was also reported in identical fashion when participants hardly distinguished between the premium confectioneries when comparing within the same segment (Camargo, 2013).

4.8.2 PC1 bias

In general, PC1 demonstrates good calibrations in achieving the goodness of fit statistics in RM despite some challenge, such as there are some participants associated with positive high fit residual because participants find PC1 too easy to discriminate.

In this study, PC1 has to achieve good models as expected by the RM. This indicates binary scales offering low risks of missed targeting and statistical error. However, there are some biases and errors occurring in PC1 because the model does not work when the stimuli are too easy to distinguish.

The summary statistics demonstrate the good internal reliability as indicating good reliability and PC1 shows all the items are free from disordered thresholds. No DIF existed for both age and gender bias.

The outcome of this study has tested and validated the PC1 features. The experiment was conducted using the PC technique demonstrating the viability of using Rasch analysis to derive the minimum effect size of bias and error. This evidence has reflected the objectives of this study.

The calibrations process was used to fix the higher fit residual issue on the preliminary dataset. However, the calibrations had to suggest removing important datasets to fit into the model. Two calibrations techniques were impractical where this approach was an intolerance to one of the pairwise and stimuli — this approach suggests the analysis should either discard the Milky-way® stimulus or discard the pairwise one (Milky Way® – Caramel®) to meet a sufficient Chi-Square value. On other hand, the calibrations technique has to remove 6.39 percent or ten participants due to extreme persons and keep the high fit residuals. However, this index was a reasonable amount compared to the huge percentage lost in the LS1 and SDS1 studies.

In the last calibrations approach, the data were registered and validated using the developed software called PairWise© Software version (Humphry, 2010; Humphry et al., 2017), which was successful in obtaining individual logit for each stimulus. Table 4.8.1 below is a characteristics comparison between SDS and LS with PC, which was initially adapted from Schutte (2005) and Farsky et al. (2017).

Table 4.8.1: Comparison between PC and SDS and LS methods

	Likert and Semantic Scales	Pair Comparison
	Non-Comparative Scaling	Comparative Scaling
Type of Scales	Polytomous	Dichotomous
Characteristics	Summated SDS and LS	Equal-Appearing Intervals
Response type	Rating scale	Choice
Estimate methods	Mean Logit	Hierarchical or Rank order logit
Stimuli accessibility	Discrete Choice	Direct Comparison
Trade-off or Greater discrimination	No	Yes
Disordered thresholds	Yes	No
Required calibration on category scaling (Rescore)	Yes	No
Fast judgement	No	Yes
Complexity	Complex Choice	Simple Choice

4.8.2.1 The advantages of PC

Analysis of PC1 has demonstrated the evidence of the advantages of using PC compared with SDS and LS. Some of the key features were highlighted in this study, using PC to measure affective responses to confectionaries, and demonstrated a significant achievement in psychometric and statistical domains compared with SDS and LS.

1) Participants found it very easy to endorse even the most difficult items. However, PC1 introduces some disadvantages in that the items can become too easy to endorse inflated a large number of higher ability persons, to the extent it is suggested this group be removed as it holds a higher mean location.
2) In terms of the response rate, the PC1 test took an average of 4.25 to 6.17 minutes to complete all seventy-two pairs of comparison items. The individual response rate for each comparison item took an average of 3.7 to 6.1 seconds per item response. Compared to the response rate for SDS and LS, which took an average of twenty-four to thirty-five minutes to complete

the one hundred and forty items with an average speed of 12.64 seconds per item response, PC1 considerably faster in making a judgement.

3) Preliminary analysis of PC1 analysis required a minimum of a calibrations process to achieve the goodness of fit statistics in RM.
4) PC1 does not require calibrations on the category scales due to the characteristic of PC offering greater discrimination. The scale will automatically be ordered regardless of any distribution weight between pairwise combinations.
5) The results highlight the unique properties structure of PC where disordered thresholds are disabled or deactivate, which means there is no discriminating over the scales required to rescore the adjacent on the category scales. This feature is helpful to reduce the number of alterations while maintaining the originality of the dataset. Thus this analysis rejects the assumption the lack of invariance could be affected by disordered thresholds.
6) A small number of persons holding high fit-residuals were recorded in the PC1 study, where seven percent or eleven participants out of one hundred and fifty-seven participants are affected. As a comparison, SDS and LS had 43.4 percent or seventy-nine participants out of one hundred and eighty-two participants reporting higher fit residuals. High residual values offer an unexpected pattern response to the participants because the scales offer an unnecessary choice of category scaling.
7) The person-item distributions demonstrate that participants find it much easier to carry out paired comparisons than with SDS and LS where confectioneries are assessed individually against Likert statements.
8) Cronbach's alpha and PSI index exhibit a good level of internal consistency and reliability.
9) The DIF analysis shows there is no gender and age bias found in PC1. The result demonstrates there is no statistical difference in ability between the age, or male and female subgroups.
10) PC1 has demonstrated good linear measurement. The relationship outcome from affective responses using PC1 does not vary within the same context when using LS2011 and offering resemblance logits.

4.8.3 Test of local dependencies

The local dependencies or item dependence analysis was not performed in this chapter. Local dependencies refer to one of the tests to check whether the items in this study are free from dependencies: in other words, the items are independent and not related to each other. This test usually meets the assumptions in the RM and all IRT models. One of the common signs in the local dependency is where the items are constructed with similar passages or meanings, which is prevalent in the reading comprehension test that can be a potential source of local item dependence. The effect for the reader can be associated with redundancy, where the items were too predictable and that may result in biases (Purya Baghaei, 2008).

In this chapter, local dependence analysis was not performed at this stage in the study because of space limitation. Additionally, the objective of this chapter is to demonstrate the advantages of using PC to derive a better scaling structure to obtained goodness of fit statistics in RM measurement. However, the local dependence will be discussed in the post hoc analysis that required multiple testing, which will be outlined in the next chapter of this thesis.

4.8.4 Delimitation - Test of unidimensionality

Local dependencies have a strong correlation in unidimensionality tests because the objective of obtaining the independent variable is purposely to meet the assumptions of unidimensionality in RM. However, the unidimensional test was not replicated in this chapter because the item pools were initially taken from previous research (Camargo and Henson, 2011), which were unidimensionally fit.

4.9 SUMMARY

This work demonstrates the viability of using Rasch analysis to derive linear measurements of affective responses from paired comparisons of products, although there remain challenges. If the products are too different and participants find it too easy to discriminate between them along the affective dimension of interest, then it is likely that the data will be a poor fit to the RM. Participants in the research found it much easier to make paired comparisons than to evaluate the products separately against Likert statements.

Unlike SDS and LS, PC used direct comparisons to stimulate strong nonverbal images (Aaker and Biel, 1993). Visual stimuli make the task easier, realistic and less fatiguing for participants to respond (Green and Srinivasan, 1978). This is an essential element to reducing the burden, especially of dealing with difficult items and complexity in polytomous scaling. PC aids greater understanding for the participants to respond correctly and offers a lower risk of missed targeting. As a comparison, SDS and LS use discrete stimuli which resulted in respondent struggling to understand and use the category scales. Items without visual clues plus discrete comparisons affected the difficulty of judgement, especially when the items were not unidimensionally fit. The response will be complicated and maybe exposed to missed targeting.

The evidence demonstrates PC and LS has derived a good linear correlation, indicating that PC is a viable mechanism to extend the smaller contrasts of scaling structure in LS to a reasonable distance in PC. Then the response can be discriminated easily. However, in this study, the comparison using confectioneries may not be suitable as the products offer a greater contrast to discriminate clearly. Therefore, the next study will use vehicle interior texture to observe whether these stimuli provide reasonable magnitude in offering mid-range contrast in meeting statistical outcomes and deriving linearity against LS. In order to demonstrate the study using interior vehicle textures, the items must be unidimensionally fit in order to meet the statistical assumption in RM. In Chapter 5, the unidimensionality test was performed using the additional sample size.

Chapter 5

Identifying Unidimensionality items for paired comparison study

In Chapter five, the report features a LS version two, after this known as Likert scale study 2 (LS2). The aim of LS2 study is to examine how well the items behave in meeting the unidimensionality conditions as expected by the Rasch model (RM). Rasch analysis was used in this study to determine whether goodness of fit can be achieved if the items are unidimensional fit. Another assumption in the RM is a larger size sample increases greater stability in terms of goodness of fit compared to smaller sample sizes. In this LS2 study, an additional dataset of thirty-eight participants was added to the previous LS1 dataset which was done previously in Chapter three.

In total, one hundred forty-five datasets in LS2 were newly examined using two series of Rasch analyses using dual structure results: namely, facet analysis and non-facet analysis. Within non-facet analysis, two series of subtests were carried out to investigate the unidimensional features of the items, while facet analysis aimed to validate the statistical stabilities. The results of the LS2 study exhibit the better statistical fit. However, this study also reported a large number of items as problematic associated with response dependencies resulting in poor T-test results which signalled a lack of unidimensionality. The outcome of this study successfully determined nine items which are unidimensionally fit for the next studies using PC2.

5.1 INTRODUCTION

Validity is an important factor in any test instrument. Its value was demonstrated in the tests and surveys that were conducted. Without validity, the information is meaningless as we cannot be sure what to measure; validity requires precision measurement in order to achieve the true score that affects the reliability of the test instrument.

In latent space, one of the major challenges in survey design is to ensure the attributes are free from multidimensional variables. If any are present, the analysis is going to be problematic to fit in the RM because the properties from multidimensional variables can yield shortcomings due to the wide dimension, complexity and interconnected factors that increase the amount of multiple meaning.

Most of the error in data collection is not because the respondent does not understand the question but the items themselves inflate the amount of multiple meaning which are difficult to connect with single dimension connotations on LS. The respondent would rather use their own interpretation to answer the tricky questions, which inflates the number of inaccurate responses. This is one reason why RM is rejected if the items hold multidimensional traits or meanings.

In common approach, items being developed using affective words will normally be associated with multiple regression processes such as adjective reduction, FA and EPA dimension of rating response as evaluation, potency and activity to position the items to elicit multiple dimensions in relation to certain classification attributes. However, this process violates the measurement structure of unidimensionality in Rasch analysis and corrupts the statistical outcome because the RM works if the items are unidimensional to achieve linearity or normal distribution in statistics.

This study focuses on the development of the items’ goodness of fit rather than focusing on a person fit because items with unidimensional fit will naturally improve person location and minimise the fit residuals logits better accuracies. However, generating unidimensionality items is not an easy task and can be tricky; the items will behave in a way they were designed. The items might be troublesome if they carry multidimensional characteristics: for example, grammatical ambiguity, complicated vocabularies such as jargon terminology and items without true linguistic contras; all these problems will affect the way participant interpreted which may tend to be biases and error. Thus, it is important that items should be designed to convey the one dimensional attributes known as unidimensionality.

This study examines how well the items behave in a way that was designed for an LS2 study to be unidimensional as expected by the RM.

5.2 HYPOTHESIS OF THE STUDY

RM works when the items are unidimensional best. This study hypothesises in the unidimensionality properties will improve the target judgement and satisfy the measurement structure of RM.

One of the reasons why data could be troublesome when the items are exposed from local dependencies such as grammatical ambiguity which carries multiple interpretations and resulted the response could be distorted.

The item with unidimensionality fit able to minimised bias and error.

5.3 OBJECTIVES

The aims of the research in this chapter are therefore as follows:

1. To determine how well data obtained from affective responses using the additional sample of LS2 study offers a better power of statistical fit and delivers good stability as expected in RM theory.
2. To test and demonstrate items obtained from faceted RM in the LS2 study is unidimensionally fit, as expected in RM theory.
3. To investigate how unfit items associated with bias and error inflate multidimensional features and corrupt the measurement structure of RM theory.
4. How to minimise the problematic effect of multidimensional items and propose unidimensionally items fit for the next studies using PC2 approach (Chapter six).

5.4 METHODS

The AE method was not used to design the test instruments since this study added the sample set and replicated the same test instruments in the previous study as in Chapter three. This analysis, however, uses the RM and binomial test to determine unidimensionality features and validate the results.

5.4.1 Unidimensionality approaches

The unidimensional test using independent T-test in RUMM2030® and validated using Binomial Test. The unidimensionality test in this study was analysed in two T-test processes (Figure 5.4.1).

In the first T-test, twenty calibrated items were analysed using T-test and were assigned in positive-negative loading using PCA1. The items consisting of pairs of positive and negative components were an overlay or equate in T-test function.

5.4.2 Unidimensionality t-test procedure and process flow

All the items were considered unidimensionally fit if the T-test results gave lower than ninety-five percent of confident interval proportion (%LB95CI) or the PST is equal to or lower than five percent. This also indicated which items fit the RM. The unidimensionality test was carried out in the RM in several steps. The first step was to register the raw data in non-facet mode or an individually textured sample. The next step to clean the raw dataset of its missing data, extreme items and higher fit residuals, and iteration for disordered thresholds known as a calibrated dataset.

Next, the calibrated dataset was assigned and loaded onto PCA1 to get the hierarchical positive–negative items ranking. Using equating test or T-test function in RUMM2030® interface software allows selection and overlay both pair positive items and negative items.

This will give a total sample and the number of observed numbers at lower PST of five percent. The result of T-test analysis was summarised on the computer screen with a graphical representation for both loading plots. The binomial test was used to check the value of the total sample and the amount of observed numbers at the lower PST of five percent. The binomial test will calculate the PST and lower ninety-five percent of the confident interval proportion (%LB95CI).

The unidimensionality result will be consider accepted if the lower ninety-five percent of confident interval proportion is equal or less than five percent; otherwise, it will fail the unidimensional test.

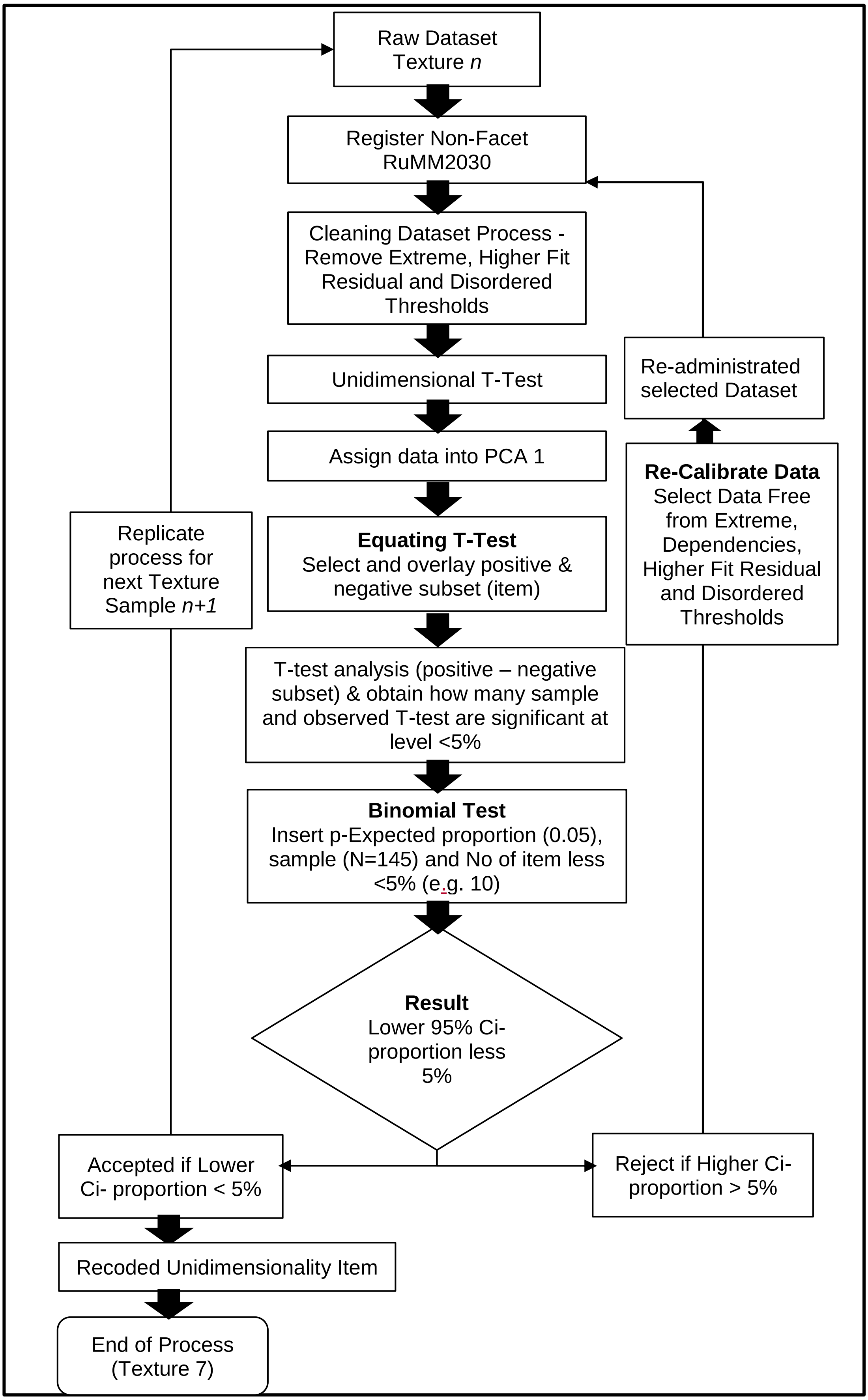


Figure 5.4.1 Unidimensional T-test Process Flow

5.4.3 Dual structure results

The additional dataset of thirty-eight participants was enhanced from the previous study completed in Chapter three. In total one hundred forty-five LS2 datasets were newly examined using the RM analysis. All the analyses were examined and compared with the original version of the LS1 and LS2 studies. Dual structure results were used to perform these studies: namely, facet analysis and non-facet analysis Figure 5.4.2. In the facet design, the aim of this analysis is to validate the statistical stabilities within all facet items and facet stimuli and compare the statistical differences in the context of the power of statistical fit, reliability and person-item separation as well as the items and stimuli facet logit between the LS1 and LS2 studies.

The reason the dual structure of results has been introduced in this chapter is that the facet analysis in the first structure failed to perform the unidimensional test because the quantity of local item in dependency matrix is too large and ties multiple items and stimuli which too complex in analysing. The multiple items blocks generated a 19600 correlation matrix that increases the complexity in calibrating and process the unidimensionality features. Thus, the second analysis, data was grouped and performed using non-facet analysis. Two series of subtests were carried out to investigate the unidimensional features of the items with independent stimuli.

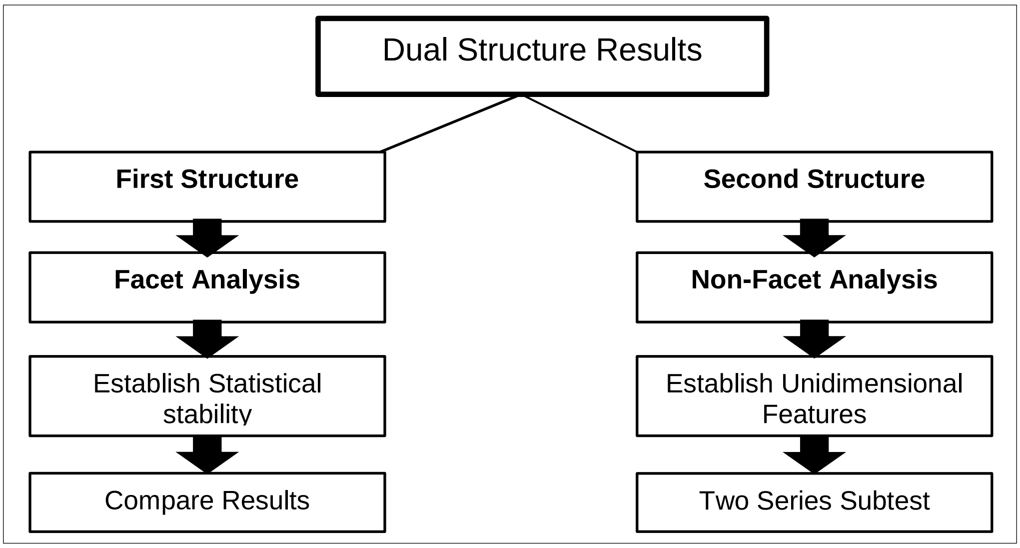


Figure 5.4.2 Dual structure results of facet and non-facet analysis

5.5 ETHICAL APPROVAL, CONSENT FORM AND RISK ASSESSMENT

The extension of ethical approval was granted from The University of Leeds with approval reference no. MEEC 14-025 version 2, conforming to participants' health and safety procedures. The ethical approval, consent forms and risk assessment are included in Appendix B.

The protocol for LS2 study was briefly explained to the participants. Information about instructions or demonstration of touching texture samples and a demonstration of ticking the semantic box in the questionnaire was also briefly explained. The protocol and questionnaire were also included in the Appendix C.

5.5.1 Data collections

In the previous LS1 study research in Chapter 3, one hundred and seven participants were asked to rate seven pieces of vehicle interior textures against twenty statements on a five-point scale.

For the extension LS2 study, thirty-eight participants (21.05 percent females and 78.94 percent male), between twenty to forty-two years of age (mean 29.3 and SD 5.98), were additionally recruited for this study. All the additional participants consisted of staff, researchers, postgraduates and undergraduates students at the University of Leeds.

The participants were asked to evaluate the quality of interior vehicle trims designs with twenty LS2 adjective statements against seven texture stimuli, this eventually creating in total hundred forty items.

The study was held in the Affective Engineering Laboratory in the Mechanical Engineering Faculty, University of Leeds. All participants were compensated for their time with £5. Each session was no longer than forty-five minutes.

5.6 APPARATUS AND EXPERIMENT SETTING

The study used seven stimuli in LS2 data collection. The stimuli were specially prepared using standard automotive plastic PP with a standard dimension of 225mm x 160mm x 3mm and were engraved with sandblasting, organically to a geometric pattern that is mostly seen in the interior of cars. Each stimulus was injected with a semi-gloss level between 1.5GU to 2.2GU in various colours such as black, matte black, brown and beige with colour pigments labelled one to seven (Figure 3.6.1, page 54)

The stimuli went through several stages of selection before being used as stimuli for data collection. To encourage the participants' understanding of the research context, all participants were required to pay attention to a few short videos about the driving experience. Three standing panels were used to evoke the study context, which illustrates the vehicle exterior and interior.

Table 5.6.1 Coding for texture samples

Texture Code	Label in LS2 questionnaire
T1	Texture 1 (K)
T2	Texture 2 (W)
T3	Texture 3 (G)
T4	Texture 4 (J)
T5	Texture 5 (R)
T6	Texture 6 (P)
T7	Texture 7 (Q)

5.6.1 LS2 Items

A similar set of LS1 questionnaires that used in Chapter three has been used for this study. Twenty statements were replicated randomly across seven pages equal to seven stimuli or texture samples.

For each texture sample, participants were asked to decide which linguistic quantifier was most suitable to describe how they felt about the texture sample if it were used in the design of the vehicle interiors.

The LS2 typically established by convention consists of a five-point scale that represents five response categories within two degrees disagreement to the left and two degrees agreement to the right, split by a neutral point. This is distinguished by a linguistic quantifier described as strongly disagree, disagree, neutral, agree and strongly agree

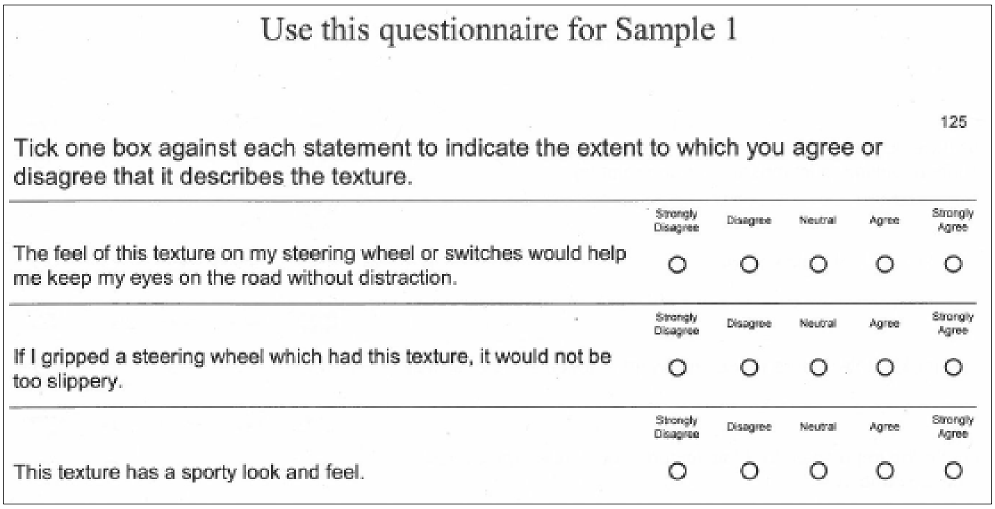
Each threshold contains values from zero to four, representing magnitude estimation on a continuum endorsing particular items. In RUMM2030®, the response categories were therefore coded where the greater point would represent a higher degree of agreement while the smaller point indicates a higher degree of disagreement.

Table 5.6.2 Original pool of items for LS2 study

Code	Descriptions
1	Before I touch this texture, I can see that it would feel grippy.
2	I have the impression that this texture would make my car feel spacious and neat.
3	I have the impression this texture is modern and contemporary-looking.
4	I would expect to see this texture with a good touch and feel in a reasonable price car.
5	If I gripped a steering wheel which had this texture, it would feel very safe.
6	If I gripped a steering wheel which had this texture, it would not be too slippery.
7	The feel of this texture on my steering wheel or switches would help me to keep my eyes on the road without distraction.
8	The feel of this texture would help me feel confident with my driving.
9	The look of this texture makes me want to touch it straight away.
10	This texture has a sporty look and feel.
11	This texture looks nice quality.
12	This texture does not look overly cheap and plasticky.
13	Touching this texture feels pleasant.
14	Touching this texture is relaxing.
15	Touching this texture makes me feel warm.
16	Touching this texture would make me feel connected when operating the switches in the vehicle.
17	Vehicle controls with this texture would give good feedback when shifting, pulling, turning and rotating.
18	When I touch this surface I get a sensation of luxury.
19	With this texture, I would be able to operate the controls without needing to look.
20	With this texture, I would feel comfortable inside the car.

Table 5.6.3 Linguistic quantifier and coding for five-point LS2

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
0	1	2	3	4



Use this questionnaire for Sample 1

125

Tick one box against each statement to indicate the extent to which you agree or disagree that it describes the texture.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
The feel of this texture on my steering wheel or switches would help me keep my eyes on the road without distraction.	O	O	O	O	O
If I gripped a steering wheel which had this texture, it would not be too slippery.	O	O	O	O	O
This texture has a sporty look and feel.	O	O	O	O	O

Figure 5.6.1 Five-point LS2 items

5.7 FACET ANALYSIS RESULTS

5.7.1 Summary of fit statistics

Initial testing of goodness of fit summarises the overall model fit, including the individual person fit and items fit, the exploration of thresholds and the probability curves of the LS2 study. In this study, the Chi-Square value of $p \leq 0.01$ indicates a lack of the desired scale of invariance, where some of the items not working as expected at discrete group levels are known as class intervals (Table 5.7.3 column *p* – page 155). This result also indicates a formal test of the LS2 was not invariance. The power of analysis fit indicated that there was a good degree of fit of the items in the dataset.

5.7.1.1 Reliability – Internal consistency

The initial fit statistics as shown in (Table 5.7.3 – page 155 column PSI and α) exhibit the exploration of person fit which the power of test fit was the indicator of reliability and was represented by the Cronbach's alpha (α) and PSI. The preliminary analysis of PSI was reported as 0.927 and illustrated the good internal consistency with the ability to statistically differentiate at least three ability groups. The class interval was equally distributed between three ability groups across the number of sample size (n145). The α value was reported to be 0.924, which indicates a good level of reliability.

5.7.1.2 Item location and fit residuals

By default, the RUMM2030® software will automatically assign all the items in which the location mean was centralised at zero (Table 5.7.3 – page 155). The items logarithm or log are on the same continuum along the x-axis using a common unit of termed logits.

While the result of this study shows the fit residuals mean was 0.443 and the SD was 0.534, these indicate that there was a goodness of fit of the items in the data as these values fall into standardised item-fit within the conventionally accepted range of +/- 2.50. There are no negative residuals, as determined in this analysis, which is normally associated with high item-total correlation in classical test theory. This would normally be interpreted to indicate the redundancy or over-discriminating of the items.

5.7.1.3 Person-location and fit residuals

The results of this study show the person location mean was 0.245 and the SD was 0.309 (Table 5.7.3 – page 155). These indicated that there was a good fit of the person location in the data, as these values fall into standardised item-fit between approximately zero and SD of approximately one. However, for the exploration of the person fit residual, the mean value was recorded as -0.403 and the SD as 3.705. This value associated some misfits where participants indicated a lack of the expected probabilistic relationship among the items within a scale. Higher SD is normally associated with a higher fit residual. This value indicated that generally, the individual person is not behaving in the same way as the other person when responding to the test. However, there were a few extreme participants identified and removed.

5.7.1.4 Facet analysis – By location order

The LS2 study involved two facet analyses using the RM. The first facet indicated twenty items of questionnaires that corresponded to the seven stimuli shown on the second facet. Table 5.7.4 – page 156 shows the calibrated analysis of two facet designs after alteration by location order. On the left facet, the column showed item location discriminated by the level of difficulty. While on the right facet, the column showed stimuli from lower degree endorsement to higher degree endorsement.

The calibrations facet illustrates item twenty *(with this texture, I would feel comfortable inside the car)* and item eighteen *(when I touch this surface I get a sensation of luxury)* was the easiest and most difficult to endorse respectively. While stimuli two and one were the lowest degree endorsement and highest perceptiveness, respectively.

5.7.2 Calibrated statistics

The RM allows some alteration of the preliminary dataset in LS2. The fit statistical model demonstrates better Chi-square value. The accepted Chi-square value of item-trait interaction in LS2 indicates a good degree of fit between the data and the model which reflects the property of invariance across the trait.

5.7.2.1 Summary of calibrated fit statistics

The calibrated Chi-square interaction value of $p \leq 0.154$ indicates goodness of fit interaction on the desired scale, indicating that there is no significant deviation between the observed data and what was expected from the model (Table 5.7.3 – page 155 data four column p). This result also indicates a formal test of the LS2 showed invariance and the items working as expected at discrete group levels known as class intervals. The power of analysis fit also indicated that there was a good degree of fit of the items in the dataset.

5.7.2.2 Reliability – Internal consistency

The calibrated fit statistics as shown in (Table 5.7.3 – page 155 data four column PSI and α) exhibit that PSI was reported as 0.950 and illustrated very good internal consistency compared with preliminary analysis. The α value was reported to be the same manner, at 0.927, which indicates a good level of reliability after calibrations have been done.

5.7.2.3 Item location and fit residuals

The calibrated items are fit to show the location of the item mean value at zero and the SD value at 0.751 (Table 5.7.3 – page 155 data four). Although the SD was slightly higher than preliminary study, the value indicated that there was a good fit of the items in the data, as these values fall into the standardised item-fit range between approximately zero and the SD of approximately one.

Meanwhile the result for the residuals mean value was 0.233 and the SD was 0.404. These indicated that the calibrated results inflate a good fit of the items in the data compared to preliminary analysis.

5.7.2.4 Person-location and fit residuals

The calibrated person fit exhibits the person location mean value at 0.463 and the SD at 0.436 (Table 5.7.3 – page 155 data four). Although the value was slightly higher than in the preliminary study, the value indicated that there was a good fit for the items in the data.

However, for the exploration of person fit residuals, the mean value was recorded at -0.258 and the SD at 2.970. Although the value was slightly better than in the preliminary study, this value was associated with a misfit, where participants indicated a lack of the expected probabilistic relationship among the items within a scale. Higher SD is normally associated with higher fit residuals. This value indicated that generally, the individual person is not behaving in the same way as the other person when responding to the test. These results suggested there were a few unfit participants that were identified and needed to be removed.

5.7.3 Statistical comparison between LS1 and LS2 scales

Another objective in facet design is to determine whether observed data from affective responses using the additional sample in LS2 study improved the power of fit statistics to hold the statistical stability in comparison with the previous LS1 study. Thus, this analysis examined both fit statistical results for comparative analysis.

Data from affective responses using the additional sample in the LS2 study has improved the power of fit statistics compared with the LS1 study. In the context of Chi-square value, person-item separation, reliability, items fit and offering better stimuli, the logit demonstrated better results and matched with sample size recommended by (Linacre, 1994) as in Table 5.7.1.

Table 5.7.1 Sufficient sample size for Items and person stability (Linacre, 1994)

Item calibrations of person measure stable within	Confidence	Minimum sample size range (Best to poor targeting)	Size for most purposes
+/- 1 logit	95%	16 - 36	30 (minimum for dichotomies)
+/- 1 logit	99%	27 - 61	50 (minimum for dichotomies)
+/- ½ logit	95%	64 - 144	100 (*n*107)
+/- ½ logit	99%	108 - 243	150 (*n*145)
Definitive or High Stakes	99% (items)	250	250
Adverse Circumstances	Robust	450 upwards	500

5.7.3.1 Better Chi-square value

The LS2 analysis results in Table 5.7.3 – page 155 showed a better statistical effect in most of the fit statistics compared with the LS1 analysis in Table 5.7.5 – page 157. The additional sample in the LS2 study offers better Chi-square value in both the preliminary and calibrated analyses. As the first comparison, the Chi-square was reported at p 0.154 for an LS2 study, which was greatly improved from the LS1 at p 0.01. This indicates that increasing the sample size produced the desired scale invariance wherein some of the items worked as expected at discrete group levels known as class intervals.

5.7.3.2 Better person-item separation

The internal reliability and PSI are also holding resembles value with a slight improvement on the LS2 study. The LS2 study was reported as good person-item separation stability at 0.950 for calibrated LS1 (Table 5.7.3 – page 155) and 0.953 for calibrated LS2 analysis (Table 5.7.5 – page 157).

5.7.3.3 Good reliability

The LS2 study also reported the α provides a better reliability value of 0.933 for the LS1 (Table 5.7.3 – page 155) and 0.927 for the LS2 studies (Table 5.7.5 – page 157).

5.7.3.4 Better fit residuals

The LS2 study also reported that the person-item fit residuals provide better item fit residuals of 0.233 compared to 0.300 on the previous LS1 study (Table 5.7.3 – page 155). The person fit residuals value increased to -0.258 from -0.548 in the previous LS1 study (Table 5.7.5 – page 157).

5.7.3.5 Individual stimulus logit illustrates high correlation logit

The study exhibits a strong correlation of the stimuli mean location logits with data observed, obtained by both LS1 and LS2 study (Table 5.7.2). The correlation value at R_2 0.99 (Figure 5.7.1), which indicates the validity of the additional sample, does not illustrate much difference in providing a similarity in responding patterns for both LS1 and LS2 studies.

Table 5.7.2 Stimuli mean location between LS1 and LS2 study

Code	LS1	LS2
T1	0.348	0.323
T2	-0.274	-0.287
T3	-0.058	-0.037
T4	-0.031	-0.052
T5	-0.165	-0.111
T6	0.340	0.310
T7	-0.159	-0.146

The facet map illustrates graphical representation for both studies. The position of the stimuli was not consistent. However, the LS2 study exhibited better distribution compared with the LS1 study, which indicates increasing the sample size increases good distribution spread and targeting. Details person-item and stimuli comparison are illustrates in Figure 5.7.2.

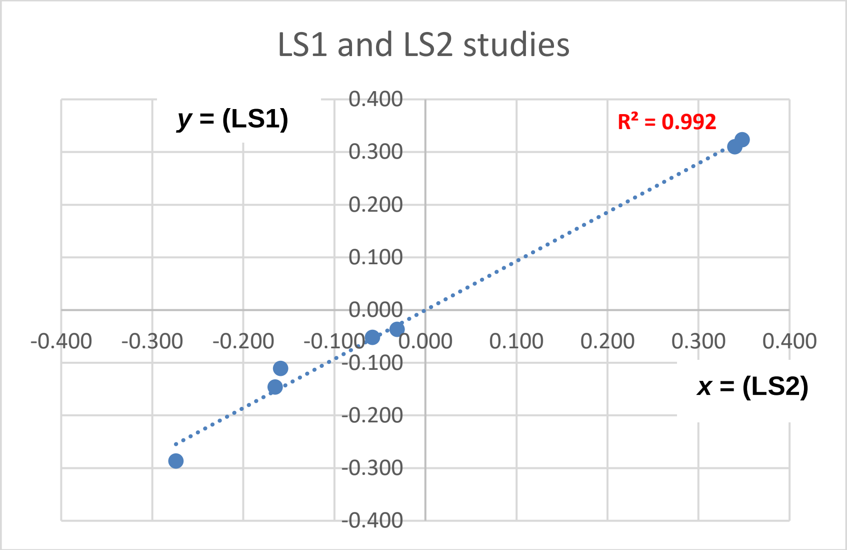


Figure 5.7.1 Strong correlation between LS1 and LS2 study

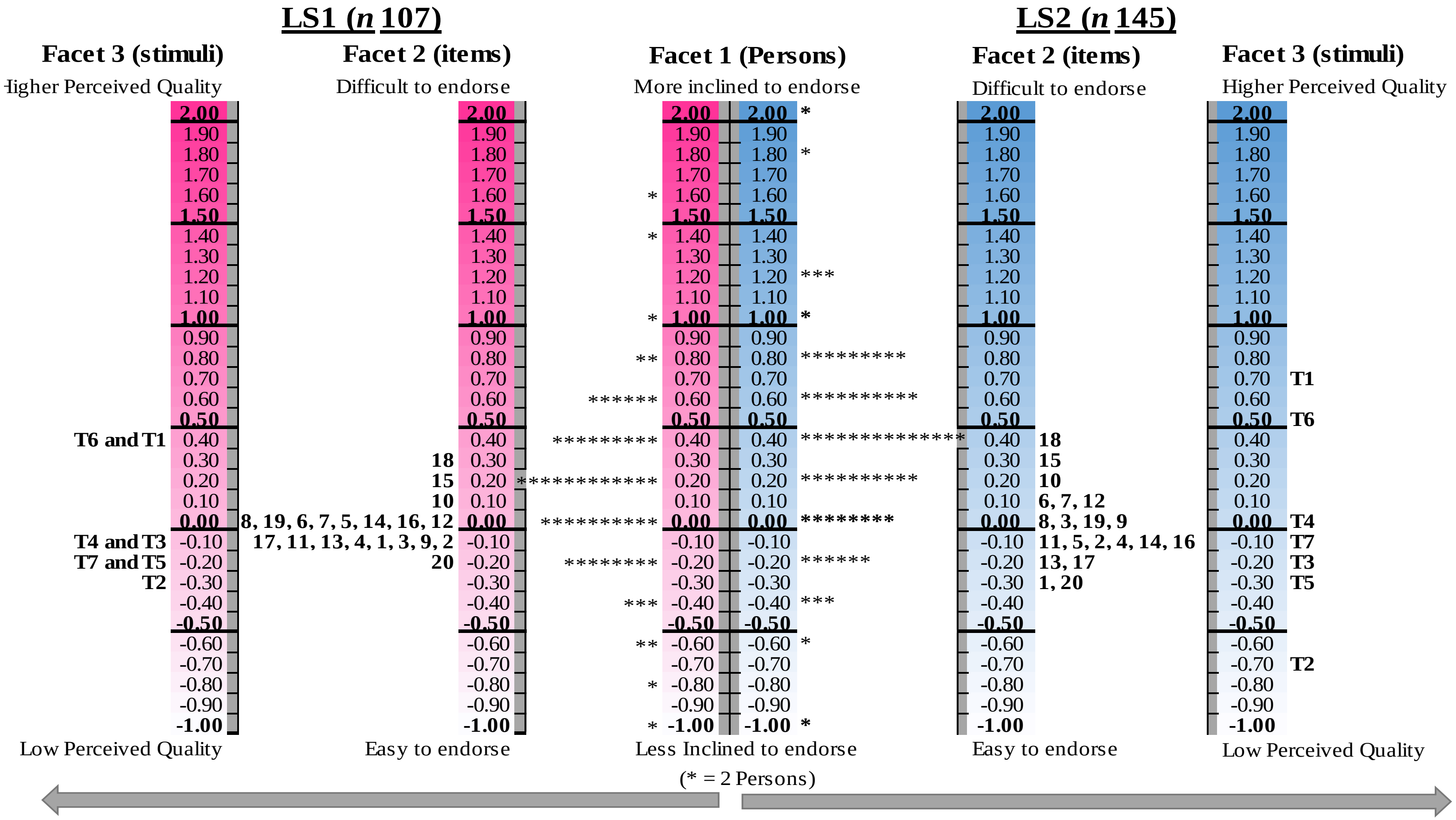


Figure 5.7.2 Stimuli Logit between LS1 and LS2 study

Table 5.7.3 Fit Statistics of Facet Analysis for LS2 study (*n*145)

Analysis	Item Loc		Person Loc		Item Fit Res		Person Fit Res		Chi-Square Interaction				
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Value**	**dF**	***p***	**PSI**	**α**
Initial Data (n145)	0.000	0.299	0.245	0.309	0.443	0.534	-0.403	3.705	342.4	280	<0.006	0.927	0.924
Calibrated Data 1 (n145)	0.000	0.381	0.292	0.394	0.273	0.435	-0.627	3.802	327.5	280	<0.027	0.952	0.924
Calibrated Data 2 (n145)	0.000	0.381	0.301	0.400	0.255	0.423	-0.652	3.815	338.2	280	<0.009	0.949	0.924
Calibrated Data 3 (n145)	0.000	0.386	0.302	0.402	0.249	0.423	-0.659	3.822	319.2	280	<0.053	0.949	0.924
Calibrated Data 4 (n131)	0.000	0.751	0.463	0.436	0.233	0.404	-0.258	2.970	304.0	280	<0.154	0.950	0.927

Table 5.7.4 Preliminary Facet Analysis LS2 study (*n*145)

Item Facet				Stimuli Facet				Metric			
Item	**Locn**	**SE**	**Fit Res**	**Stimulus**	**Locn**	**SE**	**Fit Res**	**X^2**	**df**	***p***	**PSI**
1	-0.044	0.08	0.705	Texture 1	0.323	0.08	0.413	342.4	280	0.01	0.93
2	-0.001	0.08	0.327	Texture 2	-0.287	0.09	0.354				
3	-0.027	0.08	0.726	Texture 3	-0.037	0.08	0.006				
4	-0.111	0.08	0.542	Texture 4	-0.052	0.08	0.670				
5	0.041	0.09	0.282	Texture 5	-0.111	0.08	0.207				
6	0.039	0.09	0.282	Texture 6	0.310	0.08	0.972				
7	-0.010	0.09	0.010	Texture 7	-0.146	0.08	0.458				
8	-0.068	0.09	0.010								
9	0.004	0.07	0.423								
10	0.111	0.07	0.760								
11	-0.075	0.08	0.475								
12	0.060	0.07	0.087								
13	-0.134	0.08	0.299								
14	0.087	0.08	0.085								
15	0.191	0.08	0.306								
16	-0.011	0.09	0.369								
17	-0.138	0.09	0.228								
18	0.218	0.07	0.893								
19	0.014	0.09	0.215								
20	-0.146	0.09	0.181								

Table 5.7.5 Fit Statistic of Facet Analysis for LS1 study (*n*107)

Analysis	**Item Loc**		**Person Loc**		**Item Fit Res**		**Person Fit Res**		**Chi-Square Interaction**				
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Value**	**dF**	***p***	**PSI**	**α**
Initial Data (N107)	0.000	0.320	0.226	0.335	0.450	0.526	-0.295	3.595	336.7	280	<0.011	0.937	0.934
Calibrated Data 1	0.000	0.413	0.277	0.435	0.284	0.433	-0.564	3.771	342.2	280	<0.006	0.954	0.934
Calibrated Data 2	0.000	0.414	0.277	0.435	0.286	0.442	-0.565	3.785	342.2	280	<0.006	0.953	0.933
Calibrated Data 3	0.000	0.414	0.278	0.436	0.281	0.435	-0.567	3.773	342.2	280	<0.006	0.953	0.934
Calibrated Data 4	0.000	0.417	0.271	0.431	0.300	0.496	-0.548	3.776	352.0	280	<0.002	0.953	0.933
Calibrated Data 5	0.000	0.413	0.277	0.435	0.286	0.442	-0.565	3.780	342.2	280	<0.006	0.953	0.933
Calibrated Data 6	0.000	0.413	0.278	0.436	0.281	0.435	-0.567	3.773	342.2	280	<0.006	0.953	0.934
Calibrated Data 7	0.000	0.417	0.271	0.431	0.300	0.496	-0.548	3.776	352.0	280	<0.002	0.953	0.933
Calibrated Data 8	0.000	0.413	0.277	0.435	0.286	0.442	-0.564	3.780	342.2	280	<0.006	0.953	0.933
Calibrated Data 9	0.000	0.413	0.278	0.435	0.281	0.435	-0.567	3.773	342.2	280	<0.006	0.953	0.934
Calibrated Data 10	0.000	0.417	0.271	0.431	0.300	0.496	-0.548	3.771	639.0	560	<0.011	0.953	0.933

Table 5.7.6 Preliminary Data Facet Analysis for LS1 study (*n*107)

Item Facet				**Stimuli Facet**				**Metric**			
Item	**Locn**	**SE**	**FitRes**	**Stimulus**	**Locn**	**SE**	**Fit Res**	**X^2**	**df**	**p**	**PSI**
20	-0.165	0.10	0.162	Texture 2	-0.274	0.10	0.374	336.7	280	0.01	0.94
17	-0.152	0.10	0.243	Texture 5	-0.165	0.10	0.320				
11	-0.113	0.09	0.479	Texture 7	-0.159	0.09	0.452				
13	-0.109	0.10	0.279	Texture 3	-0.058	0.10	0.035				
1	-0.086	0.09	0.747	Texture 4	-0.031	0.09	0.655				
4	-0.085	0.10	0.513	Texture 6	0.340	0.09	1.000				
8	-0.037	0.10	0.107	Texture 1	0.348	0.09	0.326				
3	-0.026	0.09	0.719								
2	-0.023	0.10	0.343								
7	-0.003	0.10	0.457								
9	0.003	0.09	0.523								
16	0.028	0.10	0.385								
19	0.028	0.10	0.219								
6	0.039	0.09	0.552								
12	0.044	0.09	0.872								
5	0.051	0.10	0.272								
14	0.067	0.10	0.131								
10	0.113	0.09	0.724								
15	0.182	0.10	0.376								
13	0.243	0.08	0.943								

5.8 NON-FACET ANALYSIS RESULTS

5.8.1 Summary of statistics

Initial testing of fit summarises the overall model fit including the individual person fit and item fit, the exploration of thresholds and the probability curves of the LS2 study of one hundred forty-five participants.

In this study, the Chi-square value of $p \leq 0.01$ indicates a lack of the desired scale invariance where some of the items not working as expected at discrete group levels known as class intervals. This result also indicates a formal test of the LS2 showed no invariance.

The calibrated result exhibited a good fit of Chi-square item-trait interaction values of $p \leq 0.44$, $p \leq 0.39$ and $p \leq 0.26$ for textures one, four and seven respectively while textures two, three, five and six against the calibrated Chi-square value exhibited a poor fit, when the value is equal or less than $p \leq 0.01$ (Table 5.8.1).

Table 5.8.1 Chi-Square results

Texture	Preliminary	Calibrated
T1	$p < 0.01$	$p < 0.44$
T2	$p < 0.01$	$p < 0.01$
T3	$p < 0.01$	$p < 0.01$
T4	$p < 0.04$	$p < 0.39$
T5	$p < 0.01$	$p < 0.01$
T6	$p < 0.01$	$p < 0.01$
T7	$p < 0.01$	$p < 0.26$

5.8.1.1 Reliability – Internal consistency

The initial fit statistics, as shown in Table 5.8.2, exhibit the exploration of person fit in which the power of the test fit was the indicator of reliability and was represented by the α and PSI. The preliminary and calibrated analysis of PSI was reported for all stimuli with more than 0.9 illustrating good internal consistency, which includes the ability to statistically differentiate between at least three ability groups.

The class interval was equally distributed between three ability groups across the number of a sample size of one hundred forty-five participants. The α value was reported to be 0.9, which indicates a good level of reliability.

Table 5.8.2 Calibrated PSI and α

Texture	Preliminary PSI	α
T1	0.935	0.929
T2	0.935	0.929
T3	0.909	0.910
T4	0.927	0.927
T5	0.906	0.909
T6	0.926	0.922
T7	0.933	0.940

5.8.1.2 Item fit and fit residuals

The results of this study exhibiting the item location to the model have examined, via individual item log residuals, the test of fit statistics of the minimum 0.071 to the maximum logit of 0.742 that falls within the acceptance range of -2.50 to 2.50 (Table 5.8.3). These indicated that there was a good fit for the items in the data. Some improvement was recorded after the calibrations were made.

Table 5.8.3 Item fit Residuals across the seven samples

Texture	Preliminary	Calibrated
T1	0.742	0.481
T2	0.072	0.071
T3	0.260	0.183
T4	0.365	0.278
T5	0.433	0.274
T6	0.460	0.074
T7	0.481	0.283

5.8.1.3 Person fit and fit residuals

The results of this study exhibiting the person location to the model were examined via individual item log residuals test of fit statistics of the minimum -0.257 to 0.577, the maximum logit that falls within the acceptance range of -2.50 to 2.50 (Table 5.8.4). These indicated that there was a good fit for the items in the data. Some improvement was recorded after the calibrations were made.

Table 5.8.4 Person Fit Residuals across the Seven Samples

Texture	Preliminary	Calibrated
T1	-0.305	-0.353
T2	-0.577	-0.432
T3	-0.544	-0.422
T4	-0.376	-0.356
T5	-0.479	-0.257
T6	-0.340	-0.413
T7	-0.353	-0.294

5.8.2 Response dependencies

In total one hundred seventy-six items was found to show problematic dependencies across seven texture samples in the preliminary analysis in this study. From twenty items registered in this study, three items were detected as the most problematic: item eighteen *(When I touch this surface I get a sensation of luxury),* item nineteen *(With this texture I would be able to operate the controls without needing to look) and* item twenty *(With this texture, I would feel comfortable inside the car)* demonstrating the most problems associated with some of the items in all textures.

In detail, item eighteen demonstrates the most problems associated with some of the items in textures four, five and seven. Item nineteen demonstrates the most problems associated with some of the items in textures one and three. While item twenty demonstrates the most problems associated with some of the items in textures two and six (Table 5.8.5).

Details of which items are affected by dependency items were illustrated in Appendix A (Subtest 1 – page 275) Table 8.3.3, Table 8.3.6, Table 8.3.9, Table 8.3.12, Table 8.3.15, Table 8.3.18 and Table 8.3.21 for the stimuli one, two, three, four, five, six and seven respectively.

Table 5.8.5 Response Dependency Items

Item holds Local Dependency													
Texture 1		Texture 2		Texture 3		Texture 4		Texture 5		Texture 6		Texture 7	
Item	Qty	Item	Qty	Item	Qty	Item	Qty	Item	Qty	Item	Qty	Item	Qty
1	0	1	0	1	0	1	0	2	0	1	0	1	0
2	0	2	0	2	0	2	0	4	0	2	0	3	0
3	0	7	0	3	0	3	0	7	0	3	0	5	0
4	0	8	0	4	0	4	0	11	0	4	0	13	0
5	0	15	0	8	0	9	0	13	0	9	0	4	1
9	0	3	1	9	0	13	0	14	0	10	0	7	1
13	0	5	1	15	0	15	0	20	0	15	0	8	1
15	0	9	1	20	0	5	1	3	1	5	1	12	1
6	2	14	1	5	1	7	1	8	1	8	1	14	1
11	2	16	1	10	1	10	1	19	1	11	1	17	1
12	2	4	2	13	1	11	1	9	2	13	1	6	2
8	3	6	2	6	2	14	1	12	2	14	1	9	2
7	3	12	2	11	2	20	1	18	3	6	2	10	2
14	3	13	3	12	2	6	2			16	4	16	2
16	3	17	3	16	2	12	2			17	4	19	2
20	3	18	3	18	3	8	3			18	4	20	3
18	4	19	3	14	4	17	4			20	4	18	4
17	4	20	5	17	4	16	4						
19	5			19	6	19	4						
						18	5						
	34		28		28		30		10		23		23

5.8.3 Unidimensionality results

5.8.3.1 First t-test and binomial result

The unidimensionality of the scale was examined using an independent T-test procedure (Figure 5.4.1, page 141). The series of T-test procedures exhibited that the items across all textures were not unidimensional as shown in the summary of statistics Table 5.8.6.

The results exhibited that paired T-test is not significant at p = 0.05. This gives the lower ninety-five percent of (%LB95CI) or the PST is greater than five percent. This indicates the items in all seven textures samples are problematic in unidimensionality construct.

The poor results are inflated by the problematic response dependencies within the items set which indicate the items highlighted constructing similar meaning with the associated items. This led to an exploration of the individual item fit.

The unidimensionality was also validated using the binomial test after T-test protocol was made. The binomial test was calculated based on sample size and the number of observed samples to get the proportion of significant test value.

The unidimensionality test exhibited some of the items in all texture samples were greater than the expected unidimensionality or binomial test of five percent; therefore, the unidimensionality result was not accepted by the model. It can be speculated that the test requires a large sample size (minimum *n* 200) and in this test the number of samples was inadequate.

Table 5.8.6 Summary of calibrated unidimensionality test across texture samples

	Paired t-tests		**Binomial Test**	**Result**
Texture	**% PST**	**%LB95CI**	**Lower 95% CI - Proportion**	
T1	28.20	24.60 %	0.246	Not acceptable
T2	22.80	19.10 %	0.191	Not acceptable
T3	27.20	23.50 %	0.235	Not acceptable
T4	27.70	24.10 %	0.241	Not acceptable
T5	20.40	16.80 %	0.168	Not acceptable
T6	21.98	18.40 %	0.184	Not acceptable
T7	30.90	27.30 %	0.273	Not acceptable

The unidimensionality test was analysed using two steps. In the first step all the items were analysed using paired t-tests, and in the second step all the analyses were validated using binomial tests. The result shows poor independent variables across all texture samples.

Details of the analysis are illustrated in Appendix A (Subtest 1 – page 275), Table 8.3.4, Table 8.3.7, Table 8.3.10, Table 8.3.13, Table 8.3.16 and Table 8.3.19 row unidimensionality for the texture sample one, two, three, four, five, six and seven respectively. Similar results were obtained for the rest of the samples.

5.8.3.2 Second t-test and binomial test

The unidimensional test was replicated in a similar way as the first t-test. The difference between the first and second tests is that on the second test the items were carefully examined through unidimensional assessment (Figure 5.8.1). Each item was examined according to the requirement and only items free from response dependencies, good fit residuals and free from disordered thresholds were select to the second t-test. Means second t-test were using pre-selected items whereas the first test was performed with all items.

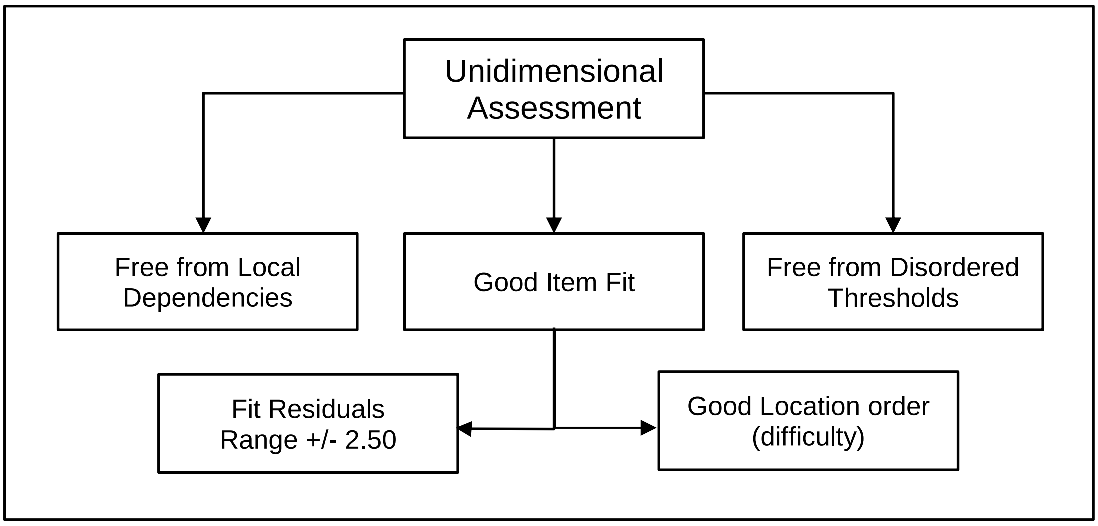


Figure 5.8.1 Unidimensional Assessment prior to Analysis in T-Test

5.8.4 Exploring t-test with multiple trials

The second t-test explored the possibilities of items that are unidimensionally fit using three approach attempts known as subtests. The objective behind these tests is to look at which approach gives better independent t-test results.

In the first subtest, two subset items were overlaid among positive loading; in order to do this, negative loading items will be removed from the dataset. The logical structure of this approach is that this study assumes positive items hold better loading just as in principle components analysis where the higher loading denotes better visibility as opposed to negative loading. By deleting negative items, this study speculates that the positive items have greater visibility to discriminate them among other items. The t-test result successfully obtained the lower ninety-five percent confident proportion is equal or less than five percent.

The second subtest was to replicate the same process, but instead of selecting positive items, the subset was selected from among the negative items which mean the positive items need to delete. The logical structure of this approach is that this study assumes negative items could be treated as the easiest to discriminate independently. The t-test result also successfully obtained the lower ninety-five percent confident proportion is equal or less than five percent (Table 5.8.7).

The second subtest of the dataset was overlaying positive and negative subset items. This result also produced a successful t-test result. Details results illustrates in Appendix A (subtest 2 - page 296) Table 8.3.22, Table 8.3.23, Table 8.3.24, Table 8.3.25, Table 8.3.26, Table 8.3.27 and Table 8.3.28 for the texture sample one, two, three, four, five, six and seven respectively.

Table 5.8.7 Second T-test Results

Texture	Sample *n*	Observed *n*	% PST	%LB95CI	Unidimensionality Accepted
1	145	12	8.27%	0.470%	Acceptable
2	138	10	7.24%	0.036%	Acceptable
3	145	12	8.27%	0.047%	Acceptable
4	145	10	6.89%	0.033%	Acceptable
5	145	11	7.58%	0.040%	Acceptable
6	145	12	8.27%	0.047%	Acceptable
7	145	10	6.89%	0.033%	Acceptable

5.8.5 Unidimensionality accepted items

Each frequency item that has a unidimensional fit across all texture samples were summarised as shown in Table 5.8.8 in the left column. As a result, item thirteen (*Touching this texture feels pleasant)* was considered as the most unidimensionally fit among all seven textures as illustrated in the hierarchy order of the right column (Table 5.8.8).

In this study it was assumed if any items fit with all textures, the items should be unidimensionally fit in all possible pairwise textures in the PC2 study. Therefore, the way in which the items were selected for the PC2 items pool is based on which appear most frequently in most of the textures.

Table 5.8.8 Items Fit Based on Second Unidimensional Accepted

T1	T2	T3	T4	T5	T6	T7		Seq	Item	Frequency	Decision
1	2	2	1	2	2	1		1	13	7	Propose
5	3	3	2	3	3	2		2	2	6	Propose
6	4	4	3	4	4	3		3	3	6	Propose
7	9	9	4	8	9	4		4	4	6	Propose
8	12	10	9	9	10	8		5	9	6	Propose
13	13	11	10	10	11	9		6	12	6	Propose
16	14	12	11	11	12	10		7	14	6	Propose
17	18	13	12	12	13	11		8	18	6	Propose
19	20	14	13	13	14	12		9	20	6	Propose
		18	14	14	15	13		10	10	5	Removed
		20	15	15	18	14		11	11	5	Removed
			18	18	20	15		12	15	4	Removed
			20	20		16		13	1	3	Removed
						18		14	8	3	Removed
						20		15	16	2	Removed
								16	5	1	Removed
								17	6	1	Removed
								18	7	1	Removed
								19	17	1	Removed
								20	19	1	Removed

Table 5.8.9 Nine unidimensionality fit items were selected for next studies

Item	Freq	Statements	Unidimensionally fit for Texture
13	7	Touching this texture feels pleasant.	1, 2, 3, 4, 5, 6 and 7
2	6	I have the impression that this texture would make my car feel spacious and neat.	1, 2, 3, 4, 5 and 6
3	6	I have the impression this texture is modern and contemporary-looking.	
4	6	I would expect to see this texture with a good touch and feel in a reasonably priced car.	
9	6	The look of this texture makes me want to touch it straight away.	
12	6	This texture does not look overly cheap and plasticky.	
14	6	Touching this texture is relaxing.	
18	6	When I touch this surface I get a sensation of luxury.	
20	6	With this texture, I would feel comfortable inside the car.	
10	5	This texture has a sporty look and feel.	1, 2, 3, 4 and 5
11	5	This texture looks nice quality.	
15	4	Touching this texture makes me feel warm.	1, 2, 3 and 4
1	3	Before I touch this texture, I can see that it would feel grippy.	1, 2 and 3
8	3	The feel of this texture would help me feel confident with my driving.	
16	2	Touching this texture would make me feel connected when operating the switches in the vehicle.	1 and 2
5	1	If I gripped a steering wheel which had this texture, it would feel very safe.	1
6	1	If I gripped a steering wheel which had this texture, it would not be too slippery.	
7	1	The feel of this texture on my steering wheel or switches would help me to keep my eyes on the road without distraction.	1
17	1	Vehicle controls with this texture would give good feedback when shifting, pulling, turning and rotating.	1
19	1	With this texture, I would be able to operate the controls without needing to look.	1

5.8.6 Proposing unidimensionality items for PC2 study

Nine items were chosen as they highlighted the proposal for the PC2 items pool which considers the goodness of fit of the unidimensionality features in most of the texture samples (Table 5.8.9). The items selected were items thirteen, two, three, four, nine, twelve, fourteen, eighteen and twenty. These items were chosen because they shared good unidimensionally features which were free from dependencies, item fit and disordered thresholds. To validate this item, an independent t-test was performed with these items. The results from the t-test and binomial test shows all nine items have good unidimensionality features which the lower ninety-five percent (%LB95CI) is equal or less than five percent.

Table 5.8.10 Final Items pool for PC2 study

Original Code in LS2	New Code in PW
Q2	Q1
Q3	Q2
Q4	Q3
Q9	Q4
Q12	Q5
Q13	Q6
Q14	Q7
Q18	Q8
Q20	Q9

All nine of these items were recorded with new coding for the PC2 study of the technical requirements of the RuMM2030 software; all the items must be in ordered or in sequence as shown in Table 5.8.10.

5.9 DISCUSSIONS

5.9.1 Test of local dependencies

The LS2 study demonstrates a higher number of local dependencies which are associated with problematic dependencies in which the value was greater than the cut-off value. In this study data from an affective response to the vehicle's interior texture exhibits a greater number of local dependencies for most of the texture samples.

Local dependencies refer to one of the tests to check whether the items in this study are free from dependencies, if in other words the items are independent, not related to each other and the participants are able to judge them independently.

One of the common signs of local dependency is where the items are constructed with a similar phrasing or meaning, which is prevalent in the reading comprehension test that can be a potential source of local item dependence. The effect for the reader can be associated with redundancy where the items were too predictable which may result in bias (Purya Baghaei, 2008).

5.9.2 Test of unidimensionality

The unidimensional test used independent t-tests in RUMM2030® and validated using binomial tests. The unidimensionality test in this study shows poor results as the independent t-test was not successful in reaching five percent of the lower ninety-five percent of confident interval proportion.

These poor results are caused by some of the items including higher fit residuals that result unfit in which the value has interfered with the aggregate in the t-test. The second t-test was performed with a pre-selected condition where the analysis looked only at items that passed the unidimensionality assessment. In this assessment, each item was examined so that the items would be free from local dependencies, fit for acceptance fit residuals and difficulties, and the items did not have any disordered thresholds. The logical sense of this test was to speculate those items are unfit in all the unidimensionality assessments inflate a greater proportion of t-test which means the items have a tendency to be multidimensional. Thus, in this assessment, the unfit items will automatically be removed from the dataset. Only the items that fit will be retained and proceed to the second unidimensionality test.

Local dependencies have a relatively perfect correlation to the unidimensionality test because the objective of obtaining the independent variable is to meet the assumption of unidimensionality in the RM.

The result from the first t-test and binomial test shows most of the items did not unidimensionally fit, thus requiring innovative strategies on how to examine each of the items independently as some of the items might be fit in certain texture samples but not in others. Thus, the notion of unidimensional assessment was created as a checkpoint to examine the characteristics of each item. Are these items holding dependency problems, fit residuals problems, disordered thresholds problems and dimensionality problem in PCA analysis? All these factors were carefully examined for each of the items independently. The results of the second t-test and binomial test analysis show some of the items are unidimensionally fit and were successfully nominated as items for PC2 study.

5.9.3 Test of items fit residual

RM assumed the items which include higher fit residuals greater than +/- 2.50 logits are potentially exposed to multidimensionality where the items were influenced by participants, who did not behave in the same way as other participants endorsed or responded. These problems were normally associate with item difficulties, ambiguity and multidimensional attributes, which resulted in the items unable to perform as expected by the model.

The value greater than 2.50 logits indicates this value is associated with some misfits where participants indicate a lack of the expected probabilistic relationship among the items within a scale which may probably be responses endorsed with careless or low motivation.

Higher negative fit residuals are normally associated with item-total correlation in classical test theory that normally indicates the redundancy or over-discriminating of items. The negative fit residuals indicate the participants could be responding according to fixed thinking.

This value greater than 2.50 logits indicated that unfit participants that do not behave in the same way as the other participants when responding to the test. It also indicates the items are less discriminating which may probably have been caused by the difficulties of the items. In common calibrations practice, RM suggests participants who hold higher fit residuals need to be removed as it will affect the overall aggregate lead to fit statistics.

5.10 SUMMARY

This chapter demonstrates a perfect correlation of the stimuli of mean logits between data observed and obtained by both LS1 and LS2 studies.

Although studies evidence of the LS2 shows that the large sample size offers greater power of statistical fit and delivers good stability in comparison to the previous LS1 study. However, this study has endeavoured with the large number items with problems that are associated with response dependencies resulted in poor t-test which signals lack unidimensionality.

Two series of analysis were performed in this study: namely, facet analysis and non-facet analysis. The objective of facet analysis is to examine the statistical stability of the person facet and item facet which are linked to all stimuli, while the non-facet were used to test the items fit in the context of response dependencies, which was done independently as a separate facet.

Two series of subtest using non-facet analysis was introduced to analyse the independent t-test. The first subtest, however, failed to perform any unidimensionality tests because some of the items were fitted, problematically fitting that associate with high fit residuals. However, the first subtest was used as items screening for second T-test via unidimensional assessment.

The result of the second subtest was successful in detecting and calibrating the multidimensional items were associated with response dependencies. The binomial test was used to validate the independent T-test, which the result is significantly matched. This analysis was successfully calibrated nine items which are unidimensionally fit for PC2 study.

Chapter 6

Validating the stability PC for interior vehicle texture

6.1 INTRODUCTION

Previous chapter findings have demonstrated the viability of using Rasch analysis to obtain measures of affective responses from paired comparisons, that participants find it easier to make paired comparisons when rating confectionery. However, the goodness of fit statistics of the data compared with the Rasch model is weak. In this study, the PC was used again to replicate similar procedures using different stimuli. The second PC namely Pair comparison study 2 (PC2) was used on seven pieces of interior vehicle textures to evaluate the PQ properties against nine items which was calibrated and unidimensionality fit. A new computer-based self-report system presented one-hundred and sixty-nine participants with a picture of pairs of texture labels and evaluated the statements in all pairwise combinations. The study demonstrates a weak correlation value at R^2 0.0033, indicating the PC2 data from affective responses using vehicle interior texture does vary within the same context in LS studies. PC2 was insufficient to hold the stability across different samples. Similar results displayed weak fit statistics due to a large number of unfit persons. The weak statistical results, however, are seen as the weakness behind the advantages of PC. The hypothesis to obtain the statistical stability in measuring affective responses using vehicle interior textures was not achievable; this can be seen as the weakness behind the advantages.

6.2 HYPOTHESIS OF THE STUDY

This study hypothesises is to determine whether the use of PC2 improves the target judgement where the binary scales allow the items to endorsed with greater discrimination.

This study formulates that the observed data from affective responses using PC2 does not vary within the same context when using SDS and LS, but it also fits as an expectation in the RM-structure.

The study formulates that PC2 is likely to have a better scaling structure to minimise bias and error when measuring participants' affective responses compared with SDS and LS.

6.3 OBJECTIVES

The aim of this thesis is to assess whether linear measurement of affective responses can be derived from PC2. A similar objective was achieved in PC1 using confectionery; however, in this chapter the stability of PC2 was examine using vehicle interior textures.

The objective of this study is as follows:

1. Administer an AE study of affective responses to the quality of vehicle interiors using PC2. This study will observe how well participants might find it easier and faster to evaluate products using PC2.
2. Assess how well PC2 satisfies the assumptions in minimising effect size of bias and error. The emphasis is to observe the targeting judgement.
3. Determine whether observed data from affective responses using vehicle interior texture do not vary within the same context using confectionery. If it does, then the linear correlation can be established to indicate the stability across the samples.
4. Evaluate whether PC2 shows a better response rate (speed).

6.4 METHODS

The method used to measure affective responses to vehicle interior texture was replicated from the study using confectionery. The PairWise© software was used to validate the mean logits of the RM. SPSS was used to validate the R-square correlation logits of individual textures.

6.5 ETHICAL APPROVAL, CONSENT AND RISK ASSESEMENT

Ethical approval was granted from The University of Leeds (Approval Reference No. MEEC 16-050), conforming to participants' health and safety procedures. The ethical approval, consent forms and risk assessment are included in Appendix B.

The protocol was briefly explained to the participants. Information about instructions and examination of the interior vehicle stimulus and a demonstration of clicking the choice using the computer-based survey was also briefly explained. The protocol and questionnaire were also included in Appendix C.

6.5.1 Data collections

In this study, one hundred and sixty-nine participants which 58.58 percent or ninety-nine males and 41.42 percent or seventy females, aged between nineteen and sixty-nine years of age (SD 5.968 and Mean 24.2) were recruited to take part in this study. All the participants consist of staff, researchers, postgraduates and undergraduate students at the University of Leeds. The study was conducted in the Affective Engineering Laboratory in the Mechanical Engineering Faculty, University of Leeds. Each participant received £5 as compensation for taking part in the study and each session was no longer than forty-five minutes.

6.6 APPARATUS AND EXPERIMENT SETTING

The stimulus was specially made using standard automotive plastic PP with a standard dimension of 225mm x 160mm x 3mm and was engraved using sandblasting into an organic, geometric pattern mostly seen in the interior passenger cars. Each stimulus was injected with a semi-gloss level of between 1.5GU to 2.2GU in various colours such as black, matte black, brown and beige with colours pigments labelled one to seven (Figure 3.6.1, page 54).

The stimuli went through several stages of selection before being used for data collection. To encourage the participants' understanding of the research context, all participants were required to pay attention to a few short videos about the driving experience. Three standing panels were used to evoke the study context, which illustrates the vehicle exterior and interior.

Seven pieces of stimuli used in this study were coded and configured according to rank-order methods, by which they were assigned into the best possible pairwise, so that every piece of stimuli will be paired fairly as shown in Table 6.6.2.

Table 6.6.1: Seven stimuli of interior vehicle

Texture Code	Label in Computer Screen
T1	Texture K
T2	Texture W
T3	Texture G
T4	Texture J
T5	Texture R
T6	Texture P
T7	Texture Q

Table 6.6.2: Preliminary pool of pairwise stimuli

Pair	Pairwise combinations
1	Texture K vs Texture W
2	Texture K vs Texture G
3	Texture K vs Texture J
4	Texture K vs Texture R
5	Texture K vs Texture P
6	Texture K vs Texture Q
7	Texture W vs Texture G
8	Texture W vs Texture J
9	Texture W vs Texture R
10	Texture W vs Texture P
11	Texture W vs Texture Q
12	Texture G vs Texture J
13	Texture G vs Texture R
14	Texture G vs Texture P
15	Texture G vs Texture Q
16	Texture J vs Texture R
17	Texture J vs Texture P
18	Texture J vs Texture Q
19	Texture R vs Texture P
20	Texture R vs Texture Q
21	Texture P vs Texture Q

6.6.1 Data administration

6.6.1.1 Building the items pool

Nine statements about measuring the dimensions of PQ attributes of vehicle interior textures was statistically validated as unidimensionality fit in Chapter 5 were used in this study. The results from the t-test and binomial test shows all nine items have good unidimensionality features with the lower ninety-five percent (%LB95CI) equal to or less than five percent.

Table 6.6.3: Nine unidimensionality fit items pool for PC2 study

LS2 Code	PC2 Code	Statement
2	1	I have the impression that this texture would make my car feel spacious and neat.
3	2	I have the impression this texture is modern and contemporary-looking
4	3	I would expect to see this texture with a good touch and feel in a reasonably priced car.
9	4	The look of this texture makes me want to touch it straight away.
12	5	This texture does not look overly cheap and plasticky.
13	6	Touching this texture feels pleasant.
14	7	Touching this texture is relaxing.
18	8	When I touch this surface I get a sensation of luxury.
20	9	With this texture, I would feel comfortable inside the car.

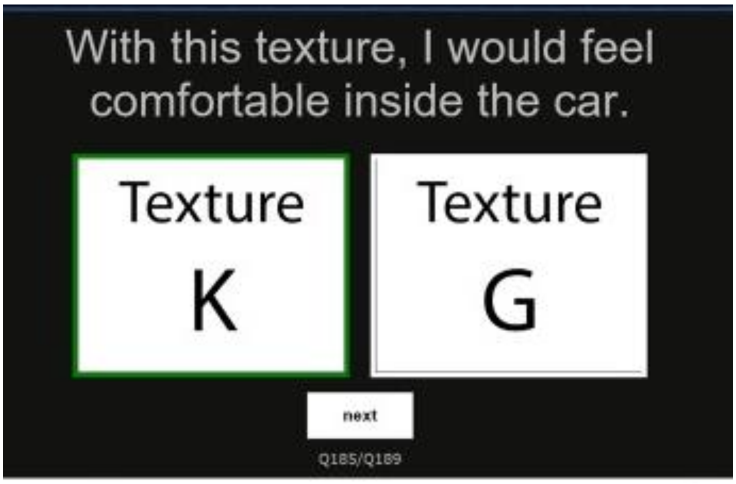


Figure 6.6.1: PC2 Self-report assessment computer interface

6.6.1.2 Computer-based survey

Data from participants' affective responses were collected using a bespoke, computer-based, self-report system using bespoke software from Microsoft Visual Basic 2016 (Figure 6.6.1).

The computer interface presented to the participant each label in a pair of vehicle interior stimuli in all Pairwise combinations. In total twenty-one Pairwise combinations need to be endorsed against the nine statements so each participant needed to endorse one hundred and eighty-nine items to complete the survey.

The participant was asked to indicate which of each pair satisfied the statement best. Instead of using a visual image of the interior vehicle texture, the computer program replaced them with alphabetical labels to avoid bias as participants might have a tendency to respond to the items without touch and feeling the physical stimuli. It was also to encourage the participants to compare the pairwise which was located next to the computer, with confident judgement. In a way the item appearing on the computer screen was randomised to minimise bias effect.

6.6.1.3 Data coding and analysis process

The data were administrated and stored in an online PC2 program database created by Justine Gallagher (Gallagher, 2016). All the data collections were exported to a Microsoft Excel spreadsheet from Microsoft Corporation for cleaning then imported to the Rasch Model software called RUMM2030® Professional Edition by RuMM Laboratory Pty Ltd (Andrich, Sheridan and Luo, 2012). The data were analysing and calibrated using Rasch-facet design approach. The data were analysed using a secondary software called PairWise© Software version 1.5.4198 Copyright Active One Software 2007-2009 (Humphry, 2010; Humphry et al., 2017) to obtain a mean location logit for each stimulus. Datasets were administered using Excel spreadsheets (Microsoft Corporation) and the unidimensionality test was performed and validated using statistical test methods of the t-test and binomial test that were performed using RUMM2030® software and Excel spreadsheets (Microsoft Corporation). R-square correlation was established using IBM SPSS statistics 21 and Matlab version R2013b.

6.7 RESULTS

In this study, two results were presented. The first results, consisting of summary statistics of all Pairwise textures, were treated as separate facets and the second results presenting summary statistics of individual textures were treated as separate facets.

6.7.1 Summary of statistics – All paired textures

Results of the preliminary analysis were presented with the Chi-Square probability value of $p \leq 0.01$ indicating a lack of the desired scale invariance across the traits (Table 6.7.8 column *p* – page 190).

Chi-Square calculated the item fit statistics based on person-item deviations and deviation by the ability level within the same group, known as a class interval. The equal class interval exclude extreme persons of (56 + 55 + 58 = 169), indicating good interval distribution.

The power of analysis fit indicated that there was a reasonable degree of fit of the items in the dataset. The initial test of fit summarises the overall model fit including the individual person fit and item fit, targeting, the exploration of thresholds and calibrations statistics.

6.7.2 Reliability – Internal consistency

The power of test-of-fit is a visual representation of the person separation index (PSI) which indicates how well the PC2 can distinguish or discriminate between the participants' latent trait locations.

The initial fit statistics in Table 6.7.8 – page 190 show the preliminary analysis of PSI was reported as 0.91 and the illustrated PC2 construct meets the acceptance level to statistically differentiate and equally distribute between three groups of participants across a number of a sample size of one hundred and fifty participants.

6.7.3 Item fit location and fit residuals

By default, the RUMM2030® software will automatically assign all the item location means centralised at zero (Table 6.7.8 – page 190 column item location). The item logarithm or log on the same continuum along the x-axis uses a standard unit of termed logits.

The overall fit residual statistics to the model were examined using the mean item log residual test. Item fit residuals refer to how easily the items are endorsed by participants. RM analysis estimates the degree of divergence, known as the residual, between the observed data from the participants and the expected data from the model. The data will consider it fits into the model when the values have a mean of zero and a SD of one. However, Rasch conventionally accepts residual log ranges of < +/- 2.50 logits.

The overall fit residual statistics were examined using the mean item log residual test (Table 6.7.8 – page 190). The mean value was recorded at -0.443 logits and the SD was 1.589 logits. These indicated that the items are associated with some misfit, which indicates the items were not functioning as intended where it was over-discriminated.

Table 6.7.1 indicates one item does not fit well into the model, which includes high positive fit residuals and it is therefore suggested it should be removed from the dataset. The positive residuals usually indicate items treated with under-discrimination, which lead to a misfit with the model expectations. However, the only item affected is item five with a log residual of 2.721 logits *(This texture does not look overly cheap and plasticky)* and was associated in pairwise sixteen *(Texture J vs Texture R).*

Table 6.7.1: Items with positive high fit residuals

Seq	Item	Pairwise	Location	SE	FitResid	DF	ChiSq	DF	Prob
1	5	16	-0.732	0.173	**2.721**	167.11	41.159	2	0

Table 6.7.2 indicates twenty-one items do not fit into the model, which contains a high negative fit residual. The negative residuals are usually associated with high item-total correlation in classical test theory, which usually associates with items with redundancy or over-discriminating of an item. RM suggests removing these items because they will violate the statistical results.

Table 6.7.2: Items with negative high fit residuals

Seq	Item	Pairwise	Location	SE	FitResid	DF	ChiSq	DF	Prob
1	3	6	0.533	0.161	**-4.705**	167.110	21.578	2	0.000022
2	7	6	0.488	0.162	**-4.513**	165.130	20.974	2	0.000028
3	5	6	0.515	0.161	**-4.449**	167.110	22.492	2	0.000014
20	9	3	0.565	0.162	**-2.557**	167.110	5.569	2	0.061766
21	5	11	0.591	0.163	**-2.551**	166.120	9.432	2	0.008953

6.7.4 Person fit location and fit residuals

Person fit location refers to participants' ability to endorse the items. In RM the participants with higher ability levels are likely to have more positive endorsements towards the difficulty items on a scale. The participant's logarithm or log on the same continuum along the x-axis uses a standard unit termed logits. In RM the fit statistics are using a Z-score of normal distribution. The overall person location fit statistics show the person location mean value at 0.199 logits and the SD value at 0.588 (Table 6.7.8 – page 190 column Person Location). This indicates the participants have higher ability levels to respond to difficult items. Despite having a sufficient fit location, the person fit includes some misfit, which indicates a lack of the expected probabilistic relationship among the items within a scale.

The overall person fit residual statistics were examined using a mean person log residual test (Table 6.7.8 – page 190). The mean value was recorded as -0.991 logits and the SD was 4.265. This value is considered unfit where the negative fit residuals indicate participants' responses are unexpected or contain too much dependence, while SD is overly spread which results in poor distribution.

Table 6.7.3: Participants holding high positive fit-residuals

No	PersonID	Tot/Exp Sc	MaxSc	Location	SE	FitResid	DegFree
1	126	110	189	0.365	0.156	**10.044**	186.9
2	43	107	189	0.292	0.156	**8.676**	186.9
3	10	87	189	-0.184	0.155	**8.624**	186.9
33	122	114	189	0.463	0.158	**2.884**	186.9
34	8	121	189	0.639	0.160	**2.777**	186.9

Further investigation was carried out to determine the source of the misfit. The individual person analysis indicated that 20.11 percent or thirty-four participants out of one hundred and sixty-nine participants were affected by higher positive fit-residuals, while 42.60 percent or seventy-two participants out of one hundred and sixty-nine participants were affected by higher negative fit-residuals (Table 6.7.3 and Table 6.7.4). However, no participants was identified as an outlier nor any participants as holding extreme values.

Table 6.7.4: Participants holding high negative fit-residuals

No	PersonID	Tot/Exp Sc	MaxSc	Location	SE	FitResid	DegFree
1	140	85	189	-0.232	0.155	**-8.050**	186.9
2	64	103	189	0.196	0.155	**-7.968**	186.9
3	71	107	189	0.292	0.156	**-7.762**	186.9
71	5	100	189	0.124	0.155	**-2.552**	186.9
72	83	106	187	0.289	0.157	**-2.505**	184.9

6.7.5 Fit statistics for facet analysis – All paired textures

Table 6.7.5 shows the preliminary analysis of all pairs of texture were treated as a separate facet before alteration by location order. The table demonstrates the preliminary analysis of two facet designs before alteration by location order. Facet one showed item location discriminated by the level of difficulty. Facet two showed stimuli locations discriminated by the ability of participants that corresponded to the stimuli. The facet level illustrates items and Pairwise combinations with hierarchical difficulties to endorse.

Facet analysis was used to demonstrates and comparing the item and stimuli or objects performance, ratting or grade. The structure in facet analysis indicates the hierarchy order based on difficulties logit in unidimensionality scale structure. Positive location indicate item or stimuli in more difficult to endorsed while negative indicate less difficult endorsement. On difference connotation or taxonomy, the positive or negative relatively refers to the degree of preferences; positive location endorsement often associates with higher preferences of particular stimuli while negative indicate with lower preferences.

Table 6.7.5: Preliminary facet analysis of all paired textures

Facet One					Facet Two				
Seq	Item	Locn	SE	FitRes	Level	Pairwise	Locn	SE	Fit Res
1	2	-0.091	0.17	-0.294	1	Pairwise 14	-0.921	0.180	-0.604
2	4	-0.034	0.16	-0.057	2	Pairwise 13	-0.897	0.180	1.288
3	8	-0.032	0.17	-0.245	3	Pairwise 4	-0.848	0.180	-0.133
4	5	0.005	0.17	-0.037	4	Pairwise 17	-0.791	0.180	-0.189
5	1	0.014	0.17	-0.273	5	Pairwise 5	-0.780	0.180	-1.531
6	9	0.020	0.17	-0.505	6	Pairwise 16	-0.740	0.170	1.820
7	6	0.036	0.17	-0.221	7	Pairwise 1	-0.422	0.170	1.363
8	3	0.040	0.17	-0.325	8	Pairwise 15	-0.319	0.160	-0.548
9	7	0.042	0.17	-0.201	9	Pairwise 18	-0.239	0.160	-0.514
					10	Pairwise 9	-0.237	0.160	0.351
					11	Pairwise 10	-0.209	0.160	-1.503
					12	Pairwise 12	-0.020	0.160	1.489
					13	Pairwise 19	0.114	0.160	-1.350
					14	Pairwise 6	0.424	0.160	-3.989
					15	Pairwise 3	0.544	0.160	-2.063
					16	Pairwise 2	0.546	0.160	0.098
					17	Pairwise 11	0.646	0.160	-2.884
					18	Pairwise 20	0.907	0.170	-1.259
					19	Pairwise 8	0.995	0.170	-0.963
					20	Pairwise 21	1.057	0.170	1.081
					21	Pairwise 7	1.191	0.180	-0.270

6.7.6 Fit statistics for facet analysis – Individual textures

Table 6.7.6 shows the preliminary analysis of individual textures that were treated as separate facets before alteration by location order. Facet one showed item location discriminated by the level of difficulty. Facet two showed stimuli locations discriminated by the ability of participants that corresponded to the stimuli. The first facet indicated sixty-three items that correspond to the seven stimuli as shown on the second facet.

Table 6.7.6 Preliminary facet analysis of individual texture

Facet Analysis of RUMM 1									
Facet One					Facet Two				
Seq	**Item**	**Locn**	**SE**	**FitRes**	**Level**	**Pairwise**	**Locn**	**SE**	**Fit Res**
1	13	-0.593	0.160	-0.325	1	Texture 3	-0.429	0.16	-0.129
2	41	-0.592	0.160	-0.310	2	Texture 1	-0.314	0.16	0.078
3	62	-0.589	0.160	-0.231	3	Texture 4	-0.294	0.16	-0.184
4	34	-0.585	0.160	-0.234	4	Texture 7	-0.024	0.16	-0.026
5	6	-0.575	0.160	-0.236	5	Texture 2	0.093	0.16	0.033
6	27	-0.573	0.160	-0.231	6	Texture 5	0.386	0.16	-0.290
7	20	-0.558	0.160	-0.247	7	Texture 6	0.583	0.16	-0.389
8	55	-0.552	0.160	-0.219					
9	48	-0.551	0.160	-0.233					
10	33	-0.385	0.160	-0.290					
11	40	-0.385	0.160	-0.310					
12	26	-0.381	0.160	-0.302					
13	61	-0.370	0.160	-0.289					
14	54	-0.369	0.160	-0.299					
15	5	-0.368	0.160	-0.305					
16	47	-0.351	0.160	-0.291					
17	12	-0.347	0.160	-0.266					
18	19	-0.344	0.160	-0.291					
19	23	-0.133	0.160	-0.040					
20	30	-0.119	0.160	-0.047					
21	44	-0.117	0.160	-0.044					
22	58	-0.105	0.160	-0.027					
23	51	-0.084	0.160	-0.039					
24	2	-0.083	0.160	-0.043					
25	9	-0.081	0.160	-0.016					
26	16	-0.081	0.160	-0.020					
27	37	-0.072	0.160	-0.052					
28	63	0.002	0.160	-0.082					
29	49	0.003	0.160	-0.082					
30	14	0.012	0.160	-0.053					
31	35	0.015	0.160	-0.071					
32	42	0.018	0.160	-0.089					
33	28	0.019	0.160	-0.054					
34	7	0.030	0.160	-0.094					
35	56	0.040	0.160	-0.032					
36	21	0.052	0.160	-0.066					
37	32	0.262	0.160	-0.181					
38	39	0.265	0.160	-0.179					
39	50	0.272	0.160	-0.005					
40	57	0.275	0.160	-0.022					
41	11	0.277	0.160	-0.186					

42	60	0.281	0.160	-0.154
43	53	0.288	0.160	-0.175
44	8	0.293	0.160	-0.035
45	29	0.301	0.160	-0.020
46	18	0.302	0.160	-0.189
47	25	0.305	0.160	-0.166
48	46	0.307	0.160	-0.183
49	1	0.312	0.160	-0.023
50	36	0.315	0.160	-0.035
51	15	0.322	0.160	-0.008
52	43	0.334	0.160	-0.043
53	22	0.343	0.160	-0.006
54	17	0.380	0.160	-0.144
55	59	0.402	0.160	-0.166
56	38	0.404	0.160	-0.154
57	52	0.408	0.160	-0.174
58	3	0.419	0.160	-0.178
59	4	0.419	0.160	-0.164
60	10	0.421	0.160	-0.169
61	24	0.453	0.160	-0.169
62	45	0.455	0.160	-0.174
63	31	0.474	0.160	-0.153

6.7.7 Expected probability targeting

The summary of fit statistics for the preliminary analysis are presented in a graphical representation to indicate how well the person and items are distributed. The person item threshold distribution in Figure 6.7.1 indicates the analyses using all Pairwise were treated as separate facets, that participants found it very easy to endorse the items, and that the matching of the difficulty of the items plotted the normal distribution kurtosis.

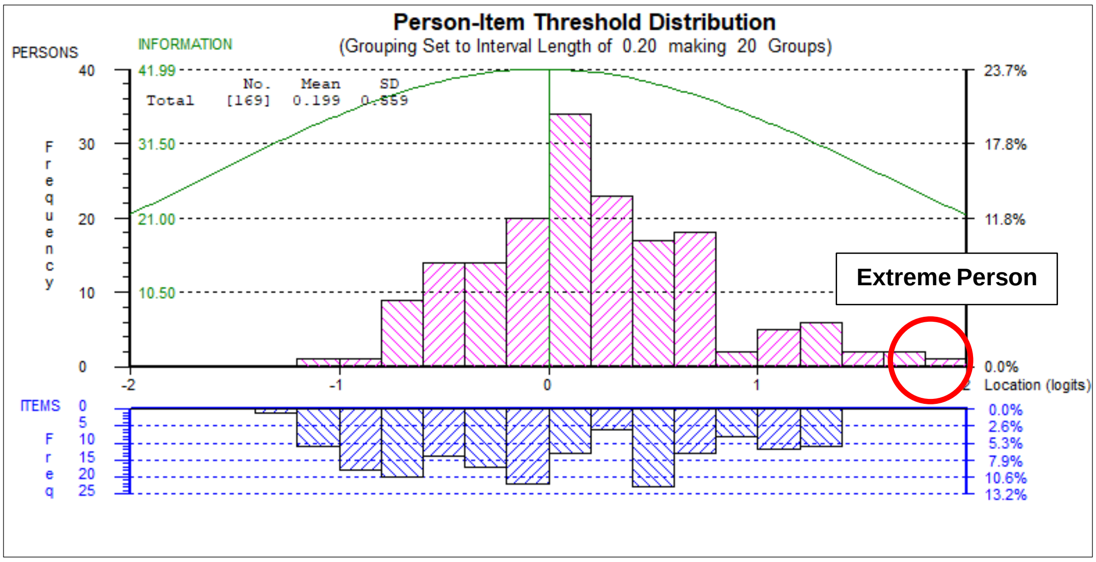


Figure 6.7.1: Person-item distribution for analysis of current data in which all pairs of textures were treated as separate facets

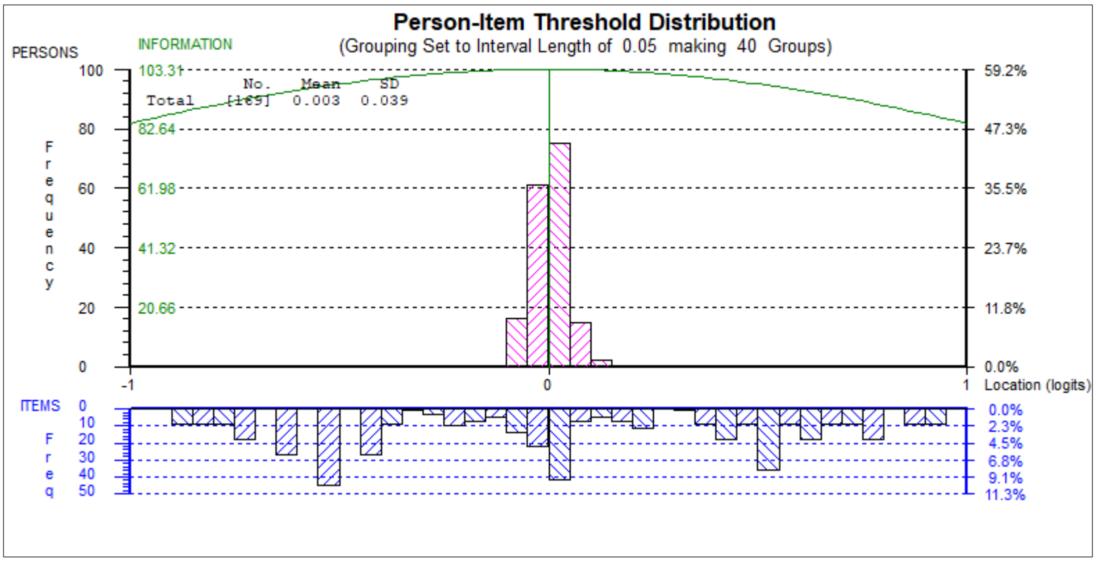


Figure 6.7.2: Person-item distribution for analysis of current data in which individual textures were treated as separate facets

The person-item distribution in Figure 6.7.2 indicates the analysis of individual textures treated as separate facets shows that approximately half of the items were very easy to endorse and the other half were very difficult to endorse.

However, because of the way the data are coded for analysis, each comparison is represented by two data points, and consequently, each easy item has a mirror-image difficult item (Table 6.7.7). In theory, therefore, the item distribution should be symmetrical. The coding of the same-confectionery pairs by random ones and zeros might account for the small amount of asymmetry in the distribution. The individual texture plot of the kurtosis distribution in leptokurtic curves or very narrow spreads of the persons' willingness to endorse shows that there was not much variation in participants' affective assessments of the textures.

Table 6.7.7: Pairwise matrix of individual textures with mirror images of difficult items

	T1	T2	T3	T4	T5	T6	T7
T1		0	1	1	0	0	1
T2	1		1	1	0	0	1
T3	0	0		0	0	0	0
T4	0	0	1		0	0	0
T5	1	1	1	1		1	1
T6	1	1	1	1	0		1
T7	0	0	1	1	0	0	

6.7.8 Disordered thresholds

Disordered thresholds occur when participants have difficulties consistently endorsing or discriminating between response options categories on a scale. The investigation of disordered thresholds in this study (Figure 6.7.3) illustrates there are no disordered thresholds found in the PC2 analysis. No overlapping thresholds or under-discriminating was identified resulting in all the category scales being ordered. The graphical representation shows ordered thresholds indicate the participants found it easy to use PC2 to discriminate between two possible items.

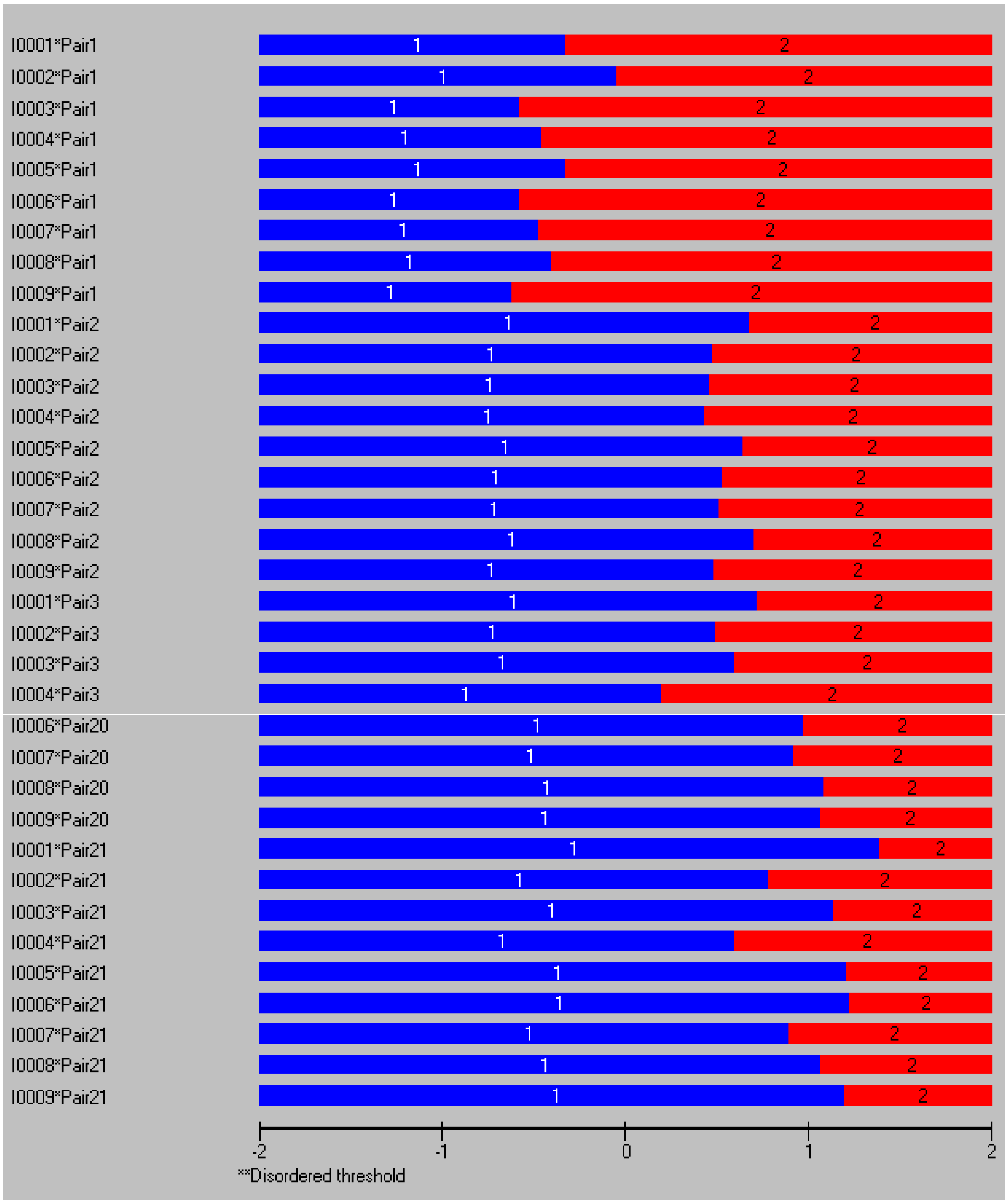


Figure 6.7.3: PC2 Thresholds map

6.7.9 Data collection speed

The average duration of each study was approximately twenty-six minutes and eleven seconds, the time taken to complete the one hundred and eighty-nine pairs of items for each participant. The individual response rate for each comparison item took an average 8.31 seconds per item to respond.

6.7.10 Calibrated statistics

The calibrations analysis of preliminary dataset has been calibrated to observe the summary statistics of the person-item fit residual, reliability and targeting performance (Table 6.7.8 – page 190).

6.7.11 Rasch model analysis results

This study has demonstrated the overall results of how participants feel about the quality of interior vehicle trims, especially in relation to tactile texture. The analysis uses the RM RUMM2030® analysis; the result has demonstrated vehicle interior texture six at 0.557 logits was endorsed as the most perceived high-quality interior texture based on the criteria or items given. Meanwhile, the less PQ value, texture three at -0.427 logits, was endorsed as the less perceived value (Table 6.7.11 – page 192). These results were most similar to the results analysed using PairWise© analysis (Table 6.7.12 – page 193).

6.7.12 PairWise© analysis results

PairWise© analysis (Humphry, 2010; Humphry et al., 2017) was used to measure algorithms from the dataset to obtain a mean location logit for each stimulus. The program was written to conform with standard statistical measurement in item response theory (IRT), which is similar to RM. Using the PairWise© analysis, the result has demonstrated vehicle interior texture six at 0.600 logits was endorsed as the most perceived high-quality interior texture based on the criteria or items given, while the less PQ value, texture three at -0.673 logits, was endorsed as the less perceived value (Table 6.7.12 – page 193). The reason PairWise© was applied in this study was to determine the mean logits which are absent in RM. Instead of a given location estimate based on each stimulus, RM offers a pairwise format which is difficult to compare and validate with the location estimates in the LS study to determine linearity.

Table 6.7.8 : Fit statistics for all PairWise© treated as separate facets

Analysis	Item Loc		Person Loc		Item Fit Res		Person Fit Res		Chi-Square Interaction			Psi Index	Alpha
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Value	dF	*p*	Psi Index	Alpha
Initial Data (n=169)	0.000	0.720	0.199	0.558	-0.443	1.569	-0.991	4.265	1312.4	378.0	<0.001	0.91	N/A
Calibrated Data 1 (n=155) Ref : n14 IQAT Missing	0.000	0.711	0.177	0.536	-0.390	1.526	-0.981	4.290	1174.6	378.0	<0.001	0.91	0.91
Calibrated Data 2 (n=123) Ref : DelHiFitResPrs1	0.000	1.194	0.191	0.652	-0.135	1.283	-0.776	3.192	957.7	378	<0.001	0.93	0.93
Calibrated Data 3 (n=96) Ref : DelHiFitResPrsn2	0.000	1.590	0.196	0.727	-0.042	1.087	-0.572	2.436	691.4	378	<0.001	0.94	0.94
Calibrated Data 4 (n=73) Ref : DelHiFitResPrsn3	0.000	1.603	0.231	0.743	0.044	1.001	-0.216	1.620	630.0	378	<0.001	0.94	0.94
Calibrated Data 5 (n=62) Ref : DelHiFitResPrsn3	0.000	1.708	0.182	0.709	0.057	0.938	-0.157	1.188	521.8	378	<0.001	0.93	0.93

Table 6.7.9: Calibrated facet design (All Pairwise)

Seq	Pairwise	Stimuli	Logit	SE	Fit-Res
1	15	Texture G vs Texture Q	-0.574	0.290	-0.130
2	14	Texture G vs Texture P	-0.411	0.280	-0.927
3	4	Texture K vs Texture R	-0.371	0.280	-0.384
4	1	Texture K vs Texture W	-0.306	0.280	1.179
5	5	Texture K vs Texture P	-0.281	0.280	-0.176
6	13	Texture G vs Texture R	-0.268	0.280	1.885
7	12	Texture G vs Texture J	-0.199	0.270	1.184
8	18	Texture J vs Texture Q	-0.157	0.270	1.310
9	6	Texture K vs Texture Q	-0.153	0.270	-1.939
10	3	Texture K vs Texture J	-0.044	0.270	-0.670
11	17	Texture J vs Texture P	0.004	0.270	-0.292
12	16	Texture J vs Texture R	0.027	0.270	2.959
13	19	Texture R vs Texture P	0.111	0.270	-1.273
14	9	Texture W vs Texture R	0.122	0.270	0.039
15	10	Texture W vs Texture P	0.157	0.270	-0.960
16	20	Texture R vs Texture Q	0.204	0.270	0.107
17	11	Texture W vs Texture Q	0.216	0.270	-1.478
18	8	Texture W vs Texture J	0.251	0.270	-0.060
19	2	Texture K vs Texture G	0.288	0.270	0.043
20	21	Texture P vs Texture Q	0.452	0.270	2.793
21	7	Texture W vs Texture G	0.930	0.280	-0.103

Table 6.7.10 : Fit statistics for the preliminary analysis of individual textures treated as separate facets

Analysis	Item Loc		Person Loc		Item Fit Res		Person Fit Res		Chi-Square Interaction				
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Value	dF	*p*	Psi Index	Alpha
RuMM 1 Initial Data	0.000	0.522	0.003	0.039	0.304	1.035	-0.820	6.840	1014	885	0.0012	-5.461	N/A
RuMM 2 Initial Data	0.000	0.523	0.001	0.039	0.299	1.042	-0.813	6.812	1080	882	0.0001	-5.300	N/A
RuMM 3 Initial Data	0.000	0.522	0.002	0.039	0.300	1.021	-0.829	6.855	752.1	882	0.9994	-5.312	N/A
RuMM 4 Initial Data	0.000	0.522	-0.004	0.039	0.311	1.025	-0.818	6.832	802.5	882	0.9734	-5.418	N/A
RuMM 5 Initial Data	0.000	0.521	-0.001	0.041	0.284	0.977	-0.819	6.837	994.7	882	0.0700	-4.726	N/A
Average Data*	0.000	0.522	0.000	0.039	0.300	1.020	-0.820	6.835	928.6	882	0.409	-5.243	N/A

Table 6.7.11: Average individual logit (RUMM 1 – RUMM 5)

Textures	Average		RuMM 1		RuMM 2		RuMM 3		RuMM 4		RuMM 5	
	Location (logit)	Standard error	Location (logit)	Standard error	Location (logit)	Standard error	Location (logit)	Standard error	Location (logit)	Standard error	Location (logit)	Standard error
Texture 3 / G	**-0.427**	0.16	-0.429	0.16	-0.421	0.16	-0.441	0.16	-0.419	0.16	-0.425	0.16
Texture 1 / K	**-0.308**	0.16	-0.314	0.16	-0.304	0.16	-0.307	0.16	-0.306	0.16	-0.312	0.16
Texture 4 / J	**-0.29**	0.16	-0.294	0.16	-0.295	0.16	-0.288	0.16	-0.293	0.16	-0.284	0.16
Texture 7 / Q	**-0.024**	0.16	-0.024	0.16	-0.034	0.16	-0.034	0.16	-0.014	0.16	-0.017	0.16
Texture 2 / W	**0.092**	0.16	0.093	0.16	0.083	0.16	0.1	0.16	0.086	0.16	0.99	0.16
Texture 5 / R	**0.381**	0.16	0.386	0.16	0.388	0.16	0.38	0.16	0.375	0.16	0.378	0.16
Texture 6 / P	**0.557**	0.16	0.583	0.16	0.583	0.16	0.59	0.16	0.571	0.16	0.56	0.16

Table 6.7.12: PairWise© results – Calibrated analysis (n123)

Textures	Preferred	Involved	Estimated	SE	Outfit	Chi-Sqr	DF	Class Interval
Texture 3 / G	2125	6642	**-0.673**	0.027	1.021	17.110	4744.2	1
Texture 4 / J	2510	6642	**-0.449**	0.026	1.010	23.360	4744.2	1
Texture 7 / Q	2937	6642	**-0.211**	0.025	0.992	36.839	4744.2	1
Texture 1 / K	2966	6642	**-0.195**	0.025	0.976	36.746	4744.2	2
Texture 2 / W	3928	6642	**0.333**	0.026	0.992	9.168	4744.2	2
Texture 5 / R	4385	6642	**0.594**	0.026	1.013	12.812	4744.2	3
Texture 6 / P	4396	6642	**0.600**	0.027	1.011	16.375	4744.2	3

Table 6.7.13: Summary statistics

Property	Value
Outer Loop Count	4
Mean Location	0.000
Variance	0.223
Mean Square Error	0.001
PSI	0.997
Sum Chi-Square X^2	76.205
DF / Element	1.000

6.7.13 Linear correlation between scales

The linear correlation in this study demonstrates the weak relationship between PC2 and LS1. The correlation R-square value exhibits R^2 0.0057, indicating a weak relationship between individual textures treated as separate facets in the PC2 study and individual textures in the LS1 study (Figure 6.7.4). The result indicates the PC2 data from affective responses using vehicle interior texture do vary within the same context.

This indicates that PC2 was insufficient to hold the stability across different samples indicates that some of the properties were an identified misfit. The investigation has determined the poor correlation likely occurred due to sampling bias, stimuli bias and scaling bias. LS1 and PC2 studies demonstrates both measurement scaling are unique, have their own strength and weaknesses. PC more likely easier to be endorsed than LS, however if the product too easy to endorsed RM will reject unfit data because greater contrast promote higher degree endorsement and resulting extreme score. Then the rejection will interfere the overall location logit. The poor correlation in this study illustrates that they are some factor associate with the degree of bias in both study.

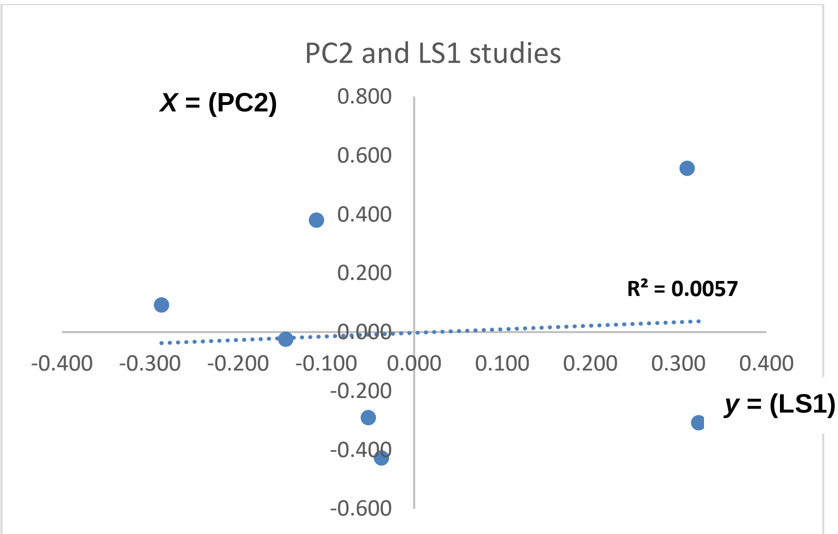


Figure 6.7.4: Weak relationship between PC2 and LS1

The similar results demonstrate a weak relationship between PC2 and within the same contexts using same stimuli in the second LS2 study at R^2 0.0033 (Figure 6.7.5).

The similar results demonstrate a weak relationship of R^2 0.0043, which indicates a weak relationship between PC2 using PairWise© software and the LS1 study (Figure 6.7.6). As a comparison, the individual stimulus logits were illustrated in Table 6.7.14.

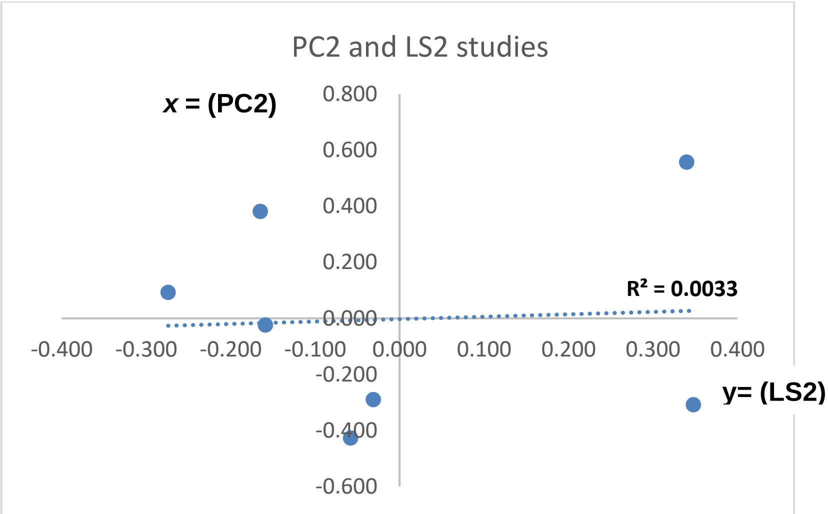


Figure 6.7.5: Weak relationship between PC2 and LS2

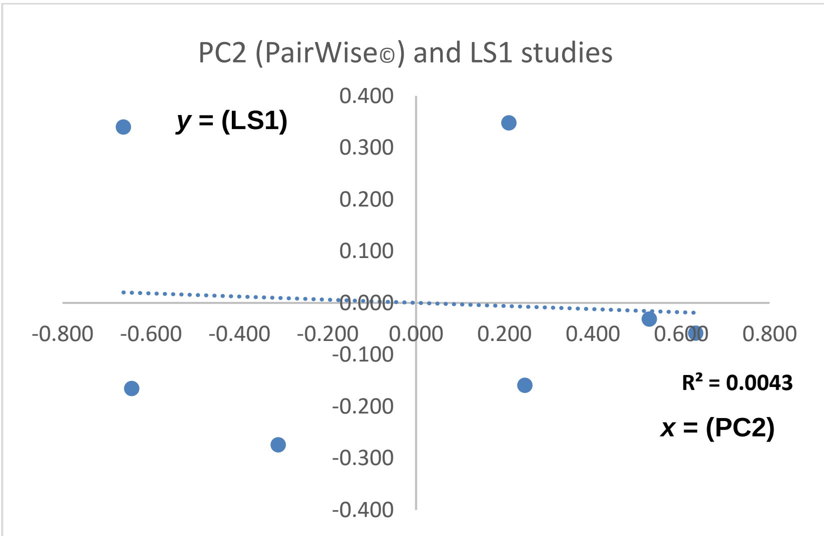


Figure 6.7.6: Weak relationship between PC2 (PairWise©) and LS1

Table 6.7.14: Individual textures logits across methods

	PC2 RUMM2030®	PC2 PairWise©	Likert (LS2)	Likert (LS1)
Texture 1 / K	-0.308	-0.195	0.348	0.323
Texture 2 / W	0.092	0.333	-0.274	-0.287
Texture 3 / G	-0.427	-0.673	-0.058	-0.037
Texture 4 / J	-0.290	-0.449	-0.031	-0.052
Texture 5 / R	0.381	0.594	-0.165	-0.111
Texture 6 / P	0.557	0.600	0.34	0.310
Texture 7 / Q	-0.024	-0.211	-0.159	-0.146

6.7.14 Pearson correlation

To verify the linear correlation, the mean location logits for individual textures, which is a standard dataset, were correlated using Pearson correlation in IBM SPSS statistics version 21. The result demonstrates the Pearson correlation value at 0.057 (Table 6.7.15) which indicates the weak relationship between LS1 and PC2 studies. The linear R-square shows correlation value at R^2 0.003 (Figure 6.7.7).

Table 6.7.15: Pearson correlation between PC2 and LS1 studies

Correlations

		LS1	PC2
LS1	Pearson Correlation	1	.057
	Sig. (2-tailed)		.903
	N	7	7
PC2	Pearson Correlation	.057	1
	Sig. (2-tailed)	.903	
	N	7	7

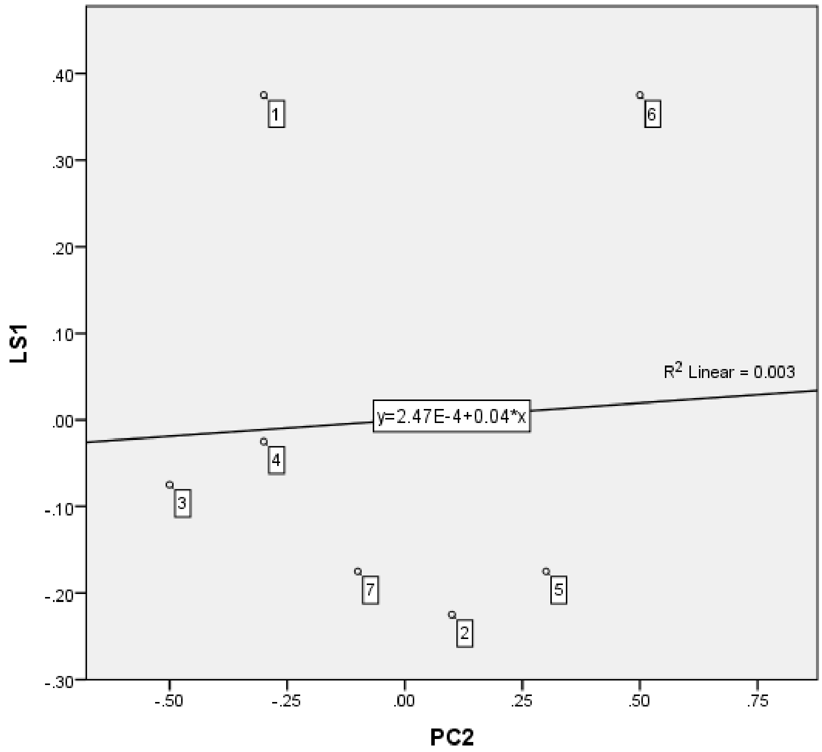


Figure 6.7.7: Linear R-square correlation in SPSS

6.7.15 Correlation validation using Matlab

Similar tests were used to verify the correlation value in Matlab. The R-square value was also exhibiting similar results at 0.003301 (Figure 6.7.8).

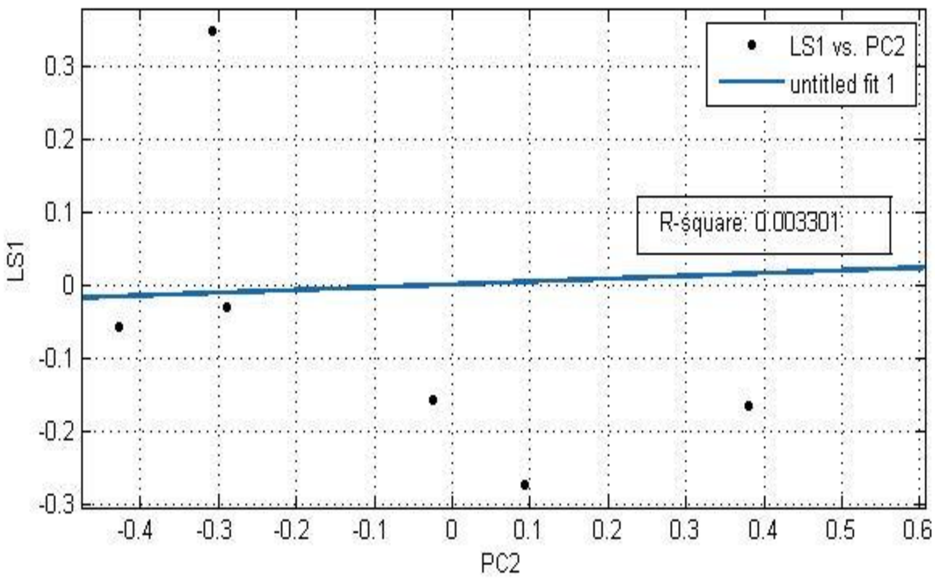


Figure 6.7.8: Linear R-square correlation in Matlab

6.8 DISCUSSIONS

6.8.1 PC2 demonstrates greater discrimination and stability against LS and SDS

The outcome from PC2 study has validated the research objective and provided the evidence that measuring affective responses to vehicle interiors textures using PC is a robust process. This study has also validated the claim that participants might find it easier and faster to evaluate products using PC compared with LS and SDS.

Two pieces of evidence associated with statistics viewpoints were used to justify the logical structure underlying the analysis. First, the targeting performance of person-item distribution, and secondly the person ability factors.

6.8.1.1 Targeting performance of person-items distribution

The graphical representation of targeting in the LS1 and LS2 studies are illustrated in the person-item distribution graphs (Figure 6.8.1 and Figure 6.8.2), respectively showing that the majority of the participants are well targeted with the set of calibrated items. However, the item difficulties are skewed more than person ability towards the x-axis direction on the scales, indicating that persons have difficulties in responding to LS items. The red box indicates that most of the difficult items were unable to be endorsed by participants.

While the graphical representation of targeting in the PC2 study (Figure 6.8.3) demonstrates the majority of the participants have better ability than item difficulties, the person ability is skewed towards the x-axis direction on the scales, indicating that participants are able to endorse PC2 more easily than LS. The red box indicates most of the difficult items were able to be endorsed by participants.

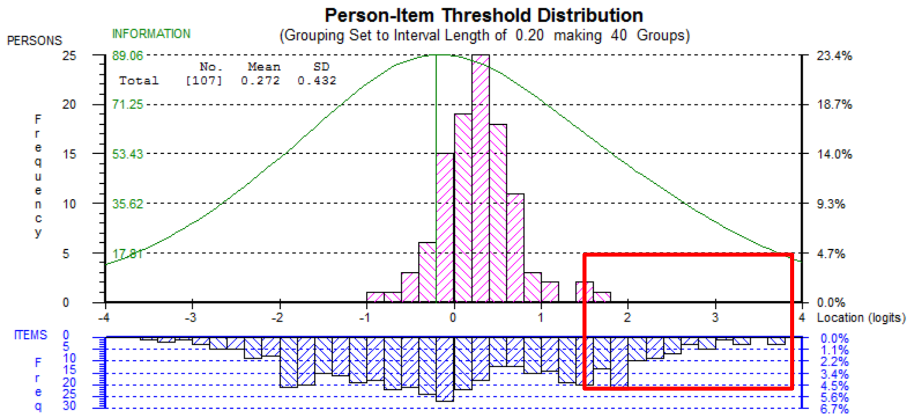


Figure 6.8.1: Graphical representation of targeting in LS1 study

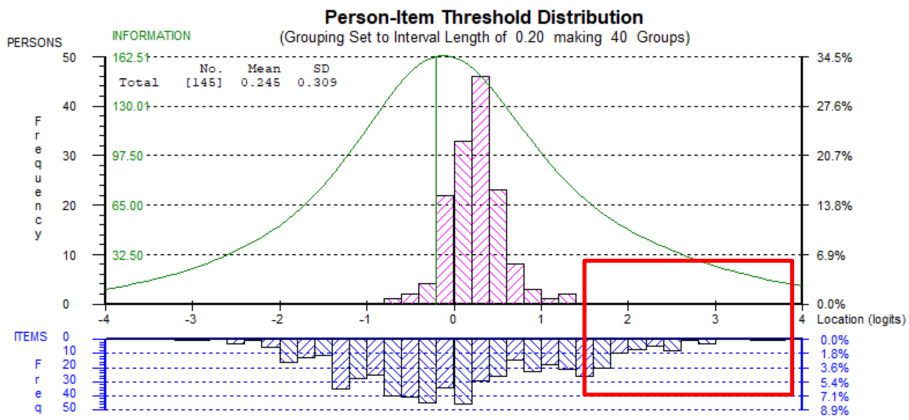


Figure 6.8.2: Graphical representation of targeting in LS2 study

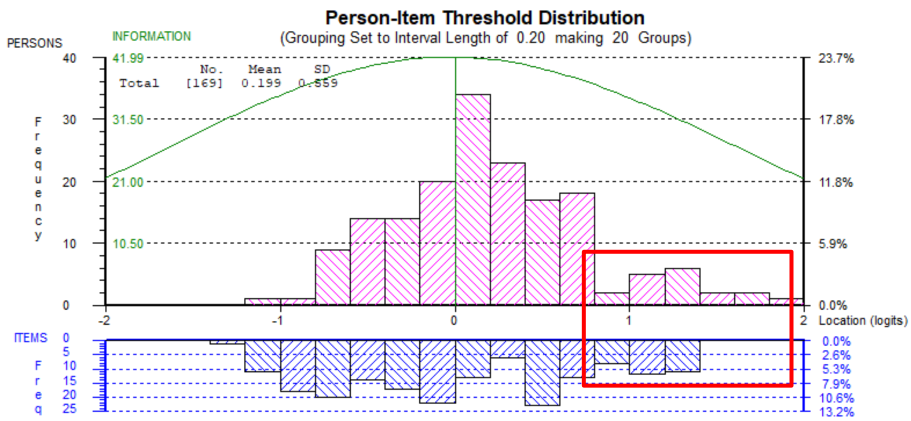


Figure 6.8.3: Graphical representation of targeting in PC2 study

6.8.1.2 Person Ability

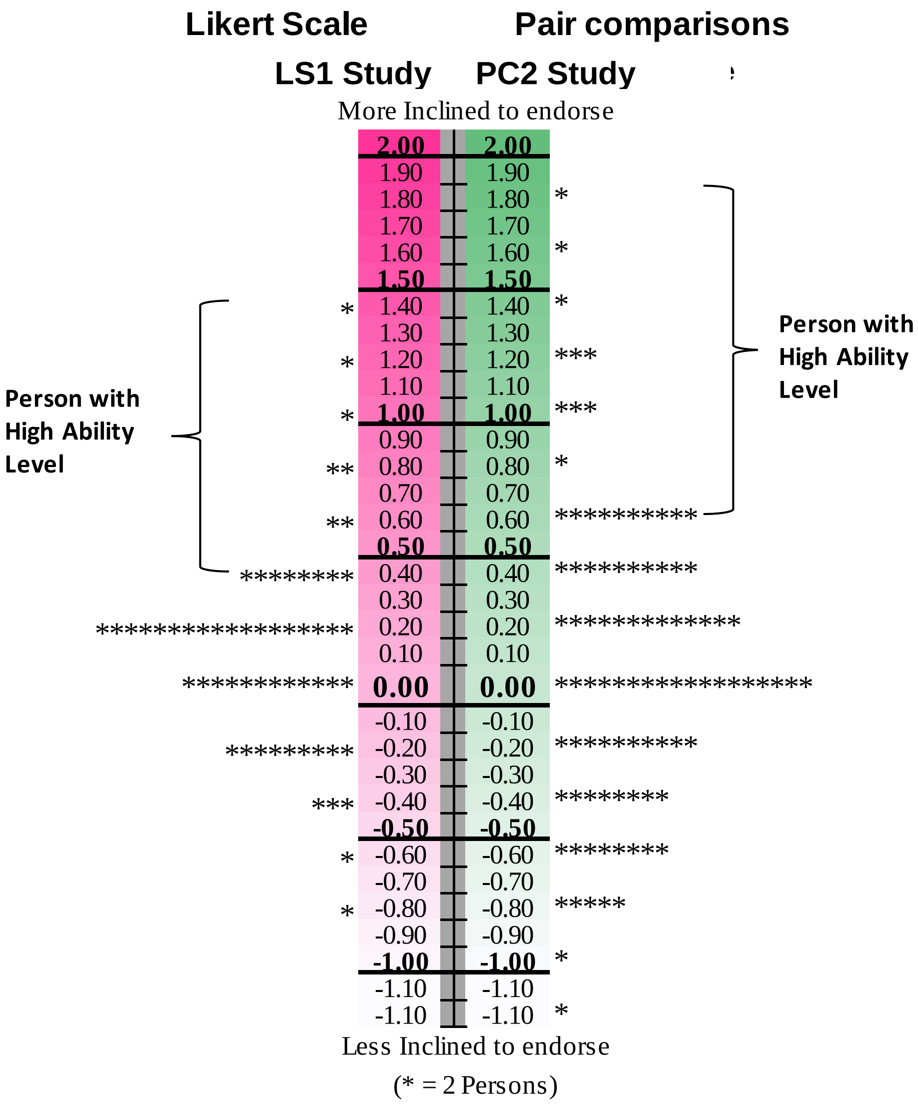


Figure 6.8.4: Comparison between person location logits LS1 and PC2

The graphical representation in Figure 6.8.4 shows the comparison of person location logits in endorsing the interior vehicle texture between individual textures on LS1 study on the left graph and all the Pairwise textures in PC2 on the right graph. The willingness to endorse in PC2 is greater than in LS as numbers of persons with higher ability levels are greater compared with LS. The location level of high ability persons is located at 1.80 logits compared with 1.40 logits in LS, indicating that measuring affective responses using PC2 is relatively effortless.

6.8.2 Delimitation – Bias in PC2

The analysis has demonstrated the poor statistical results and the correlation in this study demonstrates the weak relationship between PC2 and LS1 studies, which indicates PC2 was insufficient in holding the stability across the different samples.

Further investigation demonstrates finding the evidence why the PC2 did not demonstrate similar performance as in the PC1. The evidence is needed to understand the cause and effect in relation to the poor results. The investigation initiates looking at any anomalies in the data analysis and finding there is some evidence associated with poor statistical analysis.

The weak statistical presence of the Chi-square of item-trait interaction of *p*-value <0.01 before and after calibrations in PC2 analysis indicates that there is no significant deviation between the observed data and what was expected from the model. On other hand, this result explained there is some misfit associated with the generalisable population of the study and the observed data were not well distributed and conformed to expectations.

6.8.3 PC2 excessively discriminates the person-items fit

The investigation has identified that PC2 excessively discriminates the person-item location and fit residuals to the model resulting in misfit. The analysis has identified the person-fit residual exhibit value of -0.44 and SD 4.26 indicates a higher degree of misfit associated to the participant misfit.

Test of fit has identified a huge percentage of individual person-fit fosters the higher misfit value of fit residuals where approximately 42.60 percent or seventy-two participants out of one hundred and sixty-nine participants hold higher negative fit residuals, indicating the responses in the PC2 study are excessively over-discriminated by unfit participants.

This study has speculated that the reason for the greater fit residual was associated with targeting bias. The bias evidence demonstrates that the analysis has identified the item-fit residual exhibits at -0.443 SD 1.599, indicating the items were over-discriminated. This result has reinforced the assumption that PC is too easy to discriminate, and this has caused many participants to become over-qualified or extreme.

6.8.4 Investigating the items and person over- and under-discrimination

In PC2 study, the analysis has found the evidence associated with over-discriminating responses to the items, which is fostered by PC. RM assumes participants were considered as misfits compared with the model if they hold negative fit residuals greater than +- 2.50 logits, indicating that the items have been violated with over-discriminating responses (Pallant and Tennant, 2007; Camargo, 2013).

As evidence, over-discriminations were examined and it was found the observing dots plot a steeper curve than the expected curve line (Figure 6.8.5). Meanwhile in the opposite direction, under-discrimination shows the dots are plotting in the opposite direction from the expected curve (Figure 6.8.6). These dots represent the participants within the class intervals.

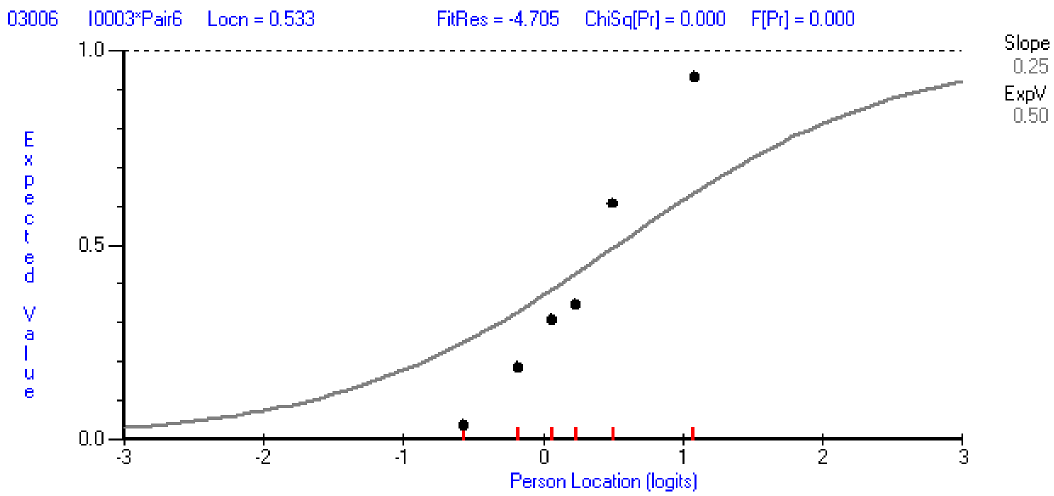


Figure 6.8.5. Over-discriminating items with high negative value

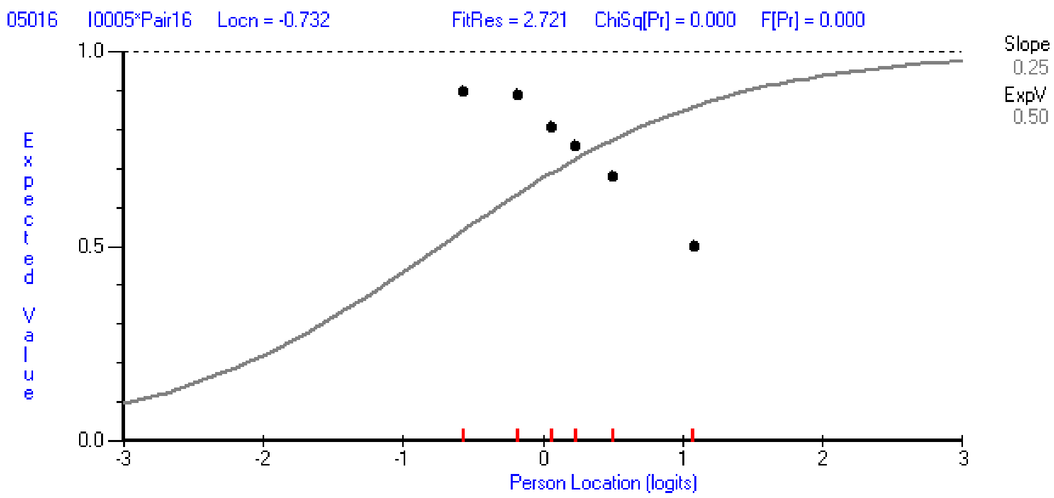


Figure 6.8.6: Under-discriminating items with high positive value

RM theory states that unfit persons are often associated with over-qualified or extreme persons, who are not considered as a valid sample; therefore, these samples will automatically be discarded through calibrations scales or procedures (Tennant et al., 2007; F.R. Camargo and Henson, 2015).

6.8.5 Source of bias

The investigation has identified sources of bias stemming from the misfit data and corrupting the measurement structure in the PC2 study. Bias was stimulating the misfit values through the following factors.

- **Non-random sampling**

Data collection in this study was performed using non-random sampling, whereby most of the participants were recruited within members of staff, undergraduates, postgraduate students and post-doctorate researchers within the University of Leeds. This could be one of the reasons the item-trait interaction is not generalised, thus producing the poor Chi-square value.

- **Poor target sampling**

Most of the participants are not familiar with the interior vehicle textures, nor do they understand what PQs are all about and how to assess them correctly. None of the participants are expert in vehicle design neither are they employed by automobile makers or working within the transport industries.

- **Stimuli Bias**

Unlike the confectionery study, in which most of participants are familiar with the products in addition to the confectionery having a strong emotional engagement in their daily lives, when it comes to the research context of using vehicle interior textures there is only poor emotional engagement for the most of participants. Additionally, the stimuli were not fast-moving consumer products, nor the finished vehicle components such as the steering wheel or door trim, but only pieces of plastic specimens, which may have resulted in vagueness for the participants. In evaluating the PQ of tactile surfaces, most of participants struggled to use familiar units to estimate the magnitude of the products. Most of them did not know how to quantify the quality of the surface. This means the evaluation might potentially carry some degree of bias. Jansari et al. (2000) reveal a lateral bias that the perceiver or participants show a bias to the stimuli, the so-called “perceived bias” (Jansari et al., 2000).

- **Small contrast ratio between stimuli**

Although the advantages of PC2, promote participants to discriminate between the difficult items and stimuli in greater contrast, some of the participants struggled to discriminate between the test stimuli when some of them had only a small contrast ratio and the image or colours may have had a distorting effect on the affective judgement.

- **Missing dataset**

Missing data were identified during analysis procedures. This may affect the degree of accuracy and impact the statistical outcome. The missing data were identified during the data process which involved 0.18 percent or forty-six data points from 31,941 data points. The missing data points were inflates 8.28 percent or fourteen participants out of one hundred and sixty-nine participants. However, the reason the data were missing was unknown, although the investigation was carried out. The missing data were coded with number 9 as shown on Table 6.8.1.

Table 6.8.1: Missing datasets were recorded

IQAT Software Missing Data value

No	Seq	Person ID	Pairwise	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	IQAT Coding
1	1	345	2	0	1	9	0	0	1	1	1	1	picture_Q3_2
2			15	1	1	0	9	1	1	0	0	1	picture_Q4_15
3	2	346	12	0	0	1	9	0	0	1	0	0	picture_Q4_12
44	14	431	3	9	1	1	0	1	1	1	1	1	picture_Q1_3
45			14	0	1	0	0	0	0	9	0	0	picture_Q7_14
46			21	1	1	0	0	0	1	9	1	0	picture_Q7_21

6.9 SUMMARY

This work demonstrates the viability of using Rasch analysis to obtain measures of affective response from PC2 and that participants find it easier to make PC compared with evaluating products separately against LS1 statements. However, PC2 is unable to hold sufficient stability across different samples where PC2 is not able to construct linear measurements of affective responses from paired comparisons using vehicle interior textures.

However, the result presents consistent performance in the context of the ability to discriminate persons and items, the reliability of testing, calibrations simplicity, and speed and quality of targeting, demonstrating a similar outcome as in the confectionery study. Participants find it too easy to discriminate the products along the affective dimension of interest. Despite having some good advantages, PC2 contains poor fit statistics, which was contributed to by unfit persons and has violated the measurement structure through over-discriminating responses to the items.

This study concludes that the poor statistical results may seem to be the weakness behind the advantages of PC, as PC2 has been found to excessively discriminate between the items and stimuli through its effortless style of PC scaling.

The study has come to the new hypothetical that, if the products are too different, participants find it too easy to discriminate. The greater discrimination contrast to the stimuli meant a large number of unfit persons, who were considered over-qualified or extreme, were rejected as these group exceed the acceptance cut-off value in measurement structure govern in RM.

The over-discrimination is technically considered one of source of bias, as in RM theory, misfit persons are normally associated with over-discriminating behaviour which means greater negative fit residuals are consider biased. This claim is supported by many RM scholars: that biased errors result from the person’s scale-checking style and incorrect responses (Paek and Wilson, 2011; Kubinger et al., 2012; Camargo, 2013; Horton, 2017).

Chapter 7

Discussions

Here the discussion findings are presented as relevant to the existing literature, along with the comparative studies of SDS1, LS1, LS2, PC1 and PC2.

7.1 VALIDITY IN AFFECTIVE ENGINEERING

Validity is an important factor in any test instrument, as was demonstrated in the value of the tests or surveys that were conducted.

In latent space, one of the major challenges in survey design is to ensure the attributes are free from multidimensional factors, because if they appear, the analysis can likely become a problem when trying to fit it to the RM, as the properties from multidimensional variables can yield shortcomings due to wide dimensional complexity and interconnectedness that inflate multiple meaning (Camargo, 2013).

In this study, the RM have observed most of the errors in data collection. The errors are associated with biases which the respondent suffer to connected the items inflate multiple meanings which are difficult to portray into single dimension or connotation on a LS. The respondent would rather not respond and leave it blank, or may answer using their own interpretation, which may introduce some bias response. This is a reason why RM rejects the items if they include multidimensional traits or meanings.

In common approach, items being developed using affective words will normally be associated with multiple regression processes, such as adjective reduction, FA and EPA dimension of rating responses as evaluation, potency and activity to position the items to elicit multiple dimensions in relation to certain classification attributes. However, this process may not be able to fully ensure that the items are unidimensional. Unidimensionality often violates the measurement structure of unidimensionality in Rasch analysis and corrupts the statistical outcome (Camargo, 2013). RM works if the items are unidimensionally fit, then achieving linearity or normal distribution in statistics is possible.

This study focuses on the items fit development using PC rather than focus on person fit, because an item with unidimensional fit will naturally improve the person location and minimise the fit residual logits for greater better accuracies. However, to generate unidimensionality items is not an easy task and complicated. Additionally, the items will behave in the way in which they were designed. The items might be troublesome if they carry multidimensional characters: for example, grammatical ambiguity, complicated vocabularies such as jargon terminology, and items without true linguistic contras. All these problems will affect the way participants interpret, which may tend to bias and error. Thus, it is important items should be designed to convey one dimensional attributes, known as unidimensionality.

7.1.1 Test of local dependencies

The LS2 study demonstrates a higher number of local dependencies which are associated with problematic dependencies in which the value was greater than the cut-off value. In this study, data from an affective response to vehicle interior texture exhibit a greater number of local dependencies for most of the texture samples.

Local dependencies refer to one of the tests to examine whether the items are associated with dependency issues: in other words, if the items are dependent, most likely participants are unable to discriminate independently. One of the common signs of local dependency is where the items are constructed with similar passages or meaning, which is prevalent in the reading comprehension can be a potential source of local item dependence. The effect for the reader can be associated with redundancy, where the items were too predictable and results may be biased (Purya Baghaei, 2008).

In AE studies, the ambiguity derived from affective words sometimes includes multiple interpretations leading to multidimensionality, which difficult for participants to respond to correctly. In developing items using the AE method, the unidimensional features offer a plus of protecting the dimension from multidimensional attributes. As an example, in this study the word "*grippy*" was constructed for both the SDS1, LS1 and LS2 studies. The words "*grippy"* or "*roughness"* contain a vagueness, whereby some participants can treat them as positive if the context is related to assessing haptic control. *"Grippy"* can be a positive cues to express the sensation, the feeling of confidence and safety while driving. While the on other hand, "*grippy"* might contain negative cues if the surface roughness is unable to stimulate sleek or smooth properties.

Another example in the study was the use of the word *"plasticky",* which was also classified as bipolar meaning and was reported among the difficult items to endorse which included a higher fit residual. The adjective was unable to quantify *"cheap"* cues because participants were not sure how to measure *"plasticky"* when the stimuli are plastic. The adjective *"plasticky"* in some contexts is dependent on whether some participants are able to discriminate with other attributes such as texture, gloss and colour.

The unidimensionality test done in Chapter 5 demonstrated poor results as the independent t-test was not successful in reaching five per cent of the lower 95 percent of confident interval proportions. The results are caused by some of the items including higher fit residuals. Underlying the unidimensional test is the assumption that those items are unfit in all the unidimensionality assessments, which will inflate a greater proportion of t-tests, which means the items have a tendency to be multidimensional.

7.1.2 Person-item fit residual

Person and item fit is a key important feature affecting the overall statistical results. These two components are often associated with misfit in data observations and violation of the scaling measurement. RM assumed the person and items including higher fit residuals greater than +/- 2.50 logits are potentially exposed to multidimensionality, where the items were influenced by participants not behaving in the same way as other participants endorsed or responded. This problematic factor was normally associated with item difficulties, ambiguity and multidimensional attributes, which resulted in items unable to perform as expected by the model.

A value greater than 2.50 logits indicates this value is associated with some misfit where participants indicate a lack of the expected probabilistic relationship among the items within a scale which will probably include the responses endorsed with careless or low motivation. Higher negative fit residuals are normally associated with item-total correlation in classical test theory, which normally indicates the redundancy or over-discrimination of items.

The negative fit residuals indicate the participants could be responding in a fixed-thinking fashion. It also indicates the items were less discriminating, which was probably cause by the difficulties of the items. In common calibrations practice, RM suggests participants with higher fit residuals need to be removed as they will affect the overall aggregate lead to fit statistics.

In this study, the proposed theory of PC mainly offers an alternative to non-comparative scaling to improve the precision measure of person ability and reduce the difficulty levels of items. Participants were always seen as responding correctly when the items were positioned in a low degree of difficulty. The level of information reflects the difficulty of the items being ask. If the information requires in-depth information it will definitely require a certain degree of item difficulty.

Since the items may be difficult to control, the available options to consider include choosing the simple scaling technique. In this case, PC can be a good alternative for minimising bias effect from item difficulties and person abilities.

7.1.3 The advantages of calibrations procedures in RM

In modern psychometric assessment, RM is a robust method of transforming an unfit dataset into good-fit statistics, where the effect bias can be manageably minimised and satisfy the statistical expectations. RM allows the original datasets to be diagnosed legally and provide appropriate treatment if the preliminary analysis identifies misfit. The calibrations are important for eliminating unnecessary work, time and additional cost if the data need be replicated a second time due to common error or an unsatisfactory statistical outcome. Calibrations is also an important procedure to be undertaken, although some data may need to be removed from the model. The calibrations process was not meant to discard important information, but RM suggests either fixing the misfit or removing the unfit dataset while maintaining the originality and will increase the reliability of the fit statistics.

It is interesting that RM is among the reliable methods of assessing item difficulties and person ability which can be derived independently on a same scale, which is robust (Bond and Fox, 2015; Horton, 2017). Many researchers use RM to examine how well the scales are developed, to assess how it will impact on the psychometric properties and for scales refinement (Horton, 2017).

However, the limitation of RM governs the greater cut-off point if the dataset goes wrong and shows misfit. Calibrations may suffer from greater discard as RM can be strict for latent datasets which are commonly associated with idiosyncratic responses, especially data in consumer research which often lacks targeting. Common rejection rate, approximately ten percent of datasets have to be removed in order to fit the model (Camargo, 2013).

Since RM works with unidimensionality features, there is a high chance unfit items will derive from dependency problems which hold that the values greater than the cut-off point need to be removed. Stimuli bias needs to be removed if it is difficult to discriminate as it may violate the scaling structure. When the scales appear under-discriminating or over-discriminating, this indicates some degree of misfit affected by stimuli and item bias.

This study uses facet design which is a popular method of assessing two or more factors with multiple levels. Usually the first factor was registered as a set of items and the second factor as different graders for each item or stimulus. Items and stimuli were bonding together as independent items which cannot be simply removed if the item was misfit. For example, item 01020 was coded as item 1 for stimulus 20 and was reported unfit during preliminary analysis and it was suggested it should be removed during calibrations procedures; however, these items cannot be simply removed as it would affect the adjacent item associated with the same stimulus: for example, 02020, 03020, and so on. By removing item 01020, all the items associated with the same stimulus were discarded. In this case, the misfit item needs to examine separately using a non-facet design known as single factor with multiple levels.

For some reason, RM works quite well in clinical, educational and psychometrics tests because of the homogeneity of participants within same target group; however, measuring consumer attitudes to products which involve a wide spectrum of participants with different backgrounds can be a challenge. Most likely, item-trait interaction could not fit into the model.

Calibration objectives in RM, seen as quality control tools for diagnosis and treatment of the bias problems associated with the items, scaling, stimuli and participant attitudes in order to achieve validity and reliability.

7.2 COMPARATIVE ANALYSIS

This chapter explains the specific analysis related to some of the important statistical issues and calibrations outcomes as a general overview as well the comparison between the SDS1, LS1, LS2, PC1 and PC2 studies (Table 7.2.1).

Table 7.2.1: Comparisons of preliminary statistics

	SDS1	LS1	LS2	PC1	PC2
No of Samples	75	107	145	157	169
Item-trait interaction	p <0.01	p <0.01	p <0.01	p <0.01	p <0.01
PSI	0.84	0.94	0.92	0.83	0.91
Cronbach α	0.83	0.93	0.92	0.86	0.91
Person fit Location	Mean 0.052 SD 0.121	Mean 0.226 SD 0.334	Mean 0.245 SD 0.308	Mean 1.90 SD 0.948	Mean 0.199 SD 0.558
Person fit residuals	Mean -0.01 SD 3.93	Mean -0.30 SD 3.60	Mean -0.40 SD 3.70	Mean -0.26 SD 1.01	Mean -0.99 SD 4.26
Item fit location	Mean 0.00 SD 0.268	Mean 0.00 SD 0.320	Mean 0.00 SD 0.299	Mean 0.00 SD 1.010	Mean 0.00 SD 0.720
Item fit residuals	Mean 0.64 SD 0.49	Mean 0.30 SD 0.50	Mean 0.44 SD 0.53	Mean -0.24 SD 1.24	Mean -0.44 SD 1.56
No of persons unfit residual	35	44	67	10	106
Total of items	140 items (7 stimuli x 20 items)	140 items (7 stimuli x 20 items)	140 items (7 stimuli x 20 items)	72 items (6 Pairwise x 12 items)	189 items (21 Pairwise x 9 items)
No of Scale	7	5	5	2	2
No of unfit residual items	0	0	0	3	22
Ordered Thresholds	8	28	29	0	0
Disordered Thresholds	132	118	111	0	0
No of item remain Disordered	77	2	2	0	0
Calibration Rescored complexity	Yes	No	No	No	No

7.2.1 Person fit residuals unfit to the model

The comparative analysis briefly explained the comparable strengths and weaknesses of each method in measuring affective responses using confectionery and vehicle interior textures.

In total, there were 578 participants within five series of study. In general overview, non-comparative scaling in SDS1 and LS1 and LS2 offer great methods in measuring participant attitudes, and the advantages of SDS and LS include that scaling was well recognised by most of participants in this survey. However, the outcome demonstrated in this study was that SDS1 and LS1 and LS2 suffered from category scales problems which was highlighted in red text illustrates the misfit values resulting some poor statistical outcome.

Table 7.2.1 (Row no of person unfit residual) demonstrates the complexity of the scaling in SDS1, LS1 and LS2 has affected the greater number of disordered thresholds, which indicate participants struggling to consistently discriminate between response option categories on a scale. It affected the score system, response pattern and person-item interaction (Camargo, 2013) and jeopardised the measurement structure if the response malfunctioned (Salzberger, 2015). The impact from the misfit is a major source and impacts on poor statistical result (Tennant et al., 2007; Salzberger, 2015).

7.2.2 Disordered thresholds

The graphical representation of the ICC graphs presents the disordered threshold in one of the misfit items in the SDS1 study as shown in Figure 7.2.1. Some of the category scales were treated as under-discriminating against the adjacent category scale while some of the items were treated as over-discriminating. This is to indicate the items and scales that did not cooperate in homogeneity to aid participants in discriminating clearly. While Figure 7.2.2 demonstrated the category scales in ordered.

The threshold map in Figure 7.2.3 is another graphical representation to indicate items before adjustment with disordered thresholds shows in double asterisk. While Figure 7.2.4 represents ordered thresholds after adjustment or calibrations where all the category scales are work together as expected in the model.

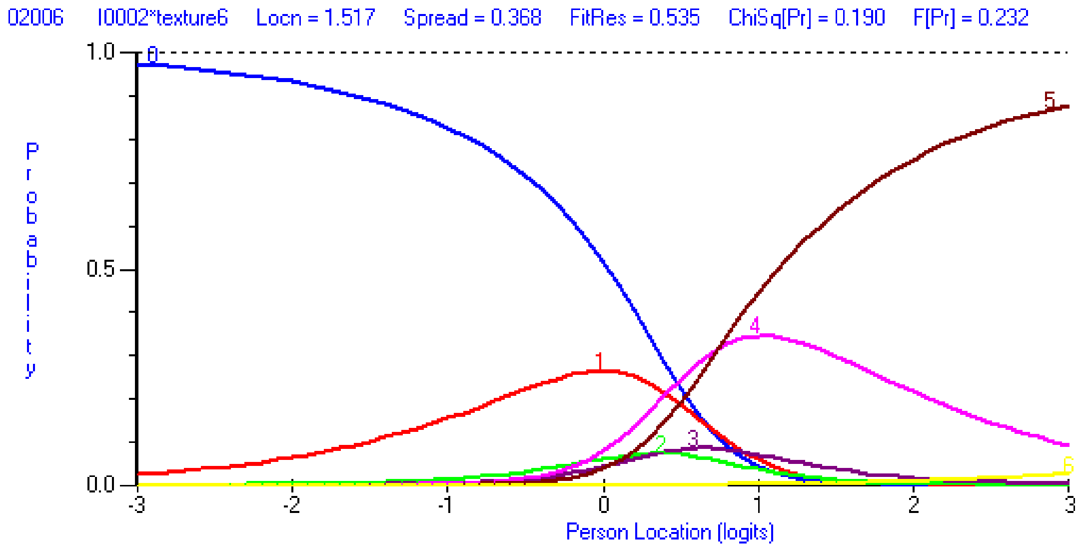


Figure 7.2.1: Category scales with disordered thresholds

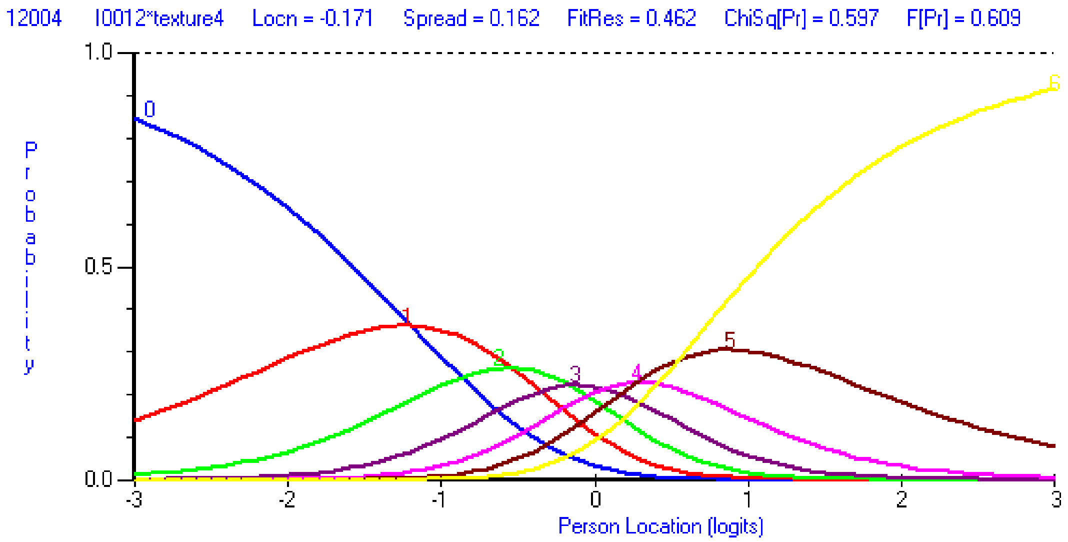


Figure 7.2.2: Category scales with ordered thresholds

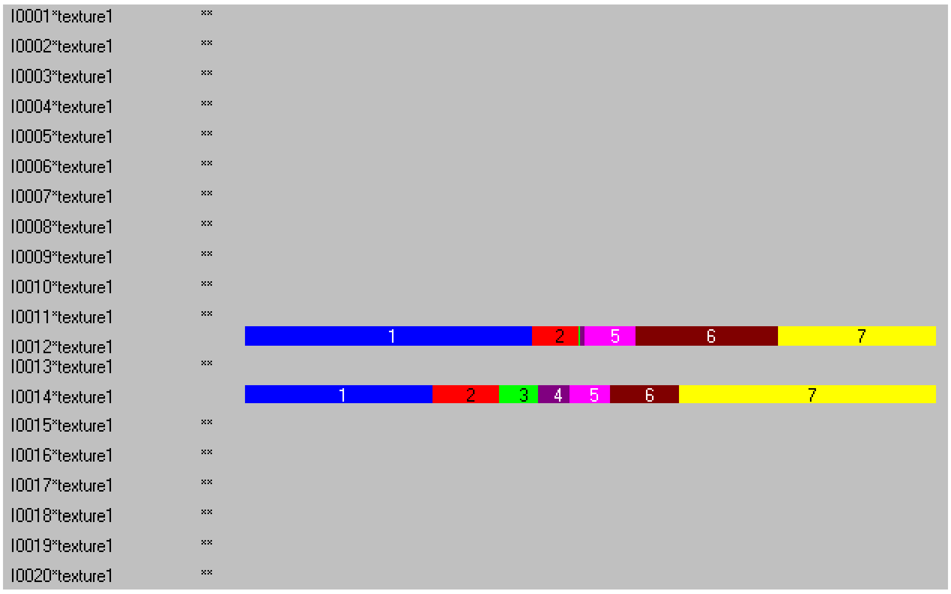


Figure 7.2.3: Disordered thresholds map before adjustment

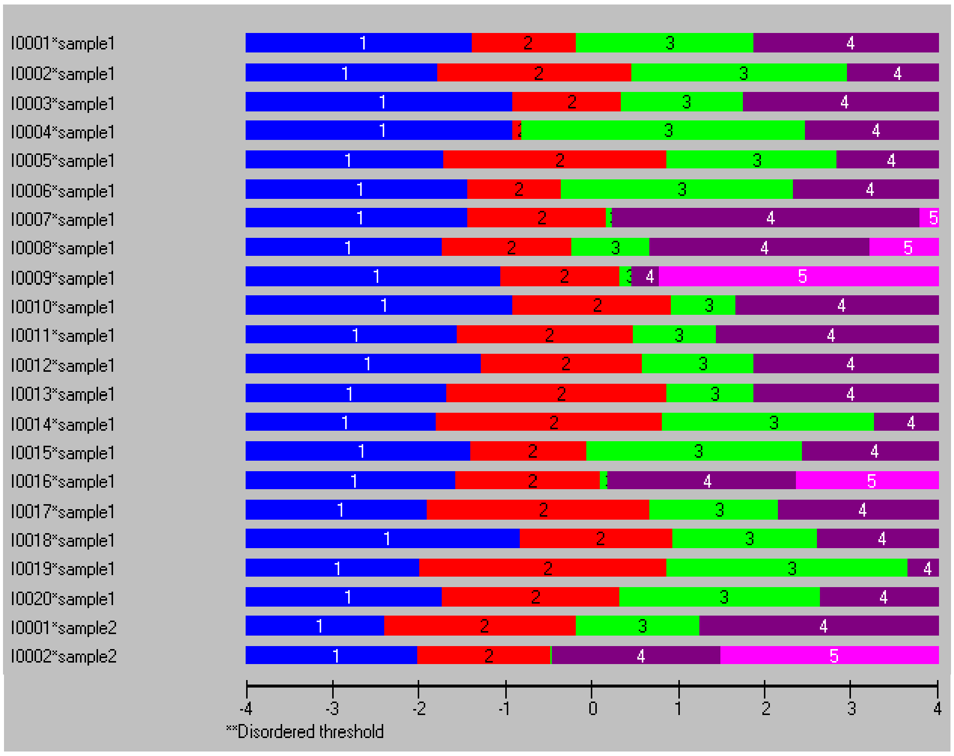


Figure 7.2.4: Ordered thresholds map after adjustment

7.2.3 Targeting performance

Targeting is a common word that indicates how well the items cooperate with the scales structure to extract the stimuli properties correctly. Poor targeting describes when the item and person ability were not on the same levels of difficulty and ability. Reasonable targeting would normally aim to have normal distribution where the items and person ability are well-targeted (Horton, 2017).

The graphical representation of targeting can be visualised in the person-item threshold distribution in the RUMM2030® software. However, in this study all the distribution is well targeted. The similar graphs were used to define how well the person ability is endorsed according to the level of difficulty items.

As comparison between LS and PC study, Figure 7.2.5 demonstrates half of the items in LS study are treated as difficult items where the item located within the difficulty region on a scale in confectioneries study where the most of participants was located in the middle of the difficulty items. Although the graphs are well shaped, some of the items were unable to response by high ability participants. This is indicate the items are most difficult than the participants' ability in LS study (Camargo and Henson, 2011).

By shifting the scaling technique from LS to PC, the targeting performance can be improved. The evidence was reported in this study that PC demonstrates participants' ability greater than item difficulties as shows in Figure 7.2.6. This indicate that PC much easier to endorsed than LS.

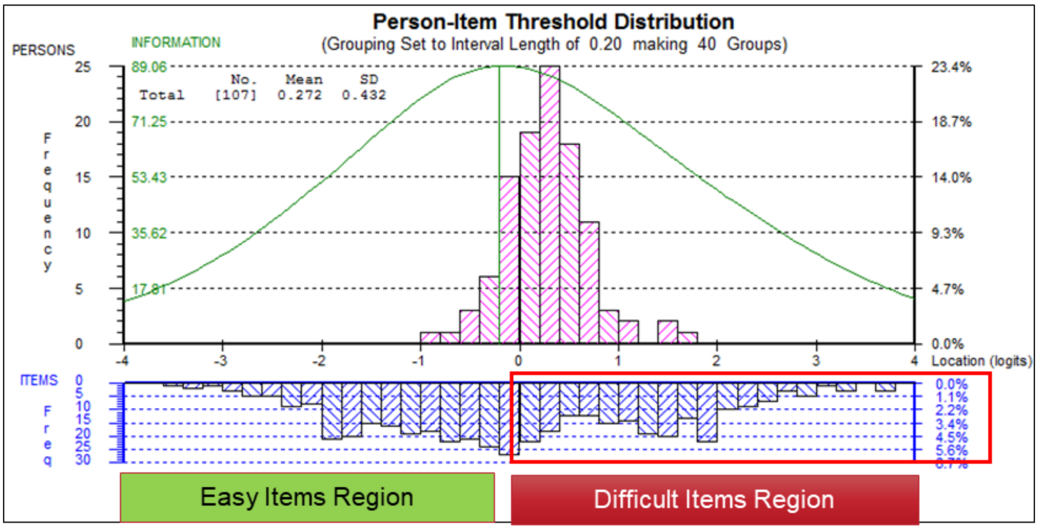


Figure 7.2.5: Graphical representation of targeting in previous LS2011

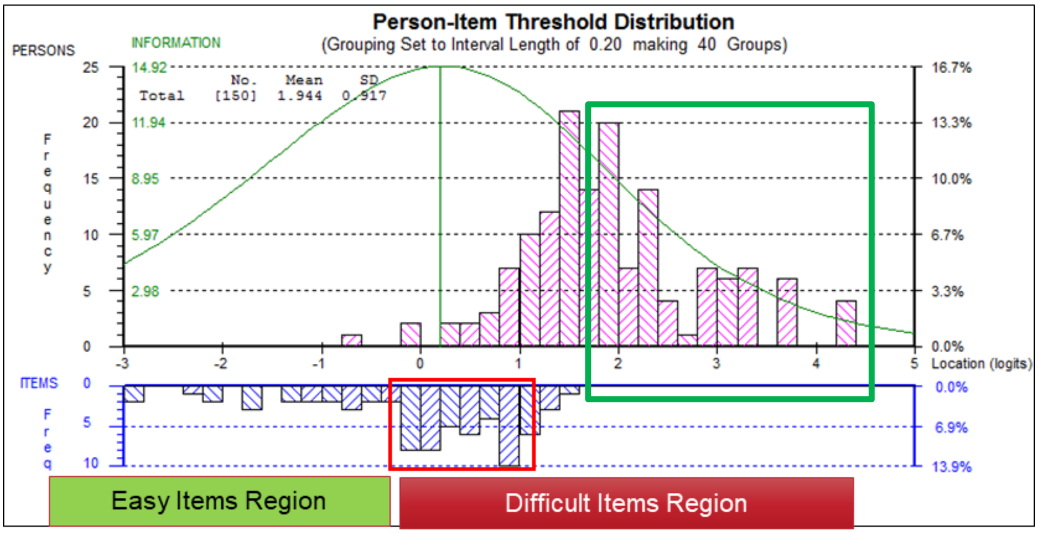


Figure 7.2.6: Graphical representation of targeting in PC1

The similar trend was reported when using interior vehicle texture studies. Figure 7.2.7 and Figure 7.2.8 demonstrates LS study hold difficulties of endorsement where the difficulty items are greater than the participants' ability in LS1 and LS2 methods respectively. In contrast, PC study, has demonstrates participants' ability greater than item difficulties as shows in Figure 7.2.9 exhibit the difficulty items greater than the person's ability in PC2 methods respectively.

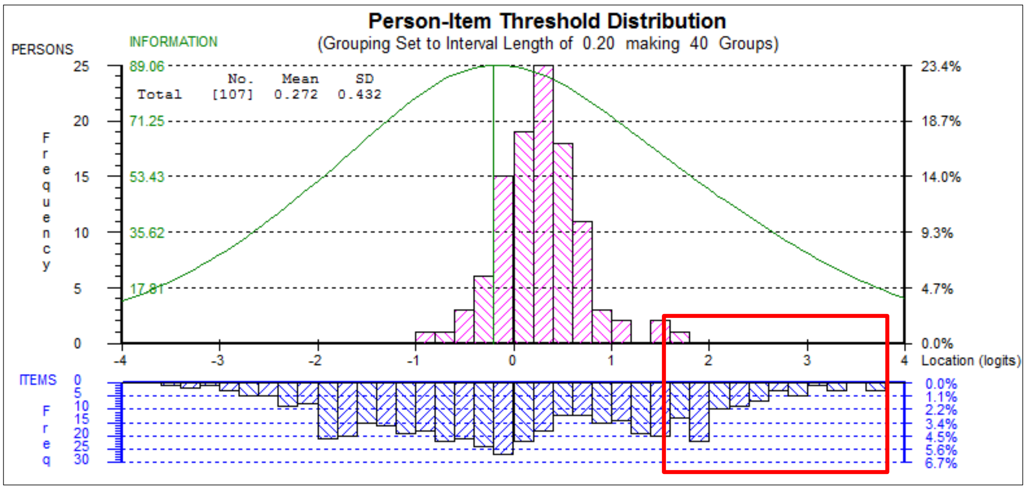


Figure 7.2.7: Graphical representation of targeting in LS1 study

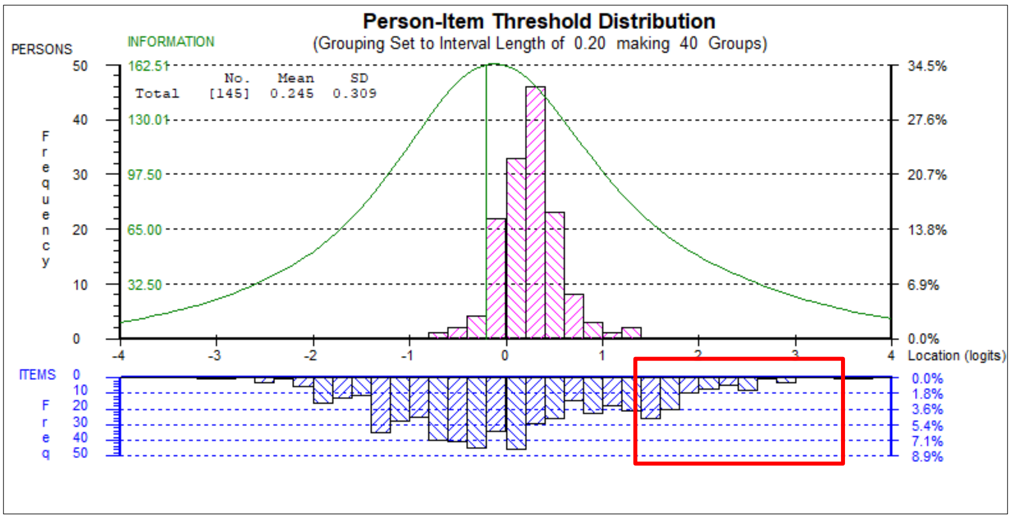


Figure 7.2.8: Graphical representation of targeting in LS2 study

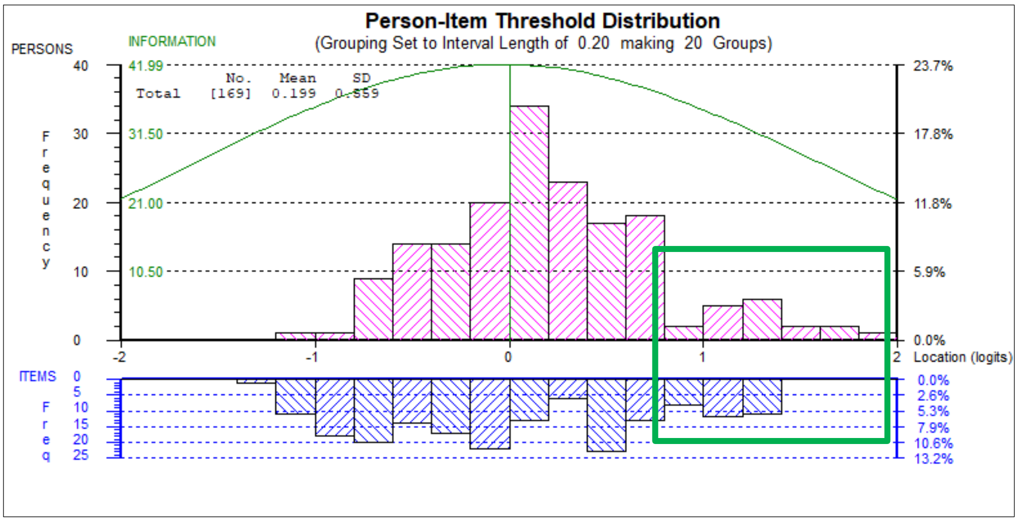


Figure 7.2.9: Graphical representation of targeting in PC2 study

7.2.4 Person ability in PC

The graphical representation in Figure 7.2.10 and Figure 7.2.11 exhibits the comparison of person location logits in endorsing the specialness pieces of confectioneries and PQ of interior vehicle textures respectively. The illustration of the willingness to endorse using PC1 is much higher than in the LS2011 study. This indicates that measuring affective response using PC is effortless.

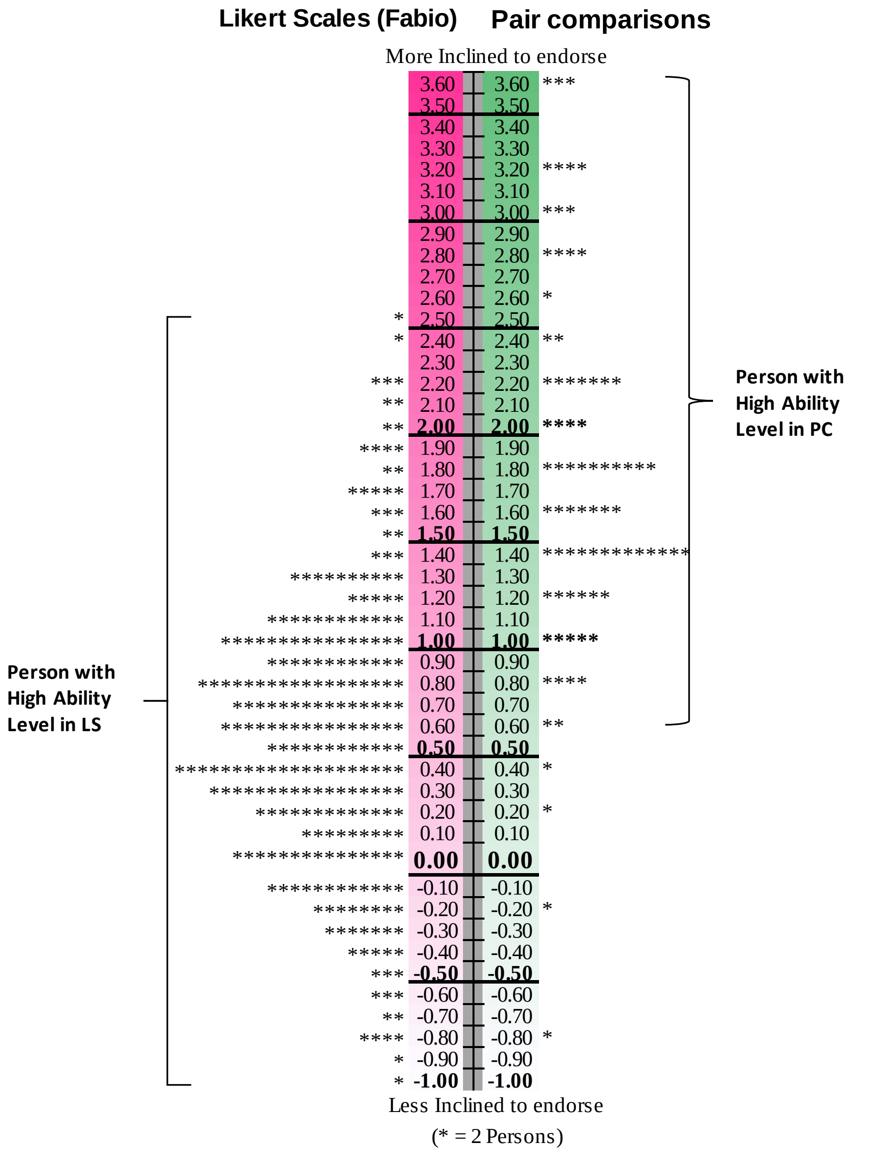


Figure 7.2.10: Comparison person locations between LS2011 and PC1 (Ahmad et al., 2018) using confectionery

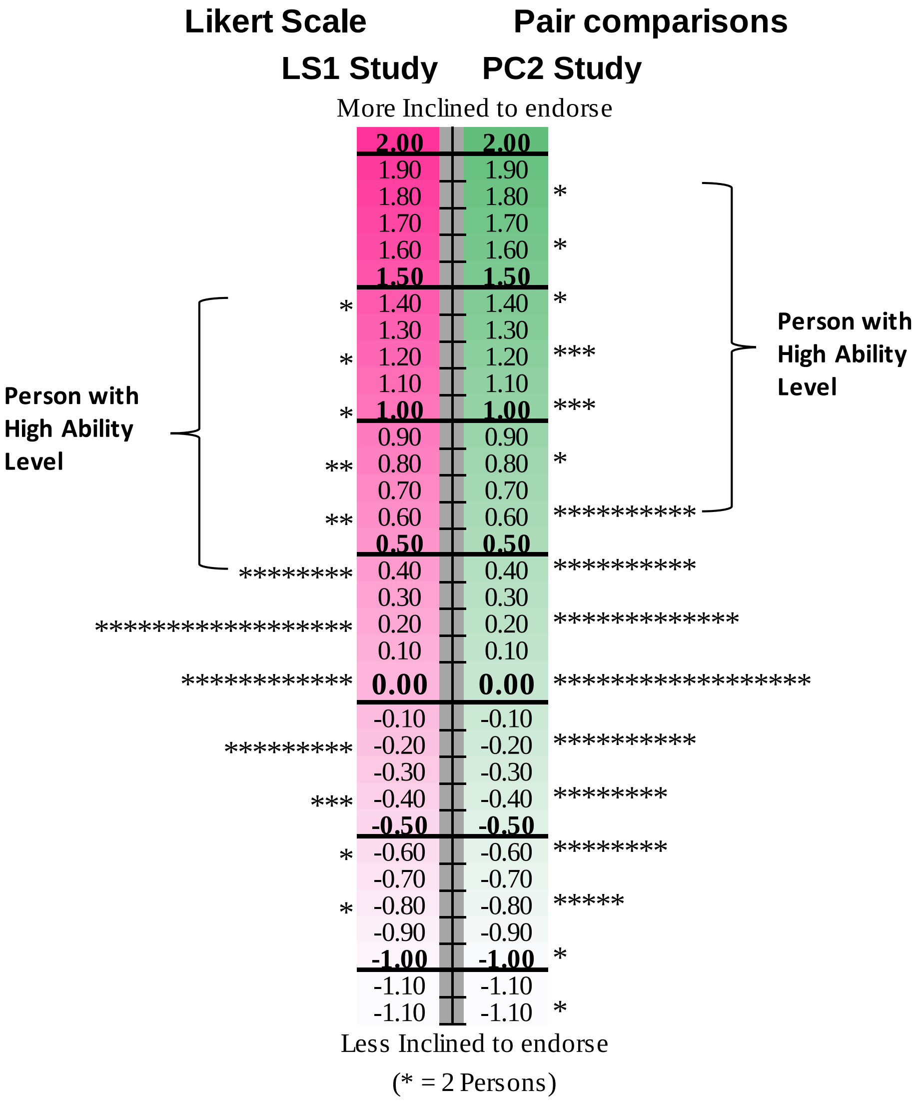


Figure 7.2.11: Comparison of person location logits between LS1 and PC2 using vehicle interior textures

7.2.5 Establishing linear correlations and validity

The strong linear correlation results for the confectionery study was unable to demonstrate similar results in the vehicle interior textures study, as this time, the PC was unable to include the strong predictive validity that PC technique needs to work well with different samples.

However, this is not valid a conclusion and correlation does not imply causality in the relation between a cause and its effect or between regularly correlated events or phenomena (Hole, 2015). First, the correlation can be considered valid depending on the number and variation. Secondly, the correlation should be a genuine association, which means there is no possibility of hidden or intervening variables (Philips Sedgwick, 2012; Hole, 2015).

In this study, bias can be one of sources of argument as to whether both tests are free from bias, although the correlation is perfect in that it is likely to occur by chance only five times in a hundred, which represents a p value < 0.05.

In this study, the weak correlation factors may be used as evidence to determine the source of bias which has stemmed from greater discrimination contrast in the PC2 study that affected a large percentage of unfit participants.

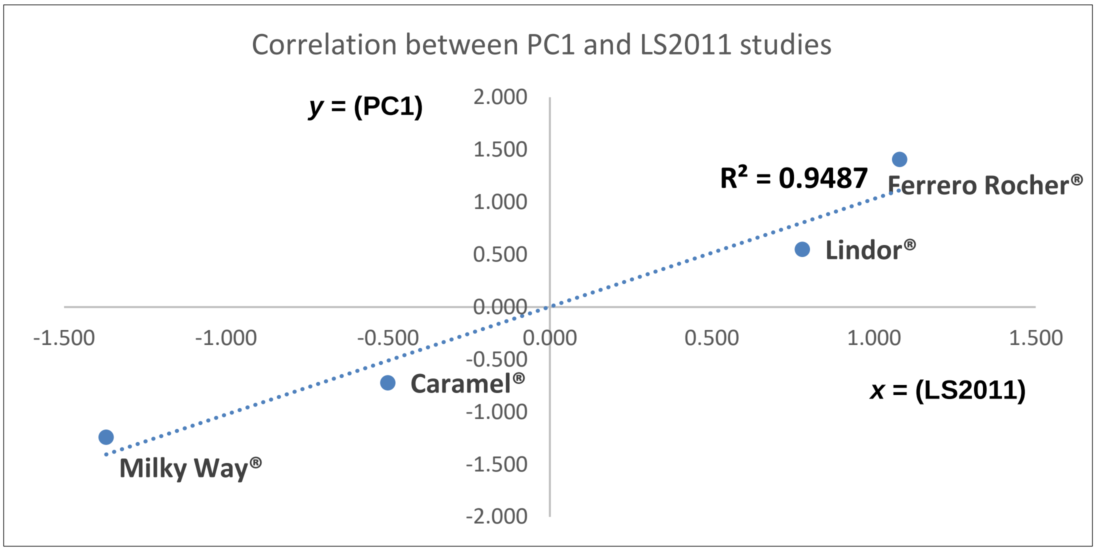


Figure 7.2.12: Strong relationship between PC1 and LS2011using confectionery

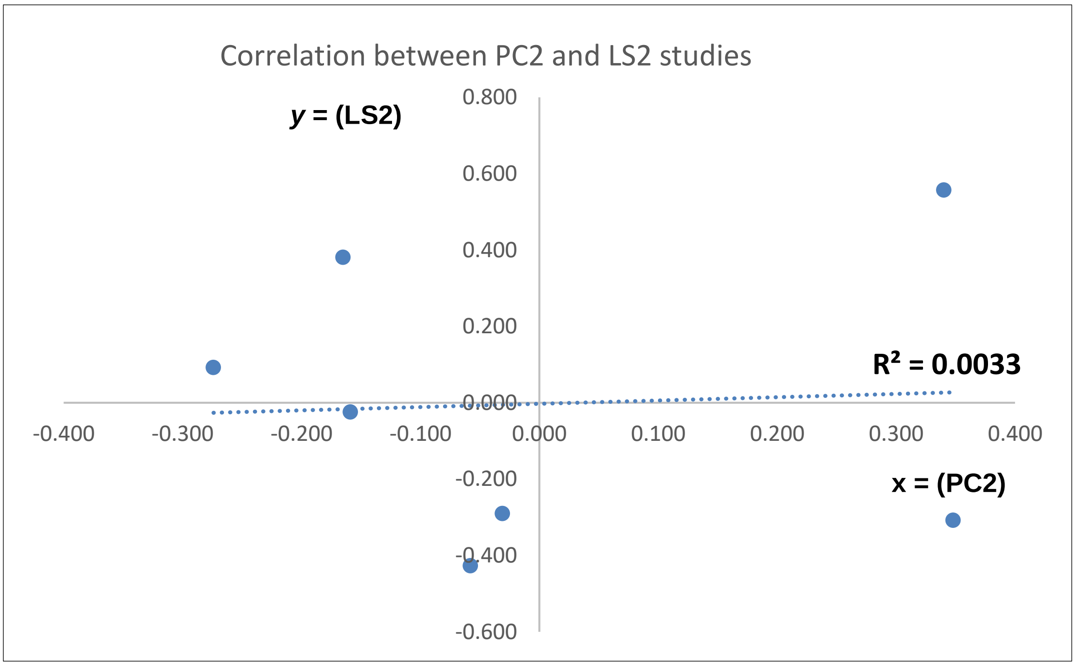


Figure 7.2.13: Weak relationship between PC2 and LS2 study

7.2.6 Data collection speed

Table 7.2.2 Average speed comparison

	Duration of data collection	No. of items	Average speed per item
SDS1	25 – 33 minutes*	140 items	12.42 sec / item
LS1	24 – 35 minutes *	140 items	12.64 sec / items
LS2	24 – 35 minutes*	140 items	12.64 sec / items
PC1	4.25 – 6.17 minutes**	72 items	3.7 sec - 6.1 sec / item
PC2	26 minutes **	189 items	8.25 sec / item
*Manually estimated **Remark : Average time based on online printed in raw database			

The average duration for SDS1, LS1 and LS2 studies was approximately twenty-four to thirty-five minutes with an average speed per item of 12 seconds, while in PC the average duration in PC1 and PC2 is 3.7 seconds to 8.25 seconds.

This clearly indicates that PC can work fast in terms of response rate, compared with the rest to avoid boredom which one of the source of bias. The faster speed in PC1 and PC2 studies were contributed to by the online questionnaire to which the required participants responding using a computer program, while the SDS1, LS1 and LS2 studies used standard paper questionnaires. However, the difficulty of the task was almost identical regardless of using the computer or paper-based options, where the participants are required to evaluate the physical stimuli in every item response.

7.3 CORRELATION COEFFICIENT WITH PHYSICAL STIMULI

The facet results from LS1, LS2 and PC2 were used to observe the correlation coefficient values to test the similarity response with physical elements of the stimuli (Figure 3.6.1, page 54).

For the physical test, three methods were used to assess stimuli with surface roughness measurement using the Hommel Etamic W5 Roughness Tester which this measurement complied with ISO 11562 and 4287 (surface profile - metrological) with the length of dynamic friction of 4.8mm, speed of 0.50mm/s and cut-off of 0.80mm. All the tests were done in a controlled environment with a room temperature of 25 degrees Celsius.

The gloss level measurements were taken using the BYK-Gardner GmbH Multi-Gloss Meter and for the colour measurement the BYK-Gardner GmbH Spectro-Photo Meter Daylight D65/10 degree was used, which comply with CIE lab data specification for automotive applications.

7.3.1 Correlation of surface roughness

Table 7.3.1: Surface roughness measurement

LS1 Logits		PC2 Logits		Surface Glossiness Measurement Rz Profile (mm) Roughness		Roughness Category
T1 / K	0.348	T1 / K	0.210	T1 / K	95	Semi-Rough
T2 / W	-0.274	T2 / W	-0.312	T2 / W	100	Rough
T3 / G	-0.058	T3 / G	0.633	T3 / G	250	Very Rough
T4 / J	-0.031	T4 / J	0.528	T4 / J	150	Rough
T5 / R	-0.165	T5 / R	-0.643	T5 / R	130	Rough
T6 / P	0.340	T6 / P	-0.662	T6 / P	60	Smooth
T7 / Q	-0.159	T7 / Q	0.246	T7 / Q	75	Semi-Rough

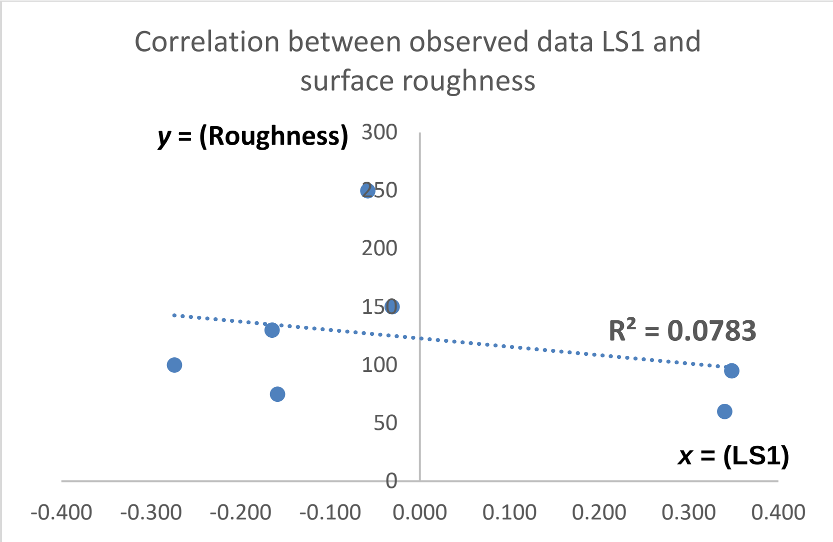


Figure 7.3.1: Weak correlation between LS1 logits and surface roughness

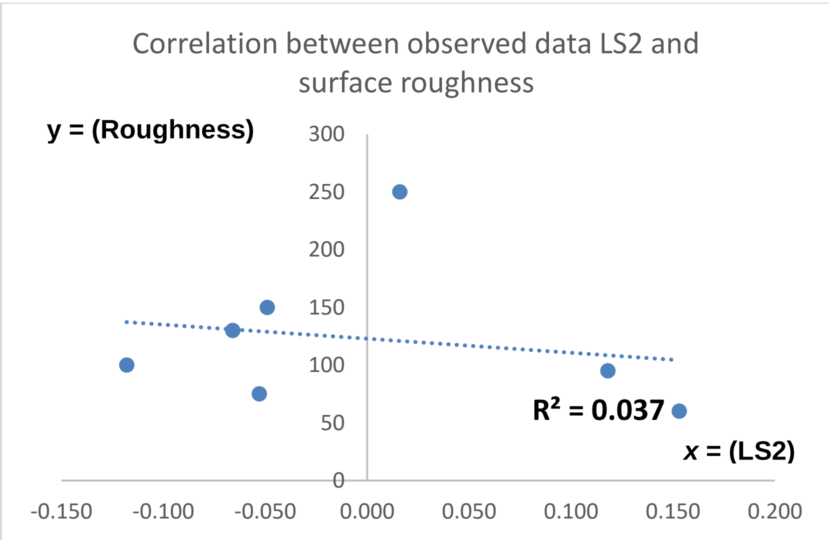


Figure 7.3.2: Weak correlation between LS2 logits and surface roughness

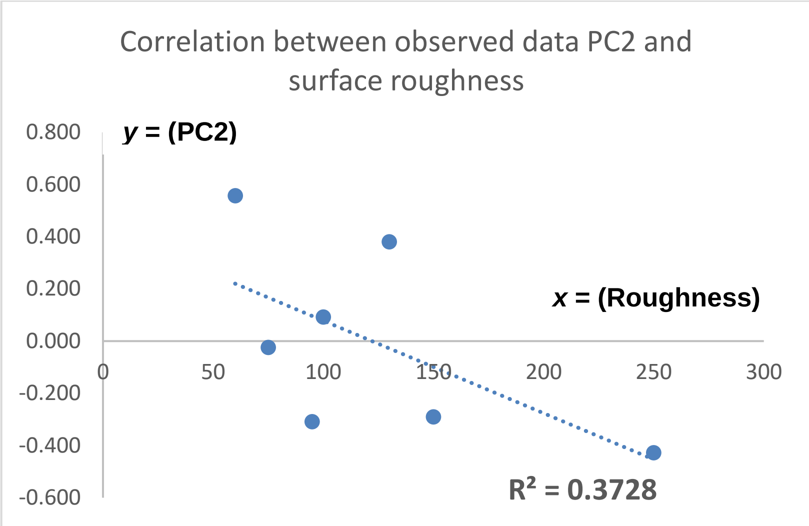


Figure 7.3.3: Weak correlation between PC2 logits and surface roughness

The correlation results demonstrate the observed data from affective responses to the surface roughness levels of the vehicle interior textures as shows in Figure 7.3.1, Figure 7.3.2 and Figure 7.3.3. All the correlation indicates a weak negative correlation between surface roughness and the stimuli logits. The negative indicates that the variability of stimuli has decreased and the variability of surface roughness is increasing. It means the surface roughness attributes did not influence the PQ judgement on the stimuli.

7.3.2 Correlation between surface glossiness

Table 7.3.2: Surface glossiness measurement

LS1 Logits		PC2 Logits		Texture	Gloss Level	Gloss Category
T1 / K	0.348	T1 / K	0.210	T1 / K	2.40	Glossy
T2 / W	-0.274	T2 / W	-0.312	T2 / W	2.70	Glossy
T3 / G	-0.058	T3 / G	0.633	T3 / G	2.30	Semi-Gloss
T4 / J	-0.031	T4 / J	0.528	T4 / J	3.50	Glossy
T5 / R	-0.165	T5 / R	-0.643	T5 / R	2.10	Semi-Gloss
T6 / P	0.340	T6 / P	-0.662	T6 / P	3.00	Glossy
T7 / Q	-0.159	T7 / Q	0.246	T7 / Q	1.70	Matt

The correlation results demonstrate the observed data from the affective responses to the gloss level of the vehicle interior textures as shows in Figure 7.3.4 and Figure 7.3.5.

The correlation indicates a weak correlation between surface glossiness and the stimuli logits. It means the glossiness attributes did not influence the PQ judgement on the stimuli. The glossiness was commonly used to measure the visual appearance of the surface textures whether there is an impact of "*Plasticky*" looks and other attribute measures associated with glossiness.

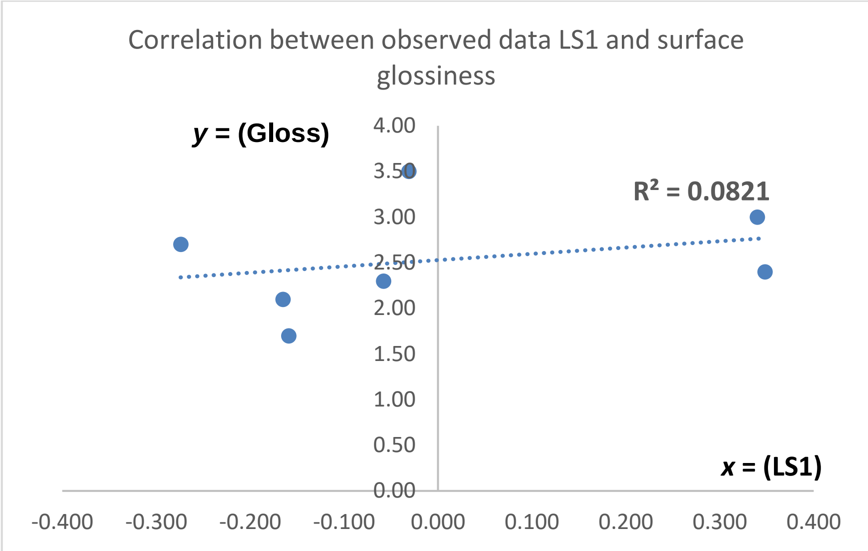


Figure 7.3.4: Weak correlation between LS1 logits and surface glossiness

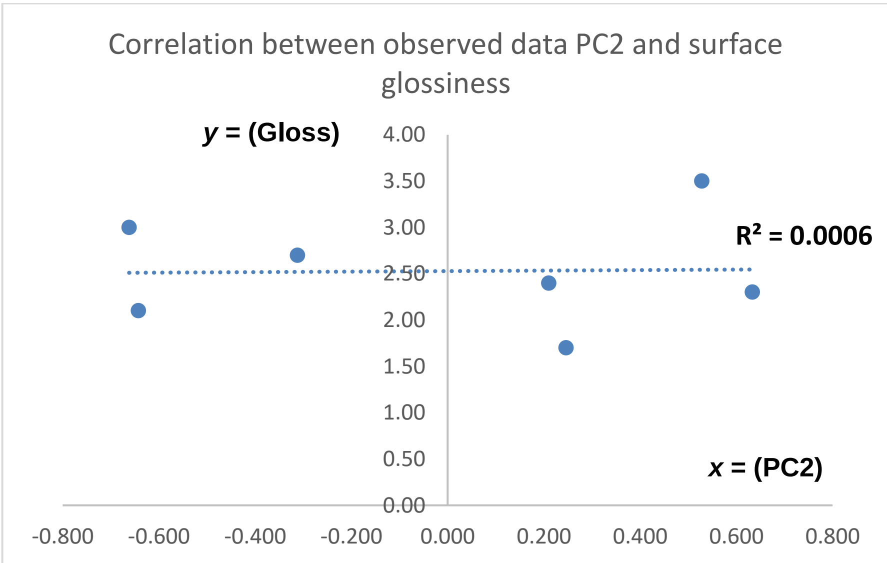


Figure 7.3.5: Weak correlation between PC2 logits and surface glossiness

7.3.3 Correlation with surface colour

Table 7.3.3: Surface colour measurement

NO	Texture Code	Standard CIE L*a*b Colour Data			Colour Sample
		Δ L Delta	Δ A Delta	Δ B Delta	
1	Texture 4 / J	20.84	0.22	2.49	Matt-Black
2	Texture 2 / W	21.66	0.02	0.43	Matt-Black
3	Texture 7 / Q	22.52	0.28	0.71	Matt-Black
4	Texture 3 / G	27.05	0.07	0.78	Grey
5	Texture 1 / K	27.98	0.25	0.58	Grey
6	Texture 5 / R	35.10	1.46	2.93	Brown
7	Texture 6 / P	55.64	4.03	11.09	Beige

The surface colour correlation could not be performed because of the dimensionality factors. The colour measurement was checked using the CIE lab based program, which consists of a three-dimensional axis (ΔL, ΔA and ΔB), which was not possible to correlate with LS1, LS2 and PC2 in one dimension logits. The data value for ΔL determines the achromatic colour value which when close to zero is black, while the value close to one hundred is white. Δa data determines red to green, while Δb data is a yellowish to bluish colour.

7.3.4 Surface colour matching between stimuli

The colour matches were used as guidelines by the auto manufacturer to obtain colour harmonisation across the interior or exterior components in vehicle especial involving difference material for example front bumper made by plastic and body made by compressed metal sheet. Colour matching was used to examine colour differences between components.

However, in this study, the colour matching was used to test whether the stimuli includes colour similarity amongst the stimuli which can affect the level of difficulty in discriminating each stimulus. If the colour is too close, it is very likely the discrimination will affect the difficulty in making good judgements, while the greater contrast seems to be easier to endorse.

Table 7.3.4: Colour matching between stimuli

	Code	Texture 5 / R	Texture 7 / Q	Texture 1 / K	Texture 6 / P	Texture 3 / G	Texture 4 / J	Texture 2 / W
1	Texture 5 / R	0.07	13.33	8.09	22.21	8.78	15.55	13.90
2	Texture 7 / Q	13.18	0.06	5.57	35.45	4.52	2.34	0.92
3	Texture 1 / K	7.96	5.51	0.02	30.36	0.92	7.76	6.85
4	Texture 6 / P	22.18	35.42	30.37	0.04	31.21	37.63	36.17
5	Texture 3 / G	9.04	4.49	0.96	31.25	0.07	6.61	5.73
6	Texture 4 / J	15.29	2.43	7.36	37.37	6.41	0.02	2.22
7	Texture 2 / W	14.09	0.74	6.15	36.04	5.43	2.17	0.01

Table 7.3.4 demonstrates the colour matching in paired matrices in which the correlation between similar colours in the grey box should be zero. However, in automotive application, these tolerances are sometimes difficult to control. The colour may still have small degree of dissimilarity because the colour pigments may not be uniform due to the injection process and the plastic material behaviour changing during the process.

The value in the pink colour box determines the similarity colour between stimuli, which are considered close to each other if the colour tolerance range runs from delta ∆E 0.00 to delta ∆E 1.50 based on Just noticeable difference *(JND)* system (Kirchner and Ravi, 2014), which means any value more than ∆E1.50 is considered greater contrast and, therefore, easy to discriminate by normal visual checking by eye. The acceptable colour tolerance differ between the manufacturers depending on their levels of quality control.

On other hand, in psychophysics the colour tolerance of less than ∆E1.50 is difficult to differentiate by the normal human eye, only for certain colourists with strong colour vision able to spot. The colour close to ∆E0.00 is considered the absolute threshold which defines the lowest level of stimuli contrast or signals of colour, light, touch and feel (Gunter, 1951). The dark grey box in Table 7.3.4 determines the colour is close to ∆E0.00, which cannot be detected by eye but can be measured using a Spectro-Photo Meter. This results demonstrates that stimulus Q is resemble with stimulus W and stimulus K is resemble with stimulus G. The result was also quite similar on opposite direction.

7.4 DICRIMINATING BIAS IN PC

7.4.1 Bias violates the response targeting

Stevens' theory (1946) dealing with stimuli magnitude states that bias may happen when observers do not know how to use familiar units, and the judgements are made by their own subjective units to discriminate (Poulton, 1989). This theory has also reflected on how familiar units can be used to discriminate according to the levels of scale measurement (Stevens, 1946). The greater endorsement might be made at nominal scales, then ordinal, interval and ratios are classified as the most accurate magnitude which equals absolute numbers.

In SDS and LS, it is assumed that the level of difficulty of the item and stimuli at ordinal scale, which participants were asked to discriminate between stimuli, assigns the best rating. However, this study demonstrates some participants incapable of quantifying the magnitude of the contrast into the category scales provided. Bias violates response targeting, as participants do not know how to use the familiar units or quantify subjective responses based on criteria given.

The study has identified that the bias was not only happened when the participants suffer to discriminate the magnitude of items and stimuli into scales in SDS or LS, but bias was also seen if endorsements are too easily; thus, from the statistical point of view we reject those participants with maximum scores, themed as extreme scores, as they affect the greater over-discrimination value of >2.50 logits in RM.

Table 7.4.1: Individual person-item fit comparisons

	PC1	PC2	LS1	LS2	SDS1
Person location	4.226 - High ability -1655 – Low ability	1.899 - High ability -1.196 – Low ability	1.407 - High ability -0.632 – Low ability	1.400 - High ability -0.616 – Low ability	0.379 - High ability -0.242 – Low ability
Person fit residuals	3.837 – Positive Fit Res -2.232 Negative Fit Res	10.044 – Positive Fit Res -8.050 Negative Fit Res	9.098 – Positive Fit Res -9.020 Negative Fit Res	9.222 – Positive Fit Res -10.782 Negative Fit Res	9.767 – Positive Fit Res -12.342 Negative Fit Res
Item fit location	1.467– Most Difficult Item -2.774– Most Easy Item	1.391– Most Difficult Item -1.222– Most Easy Item	1.007– Most Difficult Item -0.543– Most Easy Item	0.966– Most Difficult Item -0.629– Most Easy Item	1.517– Most Difficult Item -1.031– Most Easy Item
Item fit residuals	3.216 – Under discriminate -2.044 – Over discriminate	2.721 – Under discriminate -4.705 – Over discriminate	1.888 – Under discriminate -0.669 – Over discriminate	1.930 – Under discriminate -0.671 – Over discriminate	2.159 – Under discriminate -0.480 – Over discriminate

Table 7.4.1 illustrates the comparative performance of person ability and item difficulty. The higher person ability are demonstrated in the PC study, which is reported as overly-discrimination for the easiest item difficulty and resulting in the greater value of item fit residuals.

On the other hand, in comparing to item difficulty, the SDS1 study was reported as among the most difficult items to discriminate, which the item location shows at 1.517 logits, while both pairwise were exhibited among the easiest items to discriminate with the values exhibited as -2.774 and -1.222 logits, respectively.

In RM analysis, these extreme persons and items are consider misfits. Misfit data in RM will automatically be rejected because these misfit values were greater than the accepted cut-off value in RM. As common practice, misfit data must be discarded during the calibrations procedures as it will corrupt the measurement structure (Camargo, 2013). Additionally, in RMT over-discrimination is one of the major sources of bias.

7.4.2 PC over-discriminating more than SDS

The analysis has found the evidence associated with over-discriminating responses to the items fostered by PC. RM assumes participants are considered misfit to the model if there are negative fit residuals greater than +/-2.50 logits, indicating that the items has been violated by over-discriminating responses (Pallant and Tennant, 2007; Camargo, 2013). As a comparative study, over-discrimination was examined and compared between the SDS1, LS1 and PC1 and PC2 studies.

The analysis identified that PC1 and PC2 have been treated as excessively over-discriminating compared to the same analysis in SDS1 and LS1. This indicates it was too easy to discriminate on the PC versions. The PC2 analysis has illustrated the observing dots plot a steeper curve against the expected curve line than with the other methods. The dots represent that the participants within the class intervals have created a new offset at a given distance from the curve in Figure 7.4.1.

As comparison the analysis on Figure 7.4.2, Figure 7.4.3 and Figure 7.4.4 demonstrates the observing dots quite close to the curve line.

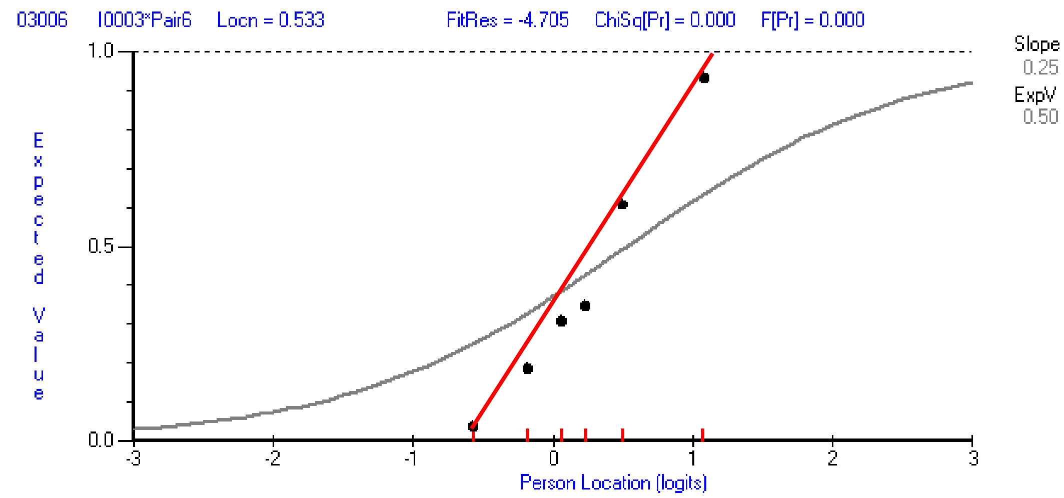


Figure 7.4.1: Over-discriminating items with high negative value in PC2 study

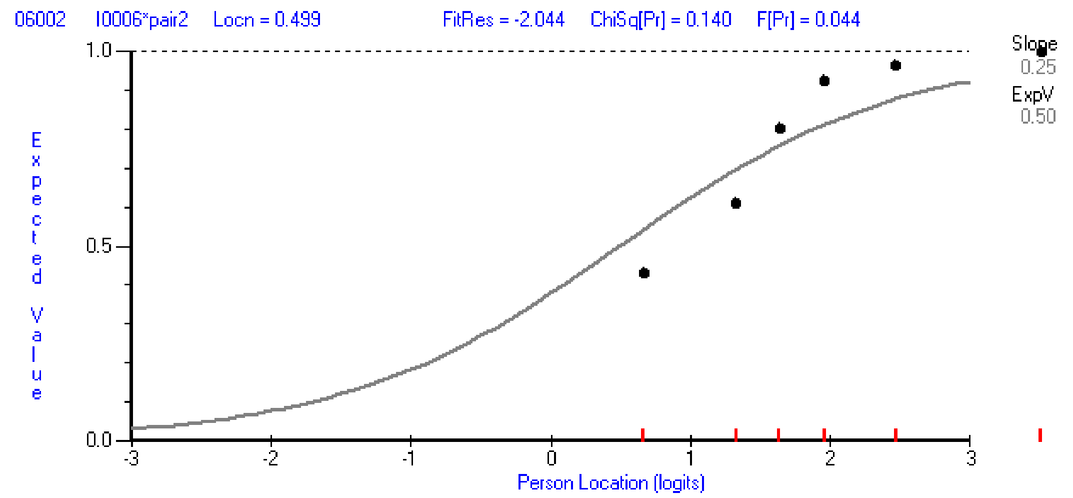


Figure 7.4.2: Over-discriminating items with high negative value in PC1 study

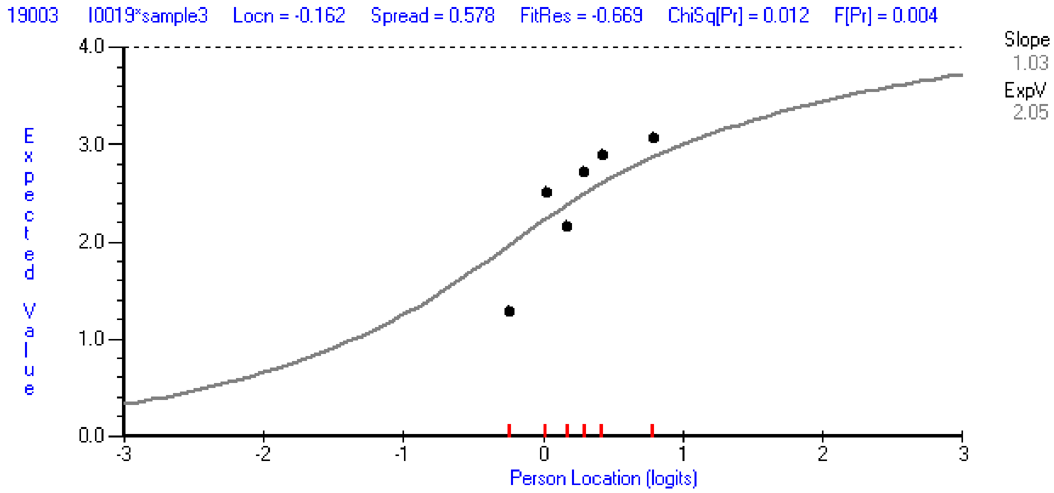


Figure 7.4.3: Over-discriminating items with high negative value in LS1 study

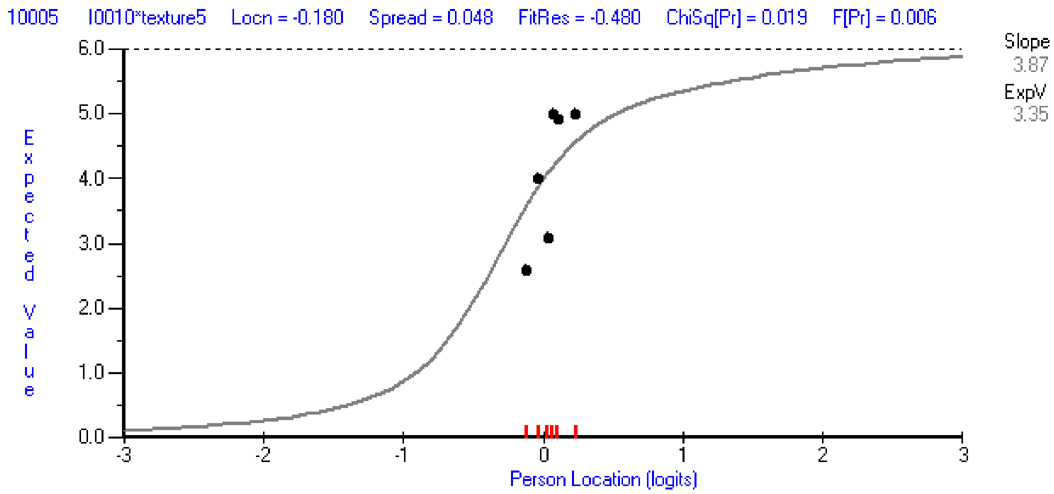


Figure 7.4.4: Over-discriminating items with high negative value in SDS1 study

7.4.3 Stimuli bias in vehicle interior textures

The investigation has determined the poor correlation likely occurred due to sampling bias and stimuli bias.

The colour bias may probably be identified as the source of bias that contributes to the higher degree of discriminating endorsements in the PC study. However, there is no statistical evidence that the colour factor is associated with bias, though there is some logical sense to make the association. Colour may influence the degree of bias because in general, colour gives an impact to visual and overall perceptions. The rationale behind this point is plausible for cases where colour is an important attribute in influencing purchase decisions (Henson and Livesey, 2006; Westerman et al., 2012; Juric et al., 2014; Hosea, 2017).

Verbal evidence during data collection reported that there are some participants that claimed the colour of the stimuli may contain some bias and affect their individual scale-checking method.

On the initial study in AE, monotone-colour stimuli were used to avoid any association with colour bias. However, the decision to use the stimuli with some colour variation is established because there was some feedback from earlier focus groups and pilot test studies claiming the monotone-colour stimuli was difficult to discriminate clearly due to poor contrast and seems to appear identical among other stimuli (Figure 7.4.5).

Figure 7.4.5: Monotone stimuli used during pilot test and focus group

The decision to change back to the colour stimuli was made when considering the attributes used in the survey may require some imaginative effort since the stimuli were not visual products. The aesthetics attributes were used in SDS1, LS1 and LS2 studies such as luxury, contemporary-looking, nice quality, sporty and revitalising are difficult to acknowledge by some of the participants if monotone-colour stimuli were used. The grain pattern and glossiness of the plastic material were not pronounced enough to be able to discriminate those attributes. Thus, during the data collection, the matt black, brown and beige colours were introduced to increase the contrast across the stimuli to aid participants to endorse more clearly (Figure 3.6.1, page 54). However, the colour stimuli unknowingly caused the poor fit residual in statistical analysis because the colour promoted more contrast in greater discrimination in conditions where PC inflates the scaling contrast.

The reason the confectionery study obtained a reasonable statistical fit is the participants are familiar and have strong emotional engagement with the products, but in context of research using vehicle interior textures there was poor emotional engagement, especially from female participants. The stimuli used were unfamiliar for most of the participants; moreover, the stimuli were not visual as completed components like a steering wheel or door trim but just pieces of plastic specimens which results in vagueness to some of the participants. In the evaluation of the PQ of tactile surfaces, most of participants struggled to use familiar units to estimate the magnitude of the products. Most of them did not know how to quantify the quality of the surface. This may result in the evaluation carrying some degree of bias in which the assessor speculates the discrimination are dominated purely in terms of the aesthetic textures and colours which are most noticeable.

The bias was also identified when one of the items asked participants to respond on how the textures would give good feedback when shifting, pulling, turning and rotating. The sample does not help the participants towards good targeting, because the stimuli is not a three-dimensional product.

7.4.4 Sampling bias in vehicle interior texture

The sampling bias is another sources of bias which potentially violates the measurement structure in the PC study.

The objective of this research is to minimise the effect size of bias in the context of test instruments involving the development of items, scaling and stimuli. The decision to narrow down the study to a smaller scope and focus on the three elements designing the questionnaire was made because of limitations of space, time and resources. The study did not cover the sampling bias because it would involve a greater scope of work within the limited space restrictions of this study.

However, during the recruitment process of data collection, the appropriate decision was taken to minimise the bias effect from sampling. The study set prerequisites for the participants who were interested to join the studies that they must meet minimal requirements as advertised through print and email advertisements, in that the participants must be familiar with vehicle interiors and regularly use a vehicle as a passenger or a driver. The advertisements were also briefs on the purpose of the survey and were followed by the appropriate email with participants' information for those interested to participate.

Although the consideration was taken to try and ensure the bias could be minimised as low as possible, it was not possible to control it 100 percent. In this study, data collection used non-random sampling where most of the participants were recruited from within members of staff, undergraduates, postgraduate students and post-doctorate researchers within the University of Leeds. This could be one of the reasons the item-trait interaction is not generalised within the sample population, reflected in the poor Chi-square value.

RM suggests greater sampling would give a more precise measure and improve the validity of the statistical outcome; however, the recruitment process recruited a total of 546 participants which was considerably below the target of 1000 participants as suggested in the RM. Two hundred and fifty samples will give 99 percent confidence that the item calibrations will be stable even under the poor targeting conditions (Horton, 2017). However, increasing the sample size does not reduce bias (Sabin, 2010).

The investigation has also identified that the quality of targeting sampling is poor. Most of the participants are not familiar with the vehicle interiors and do not understand what the PQ is all about and how to assess them correctly. None of the participants are experts in vehicle design nor are they employed by car manufacturers or working within the transport industries; however, most of them are drivers and passengers in a car on a regular basis.

7.4.5 Stimuli bias in confectionery

The discriminating factors may likely reflect the brand and price positioning that influence the specialness attributes explicitly recognised by the participants. The brand-positioning may be one source of stimuli-bias that has been identified in this survey. The quality of evaluating the specialness introduced some perception-bias because the judgements were purely based on packaging aesthetics rather than taste experience.

Participants discriminated well between confectioneries when comparing the opposing price and brand segment for instance in pair three (Milky-way® vs Lindor®) and pair five (Caramel® vs Ferrero Rocher®) because the pairwise combination holds greater contrast in order to discriminate clearly.

However, the participants find it quite difficult to distinguish when comparing between same price segments. Pair six (Lindor® vs Ferrero Rocher®) and pair one (Milky-way® vs Caramel®) involved the most difficult pairwise, because the combination contains only a poor contrast to discriminate clearly.

Participants acknowledge that two out of the four stimuli, Ferrero Rocher® and Lindor®, were classified as premium confectioneries, while Caramel® and Milky Way® were claimed as regular chocolate bars.

Some of the participants, especially those from difference geographical region or international participants, acknowledged at least one out of the four confectioneries used in this study was not available in their market region. This is probably because the particular brand was sold locally but not internationally. Some participants acknowledged that the Lindor® was unattractively packaged compared with Ferrero Rocher®. However, the Lindor® has a better taste compared with Ferrero Rocher®.

7.5 LIMITATION OF THE STUDIES

7.5.1 Insufficient sample size

The limitation in RM is that a large size of sampling is required for accuracy. RM suggests greater sampling would give a more precise measure and bring greater validity to the statistical outcome; however, the recruitment process recruited a total of 546 participants which is considerably below the target of 1000 participants suggested in RM. Two hundred and fifty samples per study will give 99 percent confidence the item calibrations will be stable even under poor targeting conditions (Horton, 2017). However, increasing sample size does not reduce bias (Sabin, 2010).

7.5.2 Higher rejection rate due to misfit persons

The calibrations procedure in RM has discarded a large number of misfit persons and items. The overall rejection rate for all studies exhibited an excessive number of 38.2 percent of the 546 participants registered in SDS1, LS1, LS2, PC1 and PC2 studies. The large rejection contributed significantly from the PC2 study where the rejection rate was 19.4 percent from a total sampling of 546 populations. This amount is quite higher than a similar study in LS done previously in 2013, with a rejection rate at 10 percent from the total participants (Camargo, 2013).

The individual rejection rates were as follows:

- SDS1 study - 35 participants out of 75 participants
- LS1 study - 44 participants out of 107 participants
- LS2 study - 14 participants out of 145 participants
- PC1 study - 10 participants out of 157 participants
- PC2 study - 106 participants out of 169 participants

The basis of RM theory is that the misfit value will not be accepted as valid and reliable data; to a certain extent these groups will corrupt the measurement structure. Therefore, the rejection is necessary in order to obtain good fit statistics.

7.5.3 DIF analysis

No DIF analysis was performed to demonstrate whether bias may be an effect with geographical sampling due to limitations of time and resources. However, in this study DIF was performed to measure demographic factors with the result demonstrating there is no bias associated with age group. DIF can be seen as an important value to the AE study.

7.5.4 Limitation of the RM application

The limitation of the RM software RUMM2030® was identified in the PC1 and PC2 studies. The software does not provide the individual stimulus at means location logits but offers the pairwise location. This is because of the way the stimuli were registered as pairwise and the observed data was administrated using facet design mode which involve three variables of person design, stimuli and item blocks. This is the main reason PairWise© was used to provide the location of each individual stimulus.

7.5.5 Data handling in RM requires expertise

To obtain the accuracies in RM is not an easy task. It’s required some years’ experience in handling complex data analysis using RM software.

The precision analysis would be possibly executed when the researcher, psychometrician or practitioner assessors have essential knowledge in multidisciplinary backgrounds such as mathematics, psychometrics, psychology and statistics as well as familiarity and experience. The developing the test instruments can be a challenge without those skills and can potentially create a bias. The accuracy is depending on who designed the test instruments and how well the survey is going. If the survey objective is unclear, it is highly likely the data collection will be meaningless, imprecise and potentially exposed to biases (Statistical Society, 2007).

7.6 SUMMARY

The comparative scales in PC1 and PC2 in this chapter has demonstrated the robust technique in measurement structure. PC offers less statistical error and provide better visibility and readability that able to make the survey experience much easier and accurate to the statistical view. Unlike Non-comparative scales in SDS1, LS1 and LS2, these category scales have made the survey experience very challenging and were potentially exposed to errors and biases. Thus in this study has recommended and suggest that PC is a viable scaling method to consider when designing the self-report questionnaire.

Chapter 8

Summary and Conclusions

Chapter 8 presents the summary and conclusion findings in relation to the existing literature works, along with the experiment studies of SDS1, LS1, LS2, PC1 and PC2.

PC has discovered the viability of using Rasch analysis to derive linear measurements of affective response, although some of the analyses remain challenges. The strength of PC was validated through two series of experimental studies using confectionery and interior vehicle textures.

RM shows that participants in the two PC studies found it much easier to make paired comparisons than to evaluate the products separately against Likert statements. The advantages of more discriminating stimuli lead to the notable findings that if the products are too different and participants find it too easy to discriminate between them along the affective dimension of interest, then it is likely the data may fit poorly to the Rasch model.

On the other hand, bias in the literature context may be associated with the difficulties of endorsing particular items then the decisions are made using subjective judgement. However, in this study another source of bias, the greater discrimination will leads to bias in the statistical outcome, where over-discriminating the products can violate the person fit residual result.

Through series of this study, some notable points was discover about the characteristic of bias. Bias can happen at any stage of the data survey from recruiting samplings, executing the series of focus groups, developing the items and stimuli, running data collection, analysing results till interpreting the analysis outcome (Sabin, 2010). Bias is a hidden vector for data corruption. Bias is difficult to control and impossible to eliminate; however, it can be predicted and significantly reduced.

Through the series of studies, some key findings may suggest that data collection required deep understanding in every aspect of the procedure. The moderate difficulty level of items, moderate stimuli contrast and the minimal number of scales could be ideal considering sustaining the validity and reliability of statistical results. The over-discriminating target may corrupt the measurement structure if the fit residual is greater than +/- 2.50 logits (Pallant and Tennant, 2007), while under-discrimination will distort the scaling structure. This is the best reason why PC is the better scaling system, offering not only the advantages of discriminating spacing but the elimination of the problematic issue of disordered thresholds in category intervals in SDS and LS. This study has also learnt and understood how bias can violate the measurement structure.

8.1 CONTRIBUTION TO RESEARCH

The rationale of this study is to promote AE in the automotive industries with valid and reliable statistical outcomes to be accepted and become a standard operating procedure in NPI. This study has demonstrated the advantages of PC despite some challenges.

The analysis demonstrates the viability of using Rasch analysis to obtain measures of affective response from PC, that participants find it easier to complete PC compared with evaluating products separately against LS statements.

1. **Minimising biases and measurement error**
   PC improves the person-item fit residual and offers simple calibrations procedures for statistical analysis. PC is also free from disordered thresholds.
2. **Improve targeting and person abilities**
   The study has demonstrated the person location logit has significantly improved in PC, which offers the greater person ability logits than the item difficulty logits.
3. **PC offers an ordinal scale that can be assessed at an interval scale**
   PC provides the best transformation solution in measuring ordinal scales of persons’ attitudes which can be assessed at an interval scales.

4. **Greater discrimination in PC**

   PC transforms poor contrast in category scales in SDS and LS into greater discrimination in PC. It is easy to respond with one click.

5. **PC derives good linearity measurement in affective responses**

   PC method offers a valid and reliable technique to measure persons' attitudes through affective responses to vehicle interior textures and confectionery and established a good linear correlation.

6. **PC offers similar reliability and internal consistency**

   PC provides a high degree of Cronbach's alpha and PSI value.

7. **Faster response rate**

   The response speed much better than in SDS and LS.

8.1.1 Contribution to the knowledge

The study in this thesis derives the following contributions to the knowledge within the field of AE:

1. To date, this is the only written evidence of a comparative study measuring participants' affective responses between SDS, LS and PC.
2. The thesis demonstrates that PC offers a better alternative to the existing SDS and LS methods of assessing affective responses. PC is able to minimise biases and the error effect and allows the responses to be instantly discriminated, more easily and faster due to the greater contrasts. PC transforms the difficulties in SDS and LS into an effortless style.
3. PC allows ordinal scales to be assessed at interval scales, deriving good linearity in affective responses, and offering a valid and reliable method to assess affective responses.
4. PC transform the difficulty of the task of ordinal level which can be biased, to be assessed at the nominal level task and provide the results at interval scales data.

8.1.2 Impact of research

1. In this thesis, the studies contained some useful information due to their novel approach. The novel approach of PC is promoted as a bridge to the knowledge gap between measuring the affective responses to physical stimuli as a viable alternative to existing SDS and LS methods in establishing a good linear correlation.
2. The work contained within this thesis was carried out as a significant contribution to promote AE in the automotive industries. The contribution of this thesis is to promote reliability with an impact on the automotive design guidelines, which comply with International Organisation for Standard ISO 4142:2003 (en) of sensory analysis – guidelines for the use of quantitative response scales (ISO, 2003).
3. PC is an ideal method that can be beneficial for academics and industry practitioners considering PC as a valid and reliable method that offers an alternative to the existing SDS and LS methods in the AE domain. The current gap of knowledge in SDS and LS associated with bias and error in measuring affective responses in consumer products can be minimised through the use of the PC method.

8.1.3 PC offers interval scales being assessed at nominal scales

The advantages of PC from the statistical point of view offer the best transformation solution in measuring the ordinal scales of persons' attitudes, which can be assessed at interval scales.

However, from a psychological point of view, the advantages of PC are robust. PC able to transform the ordinal level task to nominal level task, instead of asking participants to determine the value of the stimuli by category scale which can potentially be biased, participants were only asked to discriminate easily between the pairwise. Then, RM was used to calculate the responses and provide the estimated logits at the interval scale level which was important in statistical analysis as it is used estimate the distance between the stimuli and items.

Figure 8.1.1 illustrates the advantages of PC from the perspective of participants' willingness to endorse response in PC.

Scale of Measurement (Level)	Example Attributes	Vehicle Interior Textures	Nominal	Ordinal	Interval	Ratio
			= / ≠ Equal or Non Equal	< / > Order	+ / - Distance	· / ÷ Ratio
Nominal	High Quality – No Quality		High Quality or No Quality	1 2 3 4 5 No Quality to High Quality	✗	✗
Ordinal	Roughness		Rough or Smooth	1 2 3 4 5 How Rough to How Smooth	✗	✗
Interval	Gloss Level		Gloss or No Gloss	1 2 3 4 5 How Matt to How Gloss	0.00 No Gloss – 3.50 High Gloss	✗
Ratio	Colour contrast		Black or White	1 2 3 4 5 How Black to How White	Achromatic value 0 Black – 100 White	Colour Contrast Ratio 1:1 to 21:1
Difficulty Level		PC SD & LS	Easy ↔ Difficult Easy ↔	Difficult		
Statistical Result		PC SD & LS	→			Logits at interval level

Figure 8.1.1: PQ attributes on scale measurement

8.1.4 PC adding value to AE

The purpose of this study was to develop a systematic approach to how AE can quantify affective responses to vehicle interior textures using RM to obtain valid and reliable outcomes. By doing so, it will help to promote the reliability of AE as a robust method of measuring subjective responses, which has been used in measuring PQ in automotive application. Despite having adequate results in measuring affective responses using vehicle interior texture, PC achieved something novel when this method was able to promote an alternative to non-comparative scaling.

The importance in measuring affective responses using the Rasch model is that it provides an additional value to the PQ assessment. The value is significantly important for the automotive industries as most of the decisions involve huge investment and considerable risk. Inaccurate design proposals resulting from poor analysis can be a huge loss to car manufacturers. Due to the growing market demands and pressure, NPI is becoming shorter for developing new vehicles every year. This translates into the manufacturers working within time sensitive environments in developing products and this requires high statistical accuracy in each of the design proposals. On another hand, the manufacturer requires robust tools to measure the latent variables wherein currently quantitative data from participants is often seen as more valuable (Claudia Newton, 2017).

The vocabulary of accuracy can be interpreted as precision in measuring the affective responses in meeting the expectations in statistical analysis. Thus, in this study, the outcomes demonstrated from AE analysis using RM can be interpreted as "auditing" the design proposal, where RM is used to calibrate each item and to provide item with validity for each stimuli or design proposals.

Why does the accuracy matter in the automotive industries? Because the industry works within the ISO that requires all design works must comply with ISO. Therefore, early Chapter 1 has mentioned one of the reasons why AE was quite slow to be accepted as standard of procedures within design guidelines: the automotive industry required high reliability for quantification measurement and proven methods to use within an R&D environment which was absence in AE. (Schutte and Eklund, 2010).

Therefore, to promote AE in the automotive industry, the AE analysis needs to comply with the ISO standard. The closest standard to measuring affective response can be found in ISO 4142:2003 (en) of *Sensory Analysis – Guidelines for the Use of Quantitative Response Scales* (ISO, 2003). In this standard, the scope describes the principal need to be taken into account when designing the questionnaire using response scales when the response was objectively to obtain intensity of perception. The guidelines also recommend some codes of practice to be taken: for example, the "end effect", described as *"tendency assessors to under-used or over-use the extremities of the response scales"* (ISO, 2003), which means the questionnaire design must optimise the scaling structure to avoid bias effect from extreme samples and the highest and lowest scales value.

8.2 RECOMMENDATION FOR FURTHER RESEARCH

A strong correlation of affective response data between LS and PC1 using confectionery is a great contribution to AE study to demonstrate the concept of PC offering the same advantages to the continuous intervals in the category scales in SDS and LS.

The results in PC2 study demonstrates poor fit and the weak linear correlation when using vehicle interior. However, the result cannot be taken as valid and reliable as in RM, unless this three important properties of estimator is consistency sufficiency and unbiasedness (Engelhard, George, 1997). Additionally the contribution of this research was trivial and new; thus, a repeatable process for measuring affective responses using different application was promoted and enhances a better outcome and stability.

Thus, the future study are suggested to promote and replicate more PC-Likert comparative studies in greater scope within AE studies. To test the hypothesis, two studies proposals were focus in the following domain;

8.2.1 First PC study

The proposed study is a comparative measurement of affective response using Likert-paired comparison to vehicle steering wheels (Figure 8.2.1).

Figure 8.2.1: Likert-paired comparison used with vehicle steering wheels

The objective future work is to investigate whether the linear correlation of affective responses using Likert-paired comparisons can be establish using steering wheels. Steering wheels offer better sensitive judgement in order to extract more properties and information from both static and dynamic evaluations. Most participants are familiar with the products and the most frequent touch area was on the steering wheel, where the most PQ attributes are dedicated in the vehicle interior. Participants need to respond to the items by evaluating, touching, gripping, turning, pushing, rubbing and feeling the stimuli.

The methods are propose using enhanced PC-CAT methods. Participants were ask to compares pair of steering wheel from five units of steering wheels *(Sample A, B, C, D and E)* which creates ten possible pairwise *(AB, AC, AD, AE, BC, BD, BE, CD, CE and DE)* according to difficulties items which has program using Computer Adaptive Testing (CAT). To compare the result, the same produces will replicate using Five-point Likert scales and using a similar procedures. Results were includes fit statistics with linear correlation test.

8.2.2 Second PC study

The proposed study is a comparative measurement of affective response using Likert-paired comparison to vehicle complete seats (Figure 8.2.2).

The objective future work is to investigate whether the linear correlation of affective responses to Likert-paired comparisons and to compare the results on the first study can be established using completed seat assembly.

Seats offer better sensitive judgement to extract more properties and information from the static and dynamic evaluations, through ingress and egress movement. Most participants are familiar with the products and the most frequent touch area was on the seat where the most PQ attributes were dedicated in the vehicle interior. Participants need to respond to the items by evaluating, touching, feeling and sitting on the sample seats.

Figure 8.2.2: Five seat sample wrap with different fabrics

The methods are propose to use enhanced PC-CAT methods. Participants were ask to compares pair of seats from five units of vehicle completed seats *(Sample A, B, C, D and E)* which creates ten possible pairwise *(AB, AC, AD, AE, BC, BD, BE, CD, CE and DE)* according to difficulties items which has program using CAT.

To compare the result, the same produces will replicate using Five-point Likert scales and using a similar procedures. Results were includes fit statistics with linear correlation test with LS scales and PC in study one results.

8.2.3 Data collection methods, processing and analysing:

Future research will focus on identifying the reliable methods for data processing and analysing the data that work well in the PC technique. The following example can be used as an example.

1. **PC-CAT analysis – Paired comparison with CAT**
   To enhance the methodology in the AE study, PC-CAT is potentially the best option. The study could use CAT, which is based on presenting the participant with only the most difficult items according to his or her ability level and increasing or declining the level of difficulty items based on the

participant's ability in responding to the first item (Velozo et al., 2008; Clasen et al., 2010).

The idea is to synchronise a CAT program with PC with powerful tools to visualise the paired-stimuli in random order. CAT software is proven as the basic notion is of an adaptive test which mimics automatically what a wise examiner would do (Choppin, 1976; Velozo et al., 2008; Mike Horton, 2013). CAT uses an algorithm to assign the item difficulty using a special sort of computer-based test, exactly tailored to the participant's ability level. Rasch model would be used to observe, calibrate and validate the response data (Figure 8.2.3).

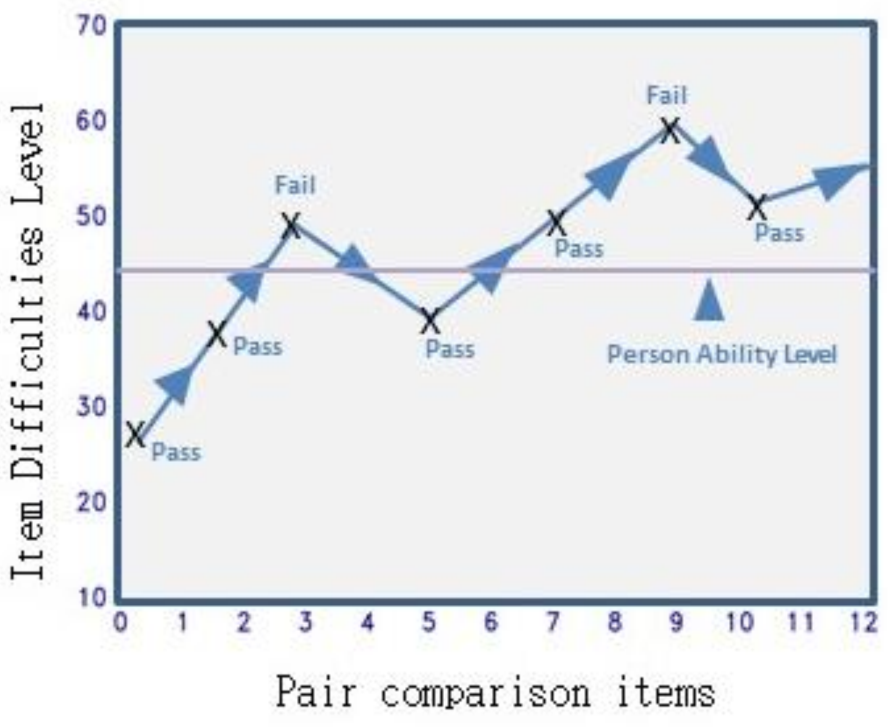


Figure 8.2.3: PC-CAT software (the graphical representation was reproduced from (Linacre, 2000)

8.2.4 Other possible PC Methods;

1. **ConQuest 2015** (Alvin and Adams, 2015) is a logistics modelling of PC software using BTL. This approach estimates a single parameter based on PC to construct a ranking of all objects at once. Judgement is made between two alternative objects, then differentiated among a large set of

objects and placed on an interval scale. ConQuest was used to estimate a paired game and sport tournament.

2. **Discrete Choice model**

   The nested logic model concept may possibly be used to analyse the PC. In the nested logic concept, the variable of vehicle attributes will group into several class or nest. Nested logic uses utility functions to calculate a logits model within the subset. Analysis can capture part of the utility U_{int} of an alternative *i* for participants *n* in choice of situation of *t* in PC (Hess et al., 2011; Heckmann et al., 2013)

8.2.5 Prerequisite suggestion for future studies

1. Promote and replicate more PC-Likert comparative studies in greater scope within the AE Eco-system and psychophysics-related studies.
2. Focusing on how PC is able to derive better statistical models with less complexity but practical.
3. Using unidimensionality items in AE focus groups and pilot test items. This means using the items from a pool of high-quality item-banking (Choppin, 1976), then calibrating the immature items if newly developed from affective words to avoid ambiguity and multidimensionality.
4. Predict the items bias, scales bias and stimuli bias at as early a phase as possible.
5. Reinforce against random sampling bias within the target population with appropriate screening methods, example stratification, and clustered and simple random samples.
6. Comparative studies should use a sampling and items, and the same number of participants and same number of items.
7. Use or design stimuli that are familiar to the participants. Good stimuli are defined as when they are able to stimulate participants' emotions and able to use familiar units to estimate the properties.

8.3 CONCLUSIONS

The overall aim of this research is reachable, although some of the objectives have resulted in an inadequate statistical fit. The notable contributions in this study are tailored to the aims and goals of the study.

The synergy between the AE and RM methods is robust in measuring affective responses. RM has added the reliability value to the AE method which were previously absent. However, this study has discovered some limitations which are associated with the scaling methods uses within AE studies.

The primary objective is to test the hypothesis in assessing bias and error when measuring affective response using SDS and LS has discovered a similar outcome as claims by AE scholars. The study has determined some tested evidence are associated with bias and error that has reinforced the validity of evidence as reported in the literature. The investigation has discovered the category scales offer in SDS and LS inflates the degree of missed targeting which resulting in the difficulty of endorsing the scale-questionnaire. The cleaning data process becoming tedious because of the complexity to calibrate the datasets in meeting good statistical fit were also challenging.

To minimise the effect size of bias and error, the following objectives are therefore to test how well the concept of PC from Bradley, Terry and Luce (1952) able to improve the problematic of category scale in non-comparative scales. Through a series of studies using confectionery, the study has demonstrated the viability of the PC method in measuring affective responses that PC works well in reducing the biases effect of missed targeting. The successful findings in when comparing historical data of LS2011 with PC1 has demonstrated notable achievements when participants in the research found it much easier to make paired comparisons than to evaluate the products separately against Likert statements.

In the context of data administrative, PC demonstrates the data calibration process is less tedious where the calibration process in achieving good linear measurement.

The aim of this study to minimising the affect size of biases and errors. One of the reasons why data could be troublesome when the items are exposed from local dependencies such as grammatical ambiguity which carries multiple interpretations and resulted the response could be distorted and biases. Thus in this study has hypothesised that the size of biases can be minimised when the items used in scale-questionnaire are unidimensional best. The outcome from this study demonstrates the new items develop for PC2 study was unidimensionally fit through a series of independent T-test and binomial test. The findings in the unidimensional study have significantly improved the target judgement and satisfy the measurement structure of RM.

The comparative assessment using confectioneries in LS2011 with PC1 study has endeavoured the used of the PC method has affecting good linear measurement. Although the notable achievement was made in PC, there are some disadvantages have been discovered when applying vehicle interior textures. Despite maintaining the similar performance when endorsement the affective response using PC, this study has discovered the disadvantages of PC where it has promoted greater contrasts to the alternative pairs that motivate the participants to excessively discriminate when responding in PC study. The impact of excessive discrimination on the items and stimuli from its effortless-style-checking in PC has inflated a large number of extremes and disqualified participants. The impact of greater discrimination in PC stems a large number of misfit persons that need to need to be removed as it has fallen outside the acceptance fit residual measures governed in RM. High residual indicates the participants most likely response to the items in unexpected pattern response.

The comparative assessment to validate the final hypotheses in this study to observe whether a linear correlation of affective response between LS1, LS2 and PC2 using vehicle interior textures studies can possibly be achieved as in the confectioneries study. However, the study was unable to carry a similar performance. PC2 is unable to hold sufficient stability across different samples where PC2 is not able to construct linear measurements of affective responses from paired comparisons using vehicle interior textures. This indicates that some of the properties were an identified misfit. The study, however, speculates the biases and errors due to poor stimuli contrast that make the discrimination difficult. If the products are a too different or greater contrast, participants found it too easy to discriminate.

However, the result cannot be taken as valid and reliable as in RM, unless this three important properties of estimator is consistency sufficiency and unbiasedness. Data will only consider valid and reliable when the estimator of true value parameter is consistence in condition where the sample size is increases and the test was repeated (Engelhard, George, 1997).

This research is therefore offering a contribution to the understanding in an important emerging area. However, the contribution of this research is was innovative; thus, a repeatable process for measuring affective responses using different application would able to promote and enhances a better statistical outcome.

This study concludes that the advantages of PC in innovating the effortless style checking to overcome the shortcoming of poor discrimination contrast in SDS and LS has motivated the participants excessively discriminate between the alternate pair of items. On other hands, the expected item correct is really depending on how the stimuli were chosen. The limitations in PC scaling in measuring affective responses will be addressed in a future study.

REFERENCES

Aaker, D.A. and Biel, A.L. 1993. *Brand equity & advertising : advertising's role in building strong brands*. Lawrence Erlbaum Associates.

Abidin, S.Z., Othman, A., Shamsuddin, Z., Samsudin, Z. and Hassan, H. 2014. The Challenges of Developing Styling DNA Design Methodologies for Car Design *In*: *DS 78: Proceedings of the E&PDE 2014 16th International conference on Engineering and Product Design, University of Twente, The Netherlands*.

AG, A. 2001. Progress You Can Touch: The 'Audi Feeling' A. World, ed. *Audi World*. [Online]. Available from: http://www.audiworld.com/news/02/haptics/content.shtml.

Ahmad, F., Holt, R. and Henson, B. 2018. Measuring Affective Responses To Confectioneries Using Paired Comparison. *7th International Conference on Kansei Engineering and Emotion Research 2018*. (146), pp.245–254.

Ajovalasit, M., Tajadura, A., Shabani, A. and Giacomin, J. 2013. Human emotional response to steering wheel vibration in automobiles. *Vehicle Noise and Vibration*. 9(Special), pp.109–128.

Al-hindawe, J. 1996. Considerations when constructing a semantic differential scale. *La Trobe working papers in linguistics*. 9(7), pp.1–9.

Alvin, V. and Adams, R. 2015. *Modelling pairwise comparison using the Bradley-Terry-Lucy (BTL) Model* [Online]. Melbourne, Australia. Available from: https://www.acer.org/files/Conquest-Notes-2-PairWiseComparisons.pdf.

Andrich, D. 1988. *Rasch Model for Measurement*. London: SAGE Publications Inc.

Anita Mohd Lokman and Kamaruddin, K.A. 2010. Kansei Affinity Cluster for Affective Product Design. *International conference on User Science Engineering (i-USEr)*., pp.38–43.

Anselmi, P., Vidotto, G., Bettinardi, O. and Bertolotti, G. 2015. Measurement of change in health status with Rasch models. *Health and Quality of Life Outcomes*. 13.

Antariksa, Y. 2018. What is Perceived Quality? *Studymarketing.org*. [Online]. Available from: http://www.studymarketing.org/articles/Brand_Management/What_is_Perceived_Quality?.html.

Arikan, C.A. and Gelbal, S. 2013. Investigation of required qualification of instructor by mean of pair-wise comparison method. *Procedia-Social and Behavioral Sciences*. 106, pp.1603–1613.

Aziz, A.A., Masodi, M.S. and Zaharim, A. 2013. *Asas Model Pengukuran Rasch, Pembentukan skala dan struktur pengukuran (Rasch Measurement Model basis, the establishment of the scale and structure of measurement)*. Bangi, Selangor, Malaysia: Penerbit Universiti Kebangsaan Malaysia.

Barnes, C. and Lillford, S.P. 2009. Decision support for the design of affective products. *Journal of Engineering Design*. 20(5), pp.477–492.

Bertram, D. 2007. Likert Scales… are the meaning of life : *University of Calagary, Department of Computer Science*., pages.cpsc.ucalgary.ca/~saul/wiki/uploads/CPSC681/.

Bhise, V.D. 2012. *Ergonomics in the automotive design process*. Michgan, USA: CRC Press.

Bhise, V.D., Sonal Onkar, Marc Hayes, Jim Dalpizzol and Dowd, K. 2007. Touch Feel and Appearance Characteristics of Automotive Door Armrest Materials. *SAE Technical Paper Series*. 01(1217).

Bjorner, J.B., Kisinki, M. and Jr, J.E.W. 2003. Calibration of an item pool for assessing the burden of headaches: An Application of item response to the Headache Impact Test. *Journal of Quality of Life Research*. 12, pp.913–933.

Böckenholt, U. and Dillon, W.R. 1997. Modeling within-subject dependencies in ordinal paired comparison data. *Psychometrika*. 62(3), pp.411–434.

Bond, T.G. and Fox, C.M. 2015. *Appliying The Rasch Model, Fundamental Measurement in the Human Science* Third Edit. New York: Routledge.

Borsboom, D. 2006. When Does Measurement Invariance Matter? *Medical*

Care. [Online]. 44(Suppl 3), pp.S176–S181. Available from: http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage&an=00005650-200611001-00023.

Bourke-Taylor, H.M., Pallant, J.F. and Law, M. 2013. Update on the Child's Challenging Behaviour Scale Following Evaluation Using Rasch Model Analysis. *Child : Care, Health and Development*. 40(2), pp.242–249.

Bradley, R.A. and Terry, M.E. 1952. Rank Analysis of Incomplete Block Designs: The methods of Paired Comparisons. *Biometrika*. 39(3/4), pp.324–345.

Bramley, T. and Black, B. 2008. Maintaining performance standards : aligning raw score scales on different test via a latent trait created by rank-ordered examinees' work. *Third International Rasch Measurement Conference, University of Western Australia, Perth*.

Bramley, T. and Oates, T. 2011. Rank ordering and paired comparisons – the way Cambridge Assessment is using them in operational and experimental work. *Research Matters*. (11), pp.32–36.

Brown, A. and Maydeu-Olivares, A. 2017. Modelling forced-choice response formats. *The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test Development*. 2–2, pp.523–569.

Burnett, G. and Irune, A. 2009. Drivers' Quality ratings for switch in cars : Assessing the role of the vision, hearing and touch senses. *Procedings of the First International Conference on Automotive User Interface and Interactive Vehcular Applications (AutomotiveUI 2009)*., pp.107–114.

Camargo, F. and Henson, B. 2015. Aligning Affective Responses with Fabric Features of Vehicle Seat: An Approach Using the Rasch Measurement Model. *International Journal of Affective Engineering*. 14(3), pp.193–202.

Camargo, F.R. 2013. *Calibrating Scales for Affective Responses to Physical Features of products using Rasch measurement theory*. Leeds: University of Leeds.

Camargo, F.R. De 2010. *Caliberating Measurement of Affective Response to Product Applying the Rasch Model*. Leeds: University of Leeds.

Camargo, F.R. and Henson, B. 2015. Beyond usability: designing for consumers' product experience using the Rasch model. *Journal of Engineering Design*. 26(4–6), pp.121–139.

Camargo, F.R. and Henson, B. 2014. Establishing Metrics for Kansei Responses : An Approach Using the Rasch Model. , pp.585–594.

Camargo, F.R. and Henson, B. 2011. Measuring affective responses for human-oriented product design using the Rasch model. *Journal of Design Research (JDR)*. [Online]. 9(4), pp.360–375. Available from: http://www.inderscience.com/offer.php?id=43363.

Carman, J.M. 1990. Consumer Perceptions Of Service Quality: An Assessment Of T." Journal of retailing 66.1 (1990): 33. *Journal of Retailing*. 66(1), p.33.

Cattelan, M. 2012. Models for Paired Comparison Data: A Review with Emphasis on Dependent Data. *Statistical Science*. [Online]. 27(3), pp.412–433. Available from: http://projecteuclid.org/euclid.ss/1346849947.

Champagne, A.J. and Shiau, N.-M. 1997. Commercial Van Diesel Idle Sound Quality. *Society of Automotive Engineers, Inc*.

Charles, E.O. and Suci, G.J. 1955. Factor Analysis of Meaning . *Journal of Experimental Psychology*. 50, pp.325–338.

Childs, T., Agouridas, V., Barnes, C. and Henson, B. 2006. Controlled Appeal Product Design: a Life Cycle Role for Affective (Kansei) Engineering. *13th CIRP International Conference on Life Cycle Engineering*., pp.537–542.

Choppin, B. 1976. Developments in Item Banking. *Monitoring National Standard of Attainment in Schools*. [Online]. (June), pp.216–234. Available from: http://www.rasch.org/memo76.pdf.

Choppin, B. 1968. Social class and educational achievement. *Journal of Education Research*. 10(3), pp.213–217.

Christensen, K.B., Makransky, G. and Horton, M. 2017. Critical Values for Yen's Q3: Identification of Local Dependence in the Rasch Model Using Residual Correlations. *Applied Psychological Measurement*. 41(3), pp.178–194.

Clasen, S., Winter, S.M., Velozo, C.A., Bedard, M., Landford, D.N., Brumback,

B. and Lutz, B.J. 2010. Item Development and Validity Testing for a Self-and Proxy Report : The Safe Driving Behavior Measure. *The American Journal of Occupational Therapy*. 64(2), pp.296–305.

Claudia Newton 2017. *Towards Sustainable Luxury Materials Selection : Measuring the Perceived Quality of Automotive Interior*. Te University of Warwick.

Cliff, N. 1973. SCALING. *Annual Review of Psychology*. [Online]. 24, pp.473–506. Available from: www.annualreviews.org.

Cook, J.D., Hepworth, S.J., Wall, T.D. and Warr, P.. 1981. *The experience of work*. London Academic Press.

Cortes, M. 2013. Thesis Questionnaire. Available from: https://www.scribd.com/presentation/177310480/Meljun-Cortes-Thesis-Questionnaire.

Cox, E.P. 1980. The optimal number of response alternative for a scale : A review . *Journal of Marketing Research*. 17(4).

Dalitz, C., Wilberg, J. and Bednarek, K.E. 2018. Paired Comparison Sentiment Scores. Available from: http://arxiv.org/abs/1807.03591.

Delgado-Rodriguez, M. 2004. Bias. *Journal of Epidemiology & Community Health*. [Online]. 58(8), pp.635–641. Available from: http://www.jstor.org/stable/25570444.

Desmet, P. and Hekkert, P. 2007. Framework of Product Experience. *International Journal of Design*. 1(1), pp.57–66.

Desmet, P.M.A. and Hekkert, P. 2009. Design & Emotion. *International Journal of Design*. 3(2), pp.1–6.

Dolan, P., M. Hallsworth, D. Halpern, G.King, R. Metcalfe and Vlaev, I. 2012. Influencing Behaviour : The mindspace way. *Journal of Economic Psychology*. 33, pp.191–208.

Dolnicar, S. and Grun, B. 2013. Translating Between survey answer formats. *Journal of Business Research*. 66(9), pp.1298–1306.

Dunn-Rankin, P., Knezek, G.A., Wallace, S. and Zhang, S. 2004. *Scaling*

Methods Second edi. New Jersey, London: Lawrence Erlbaum Associates Publishers.

Ekman, G. and Lennart, S. 1964. Scaling. *Annual Review of Psychology*. 16(May), pp.451–474.

Engelhard, George, J. 1997. Rasch Measurement Theory, the Method of Paired Comparisons, and Graph Theory. *The annual Meeting of the Southwest Educational Research Association*. [Online]. 23(77843–4225), pp.1–33. Available from: http://ericae.net/ft/tamu/STEPWIS.htm.

Farsky, M., Schnittka, O., Sattler, H., Höfer, B. and Lorth, C. 2017. Brand-anchored discrete choice experiment (BDCE) vs. direct attribute rating (DAR): An empirical comparison of predictive validity. *Marketing Letters*. 28(2), pp.231–240.

Fernandes, M.M., Rosati, A.C., Neto, didimo G., Goto, F.K., Jr, H.M. and Mologni, J.F. 2008. Quality Function Deployment (QFD) and Pugh Matrix on Innovative Concept Selection : An application in automotive sector. *SAE Technical Paper Series*. 36(0017).

Field, A. 2013. *Discovering Statistics Using IBM SPSS Statistic 4th Edition*. London: Sage Publications Ltd.

Fischer, G.H. and Molenaar, I.W. 1995. *Rasch model : Foundation, recent developments, and applications* Papers ori. New York: Springle - Verlag.

Florin, R. 1999. *A Rasch Paired Comparison of Total Physician Work* [Online]. Waltham, MA. Available from: http://rasch.org/florin.htm.

Floyd J. Fowler, J. 1995. *Improving Survey Questions, Design and Evalutation*. California: SAGE Publications Inc.

Gallagher, J. 2016. Pair Comparison Program. Available from: http://www.phdproject.co.uk.

Garner, M. and Engelhard Jr., G. 2009. Using paired comparison matrices to estimate parameters of the partial credit Rasch measurement model for rater-mediated assessments. *Journal of applied measurement*. 10(1), pp.30–41.

Glickman, M.E. 1999. Parameter estimation in large dynamic paired comparsion expeiments. *Journal of the Royal Statistical Society. Series C (Applied Statistic)*. 48(3), pp.377–394.

Green, P.E. and Srinivasan, V. 1978. Conjoint Analysis in Consumer Research: Issues and Outlook. *Journal of Consumer Research*. [Online]. 5(2), p.103. Available from: https://academic.oup.com/jcr/article-lookup/doi/10.1086/208721.

Grider, R.C. and Malmberg, K.J. 2008. Discriminating between changes in bias and changes in accuracy for recognition memory of emotional stimuli. *Memory and Cognition*. 36(5), pp.933–946.

Guan, S.S. and Luo, M.R. 1999. Investigation of parametric effects using large colour differences. *Color Research and Application*. 24(5), pp.356–368.

Gulliksen, H. 1958. How to make meaning more meaningful *In*: A. P. Association, ed. *Contemporary psychology* . Chicago: Aldine Publishing Company, pp.115–118.

Gunter, B.Y.R. 1951. The Absalute Thresholds for Vision in the CAT. *Journal of physiol*. 114, pp.8–15.

Hassan, H. and Samsudin, Z. 2012. *Proton Perceived Quality* [Online]. Shah Alam, Malaysia. Available from: www.proton.com.

Heckmann, C.G., Michalek, J.J., Morrow, W.R. and Liu, Y. 2013. Sensitivity of Vehicle Market Share Predictions to Alternative Discrete Choice Model Specifications. *Proceedings of the ASME 2013 ASME International Design Engineering Technical Conferences*. 136(December), pp.1–9.

Heise, D.R. 1969. Some Methodological Issues in Semantic Differential Research. *Psychology Bulletin*. 72(6), pp.406–422.

Hekkert, P. 2006. Design Aesthetics : Principles of pleasure in design. *Psychology Science*. 48, pp.157–172.

Helander, M. and Khalid, H. 2014. *Human Factors in Affective Product Design*. Kuala Lumpur: Human Factors and Egronomics Society Malaysia.

Helander, M.G., Khalid, H.M., Lim, T.Y., Peng, H. and Yang, X. 2013. Emotional needs of car buyers and emotional intent of car designers. *Theoretical Issues in Ergonomics Science*. 14(5), pp.455–474.

Heldsinger, S. and Humphry, S. 2010. Using the Method of Pairwise Comparison to Obtain Reliable Teacher Assessments. *Australian Educational Researcher*. 37(2), pp.1–19.

Hendriks, J., Fyfe, S., Styles, I., Rachel Skinner, S. and Merriman, G. 2012. Scale construction utilising the Rasch unidimensional measurement model: A measurement of adolescent attitudes towards abortion. *Australasian Medical Journal*. 5(5), pp.251–261.

Henson, B. and Camargo, F.R. 2014. Verification of Quantitative Properties of Kansei Scales Using The Rasch Model. *Journal of Japan Society of Kansei Engineering*. 13(1), pp.13–20.

Henson, B. and Livesey, R. 2006. Affective Consumer Requirements: A Case Study of Moisturizer Packaging. *Concurrent Engineering*. 14(3), pp.187–196.

Hess, S., Stathopoulos, A. and Daly, A. 2011. Allowing for heterogeneous decision rules in discrete choice models: an approach and four case studies. *Transportation*. 39(3), pp.565–591.

Hinkin, T.R., Tracey, J.B. and Enz, C.A. 1997. Scale construction: Developing reliable and valid measurement instruments. *Journal of Hospitality & Tourism Research*. 21(1), pp.100–120.

Hole, G. 2015. Eight things you need to know about interpreting correlations. . 1, pp.1–9. Available from: http://users.sussex.ac.uk/~grahamh/RM1web/Eight things you need to know about interpreting correlations.pdf.

Horatiu, B. 2010. *Exclusive Interview : Adrian Von Hooydonk, BMW Design Director* [Online]. Available from: http://www.bmwblog.com/2010/01/20/exclusive-interview-adrian-von-hooydonk-bmw-design-director/.

Horton, M. and Perry, A.E. 2016. Screening for depression in primary care: a Rasch analysis of the PHQ-9. *BJPsych Bulletin*. [Online]. 40(5), pp.237–

243. Available from: http://pb.rcpsych.org/cgi/doi/10.1192/pb.bp.114.050294.

Horton, M.C. 2017. *Screening for the risk of self-harm in an adult offender population*.[Online] University of Leeds. Available from: http://etheses.whiterose.ac.uk/id/eprint/20713.

Hosea, M. 2017. How brands are using colour to influence purchase decisions and change perceptionst. , pp.8–11. Available from: http://www.to-market.co.uk/brands-using-colour-influence-purchase-decisions-change-perceptions/.

Hsee, C.K. 1996. The evaluability hypothesis: An explanation for preference reversals between joint and separate evaluations of alternatives. *Organizational Behavior and Human Decision Processes*. 67(3), pp.247–257.

Humphry, S. 2010. Using the methods of pairwise comparison to obtain reliable teacher assessments. *The Australian Education Researcher*. 37(2), pp.1–19.

Humphry, S., Heldsinger, S. and Dawkins, S. 2017. A two-stage assessment method for assessing oral language in early childhood. *Australian Journal of Education*. 61(2), pp.124–140.

ISO 2003. International Organization for Standard - ISO 4142:2003 (en) for Sensory analysis - Guidelines for the use of quantitative response scales. *Online Browsing Platform (OBP)*. [Online]. [Accessed 20 December 2018]. Available from: https://www.iso.org/obp/ui/#iso:std:iso:4121:ed-2:v1:en.

Jansari, A., Tranel, D. and Adolphs, R. 2000. A valence-specific lateral bias for discriminating emotional facial expressions in free field. *Cognition and Emotion*. 14(3), pp.341–353.

Jindo, T. and Hirasago, K. 1997. Application studies to car interior of Kansei engineering. *International Journal of Industrial Ergonomics*. 19(2), pp.105–114.

Jordan, P.W. 2000. *Designing Pleasurable Products : An Introduction to the New Human Factors*. London: Taylor & Francis.

Joseph, L.Z. 1969. Scaling. *Annual Review of Psychology*. [Online]. 20, pp.447–497. Available from: www.annualreviews.org.

Joubert, T., Inceoglu, I., Bartram, D., Dowdeswell, K. and Lin, Y. 2015. A comparison of the psychometric properties of the forced choice and likert scale versions of a personality instrument. *International Journal of Selection and Assessment*. 23(1), pp.92–97.

Juric, B., Vuković, D. and Šuput, B. 2014. The influence of colours on the purchase decision making process when developing sustainable products. *Faculty of Tourism and Hospitality Management in Opatija. Biennial International Congress. Tourism & Hospitality Industry*. [Online], pp.566–583. Available from: http://search.proquest.com/openview/5313e00af66130c8f4c7a00115b53501/1?pq-origsite=gscholar.

Kamijo, K., Tsujimura, H., Obara, H. and Katsumata, M. 1982. Evaluation of seating comfort. *SAE Technical Paper Series*.

Karlsson, B.S.A., Aronsson, N. and Svensson, K.A. 2003. Using semantic environment description as a tool to evaluate car interiors. *Ergonomics*. 46(13/14), pp.1408–1422.

Kersten, P., Küçükdeveci, A.A. and Tennant, A. 2012. The use of the Visual Analo gue Scale (VAS) in rehabilitation outcomes. *Journal of Rehabilitation Medicine*. 44(7), pp.609–610.

Kikuta, M., Ishihara, S., Yanese, T. and Ishihara, K. 2008. Exploring Texture Pattern Features and Relations to Kansei with 2D FFT - Wallpapers and Dashboard Leather Grain Patterns. *Second Conference on Affective Design and Kansei Engineering.* Helsingbor.

Kingdom, F. and Prins, N. 2010. *Psychophysics: A Practical Introduction* First Edit. London: Elsevier Ltd.

Kirchner, E.J.J. and Ravi, J. 2014. Setting tolerances on color and texture for automotive coatings. *Color Research and Application*. 39(1), pp.88–98.

Krosnick, J.A. and Presser, S. 2009. Question and Questionnaire Design P. V Marsden, ed. *Handbook of Survey Research 2nd Edition*.

Kubinger, K.D., Steinfeld, J., Reif, M. and Yanagida, T. 2012. Biased (conditional) parameter estimation of a Rasch model calibrated item pool administered according to a branched testing design. *Psychological Test and Assessment Modeling*. 54(4), pp.450–460.

Larson, M. 1999. Replacing the quality craftsman. *Quality Progress*. 38(Apr), pp.48–51.

Levy, P. 2013. Beyond Kansei engineering: The emancipation of Kansei design. *International Journal of Design*. 7(2), pp.83–94.

Likert, R. and Likert, J.G. 1976. *New ways of managing conflict* . America: McGraw-Hill, Inc.

Linacre, J.M. 2000. Development of Computerized Middle School Achievement Test [in Korean]. *MESA Memorandum No. 69*. (69).

Linacre, J.M. 1994. Sample Size and Item Calibration or Person Measure Stability. *Rasch Measurement Transactions*. [Online]. 7(4), p.328. Available from: https://www.rasch.org/rmt/rmt74m.htm.

Lokman, A.M. 2010a. Design & Emotions: The Kansei Engineering Methodology. *Kansei Engineering International Journal*. 9(1), pp.21–27.

Lokman, A.M. 2010b. Design & Emotions: The Kansei Engineering Methodology. *Kansei Engineering International Journal*. 9(1), pp.21–27.

Lokman, A.M., Haron, M.B.C., Abidin, S.Z.Z., Khalid, N.E.A. and Ishihara, S. 2013. Prelude to Natphoric Kansei Engineering Framework. *Journal of Software Engineering and Applications*. [Online]. 6(12-December), pp.638–644. Available from: http://www.scirp.org/journal/jsea.

Lokman, A.M. and Nagamachi, M. 2010. *Kansei Engineering : A Beginners Perspective*. Selangor Malaysia: University Publications Centre (UPENA), UiTM .

Lorraine M. Uhlaner 2005. The Use of the Guttman Scale in Development of a Family Orientation Index for Small-to-Medium-Sized Firms. *Family Business Review*. [Online]. 18(1), pp.41–56. Available from: http://fbr.sagepub.com/cgi/doi/10.1111/j.1741-6248.2005.00029.x.

Loss, M.L.V. and Jansen, L.P. 2015. Interior Plastic Surface Grain Development - From Design Concept up to Global Vehicle Launch (Texture). *SAE Technical Paper Series*. 36(0316).

Lu, W. and Petiot, J.-F. 2014. Affective design of products using an audio-based protocol: Application to eyeglass frame. *International Journal of Industrial Ergonomics*. 44(3), pp.383–394.

Luce, R.D. 1959. *Individual Choice Behavior* . New York: John Wiley & Sons, Inc.

Manheim, H.L. 1977. *Sociological Research : Philosophy and methods* [Online] (I. 60430 The Dorsey Press, Homewood, ed.). Illinois. Available from: https://books.google.co.uk/books?id=7BhHAQAAIAAJ&q=Manheim,+1977&dq=Manheim,+1977&hl=en&sa=X&ved=0ahUKEwi1xZyHo6PfAhUJ-YUKHeaqA10Q6AEIKDAA.

Maslow, A.H. 1971. *The Farther Reaches of Human Nature*. New York: Penguin Arkana.

Mattacola, C.G., Perrin, D.H., Gansneder, B.M., Allen, J.D. and Mickey, C.A. 1997. A Comparison of Visual Analog and Graphic Rating Scales for Assessing Pain Following Delayed Onset Muscle Soreness. *Journal of Sport Rehabilitation*. 6(1), pp.38–46.

Mike Horton 2013. Item Banking & Computer Adaptive Testing (CAT) - Intermediate Training Course.

Mitsuo Nagamachi 2011. *KANSEI/AFFECTIVE ENGINEERING*. Boca Raton, FL: Taylor and Francis Group, LLC.

Munn, P. and Drever, E. 2004. *Using Questionnaires in Small-Scale Research - A Beginner's Guide* (V. Wilson & J. Lewin, eds.). Glasgow: The SCRE Centre, University of Glasgow.

Myung Hwan Yun, Heecheon You, Wooyeun Geum and Kong, D. 2004. Affective Evaluations of Vehicle Interior Craftmanship _ Systematic checklist for touch feel quality of surface covering material. *Proceedings of the Human Factors and Ergonomics Society 48th Annual Meeting - 2004*., pp.971–975.

Nagamachi, M. 2010. *Kansei/Affective Engineering* First Edit. (C. Press, ed.). Boca Raton, FL: Taylor and Francis Group, LLC.

Nagamachi, M. 2008. Perspectives and the new trend of Kansei /affective engineering. *The TQM Journal*. [Online]. 20(4), pp.290–298. Available from: http://www.emeraldinsight.com/doi/full/10.1108/17542730810881285.

Nissan Motor, C. 2015. Nissan Perceived Quality. *Quality Initiative*. [Online]. 2015. Available from: http://www.nissan-global.com/EN/QUALITY/SENSIBILITY/.

Nissan, R. and Bernhard, W. 2006. IM FOKUS : VOLKSWAGEN Im Fokus : Volkswagen Kernkompetenz : Sparen. , pp.14–15.

Okamoto, S., Nagano, H. and Yamada, Y. 2013. Psychophysical dimensions of tactile perception of textures. *IEEE Transactions on Haptics*. 6(1), pp.81–93.

Osgood, C.E. 1971. Exploration in Semantic Space:A Personal Diary. *Journal of Social Issues*. 27, pp.5–63.

Osgood, C.E., Suci, G.J. and Tannenbaum, P.H. 1957. The measurement of meaning *In*: J. G. Snider and E. O. Charles, eds. *Semantic Differential Technique*. Urbana: University of Illinois Press, pp.318–331.

Özmercan, E.E. and Kumandaş, H. 2016. A Scaling Study by Pair-Wise Comparison Method: Friend Choosing in Adolescents. *Journal of Education and Training Studies*. [Online]. 4(6), pp.155–163. Available from: http://redfame.com/journal/index.php/jets/article/view/1494.

Paek, I. and Wilson, M. 2011. Formulating the Rasch Differential Item Functioning Model Under the Marginal Maximum Likelihood Estimation Context and Its Comparison With Mantel-Haenszel Procedure in Short Test and Small Sample Conditions. *Educational and Psychological Measurement*. 71(6), pp.1023–1046.

Palasek, R.A. and Goodall, S.C. 1990. Consumer Impressions of Plastics in Vehicle Interiors. Available from: http://dx.doi.org/10.4271/900091.

Pallant, J.F. and Tennant, A. 2007. An introduction to the Rasch measurement model: An example using the Hospital Anxiety and Depression Scale (HADS). *British Journal of Clinical Psychology*. 46(1), pp.1–18.

Paulhus, D.L. 1984. Two-component models of socially desirable responding. *Journal of Personality and Social Psychology*. 46(3), pp.598–609.

Petrillo, J., Cano, S.J., McLeod, L.D. and 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*. [Online]. 18(1), pp.25–34. Available from: http://dx.doi.org/10.1016/j.jval.2014.10.005.

Phelps, A.S., Naeger, D.M., Courtier, J.L., Lambert, J.W., Marcovici, P.A., Villanueva-Meyer, J.E., MacKenzie, J.D., Andrew S. Phelps, David M. Naeger, Jesse L. Courtier, Jack W. Lambert, Peter A. Marcovici, and, J.E.V.-M. and Mackenzie., J.D. 2015. Pairwise Comparison Versus Likert Scale for Biomedical Image Assessment. *American Journal of Roentgenology*. 204(1), pp.8–14.

Philips Sedgwick 2012. Pearson's correlation. *BMJ Open*. [Online]. 345(July 2014). Available from: https://0-www-bmj-com.wam.leeds.ac.uk/content/345/bmj.e4483.

Pollitt, A. 2012. The method of Adaptive Comparative Judgement. Assessment in Education: Principles, Policy & Practice. . 19(3), p.281.

Poulton, E.C. 1989. *Bias in Quantifyling Judgements*. East Sussex, United Kingdom: Lawrence Erlbaum Associates.

Purya Baghaei 2008. Local Dependency and Rasch Measurement. *Rasch Measurement Transactions*. [Online]. 21(3), pp.1105–1106. Available from: https://www.rasch.org/rmt/rmt213b.htm.

Rasch, G. 1980. *Probabilistic model for some intelligence and attainment test* (1960. Originally published: Danish Institute for Educational Research, ed.). Chicago: University of Chicago Press.

Rosenthal, S.R. and Capper, M. 2006. Rosenthal, S. R., & Capper, M. (2006). Ethnographies in the Front End: Designing for Enhanced Customer

Experiences. *Journal of Product Innovation Managemen*., p.1312155.

Sabin, C.A. 2010. *What is bias and how can it affect the outcomes from research ?* London: University College London.

Salzberger, T. 2015. The validity of polytomous items in the Rasch model – The role of statistical evidence of the threshold order. *Psychological Test and Assessment Modeling*. [Online]. 57(3), pp.377–395. Available from: http://www.psychologie-aktuell.com/fileadmin/download/ptam/3-2015_20150925/05_Salzberger.pdf.

Sarle, W.S. 1995. Measurement theory : Frequently asked questions. *Disseminations of the International Statistical Aplications Institute*. 1(4), pp.61–66.

Schneider, B. 1980. Individual loudness functions determined from direct comparison of loudness intervals. *Perception and Psychophysics*. 28(6), pp.493–503.

Schutte, S. 2005. *Engineering Emotional Values in Product Design - Kansei Engineering in Development*. Linkoping Sweden: Linkoping Universitet Insititute of Technology.

Schutte, S. and Eklund, J.A.E. 2010. Ratingscales in Kansei Engineering : Modification for an European Context. *International Conference on Kansei Engineering and Emotion Research*. [Online]. (February). Available from: https://www.researchgate.net/publication/279640757%0D.

Schütte, S., Schütte, R. and Eklund, J. 2005. Affective Values of Lift Trucks: an Application of Kansei Engineering *In*: *QMOD Conference 2005*. Palermo, Sicilien.

Snider, J.G. and Osgood, C.E. 1969. *Semantic Differential Technique*. Chicago: Aldine Publishing Company.

Statistical Society, R. 2007. RSS Ordinary Certificate in Statistics Module 1 : Research Methods , Data Collection Methods and Questionnaire Design. *Office for National Statistics*., pp.39–43.

Stevens, S.S. 1946. On the Theory of Scales of Measurement. *Science*. 103, pp.677–680.

Suhonen, R., Schmidt, L.A., Katajisto, J., Berg, A., Idvall, E., Kalafati, M., Land, L., Lemonidou, C., Välimäki, M. and Leino-Kilpi, H. 2013. Cross-cultural validity of the Individualised Care Scale - A Rasch model analysis. *Journal of Clinical Nursing*. 22(5–6), pp.648–660.

Sung, Y.T. and Wu, J.S. 2018. The Visual Analogue Scale for Rating, Ranking and Paired-Comparison (VAS-RRP): A new technique for psychological measurement. *Behavior Research Methods*. 50(4), pp.1694–1715.

Talikoti, M. 2016. Advantages and disadvantages of Semantic Differential scales. *QuestionPro*. [Online]. Available from: https://www.questionpro.com/blog/advantages-and-disadvantages-of-semantic-differential-scale/?utm_content=34129687&utm_medium=social&utm_source=googleplus.

Tennant, A., Conaghan, P.G., Tennnant, A. and Gonaghan, P.G. 2007. The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *American college of Rhematology*. [Online]. 57(8), pp.1358–1362. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18050173.

Thurstone, L.. 1987. Psychophysical Analysis. *The American Journal of Psychology*. [Online]. 100(3/4), pp.587–609. Available from: https://www.jstor.org/stable/1422696.

Thurstone, L.L. 1927. Psychophysics Analysis. *The American Journal of Psychology*. 100(3/4), pp.587–609.

Train, K.E. 2009. *Discrete Choice Methods With Simulation* Second Edi. Cambridge: Cambridge University Press 2009.

Velozo, C.A., Wang, Y., Lehman, L. and Wang, J.-H. 2008. Utilizing Rasch measurement models to develop a computer adaptive self-report of walking, climbing, and running. *Disability and rehabilitation*. 30(6), pp.458–467.

Warell, A. 2001. *Design Syntactics : A functional Approach to Visual Products*

Form . Gothenburg : Chalmers University of Technology.

Wellings, T., Williams, M. and Tennant, C. 2005. Tactility, Craftsmanship and the NPI Process. Available from: http://dx.doi.org/10.4271/2005-01-0977.

Westerman, S.J., Sutherland, E.J., Gardner, P.H., Metcalfe, R., Nash, J., Palframan, S. and Woodburn, N. 2012. Ecommerce interface colour and consumer decision making: Two routes of influence. *Color Research and Application*. 37(4), pp.292–301.

Wilson, T.D. and Gilbert, D.T. 2005. Affective Forecasting Knowing What to want. *American Psycological Society*. 14(3), pp.131–134.

Winstanley, J.B., Saw, R., Boyle, F. and Thompson, J. 2013. The FACT-Melanoma Quality-of-life Instrument : comparison of five-point and four-point scale using the Rasch Measurement model. *Melanoma Research*. 23(1), pp.61–69.

Winter, D. 2003. Shrinking Product Development Time. *Wards Auto*. [Online]. Available from: https://www.wardsauto.com/news-analysis/shrinking-product-development-time.

Wright, B.D. and Linacre, J.M. 1989. Observations are always ordinal; measurements, however, must be interval. *Arch Phys Med Rehabil*. [Online]. 70(12), pp.857–860. Available from: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=2818162.

Wright, B.D., Mead, R. and Draba, R. 1976. Detecting and Correcting Test Item Bias with a Logistic Response Model. *MESA Memorandum No. 69*. (22).

Wright, B.D. and Stone, M.H. 1999. MEASUREMENT 2nd Edition. , p.205.

Appendix A - Subtest 1

Table 8.3.1 Summary Statistics of Non-Facet Analysis for Texture 1

	Location				Fit Residuals				Item-trait interaction						
	Item		Persons		Item		Persons		Chi square				RMSEA		
Analysis	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Value**	**df**	**p**	**Bonferroni's alpha**	**N**	**Value**	**Interpr**
Initial Data	0.000	0.256	-0.122	1.131	0.742	2.059	-0.305	1.855	71.4	40	0.002	0.0025	145	0.074	Inadequate fit
Calibrated Data 1	0.000	0.325	-0.174	1.336	0.328	1.360	-0.126	1.453	50.6	40	0.122	0.0025	144	0.043	Inadequate fit
Calibrated Data 2	0.000	0.326	-0.151	1.354	0.360	1.381	-0.387	1.740	51.3	40	0.010	0.0025	142	0.045	Inadequate fit
Calibrated Data 3	0.000	0.334	-0.139	1.403	0.481	1.760	-0.353	1.788	38.6	38	0.443	0.0025	142	0.010	Good fit

Reliability

	PSI		Alpha		
Analysis	**With extrms**	**No extrms**	**With extrms**	**No extrms**	**Interpretation**
Initial Data	0.940	0.935	0.936	0.932	Single person measurement
Calibrated Data 1	0.939	0.934	0.934	0.930	Single person measurement
Calibrated Data 2	0.935	0.935	0.930	0.930	Single person measurement
Calibrated Data 3	0.935	0.935	0.929	0.929	Single person measurement

Unidimensionality

	Paired t-tests				Binomial Test	
N Significant tests	Sample	% PST	%LB95CI	Is	P-expected Proportion	0.05
46 (Initial)	145	31.70%	28.20%	Not acceptable	Sample Size	142
40 (Final Calibrations)	142	28.20%	24.60%	Not acceptable	Observed proportion	40
						0.28169
					Lower 95% CI - Proportion	0.246

Table 8.3.2 Unidimensionality Test on each item performance of Non-Facet Analysis – Texture 1

Local Dependency				Item Fit				Unidimensionality						Thresholds info			
Item	Item Hold Dependency	Item hold Independent		Item Hold High Fit Res		Most Difficult Item		High PC1 Loading		Positive Load		Negative Load		Item Hold Disordered?	Item	Max Score	Rescored
1	6,7,19	1	0	1	2.343	18	0.797	6	0.646	6	0.646	15	-0.092	Yes	1	3	Yes
2	11,14,20	2	0	12	2.245	19	0.534	7	0.633	7	0.633	9	-0.196	Yes	3	3	Yes
3	11,12,18	3	0	4	1.95	7	0.31	19	0.584	19	0.584	13	-0.289	Yes	4	3	Yes
4	18	4	0	7	1.705	14	0.252	16	0.534	16	0.534	4	-0.37	Yes	5	3	Yes
5	6,7,8,17,19	5	0	9	1.643	5	0.244	17	0.507	17	0.507	12	-0.396	Yes	6	3	Yes
6	7,8,16,17	9	0	17	1.176	4	0.125	5	0.481	5	0.481	2	-0.423	Yes	11	3	Yes
7	8,16,17,19	13	0	19	0.984	2	0.093	8	0.474	8	0.474	3	-0.426	Yes	12	3	Yes
8	16	15	0	16	0.679	15	0.077	1	0.471	1	0.471	20	-0.427	Yes	18	3	Yes
9		6	2	6	0.314	3	0.023	15	-0.092			18	-0.497	Yes	19	3	Yes
10	Item Deleted	11	2	2	0.193	12	0.021	9	-0.196			14	-0.559	No	2	4	No
11	12,14,18,20	12	2	20	-0.123	20	-0.097	13	-0.289			11	-0.618	No	7	4	No
12		8	3	15	-0.198	8	-0.143	4	-0.37					No	8	4	No
13	14	7	3	8	-0.281	11	-0.191	12	-0.396					No	9	4	No
14	18,20	14	3	14	-0.428	17	-0.207	2	-0.423					No	13	4	No
15		16	3	5	-0.523	6	-0.26	3	-0.426					No	14	4	No
16	17, 20	20	3	3	-0.809	13	-0.275	20	-0.427					No	15	4	No
17	19	18	4	13	-1.223	9	-0.375	18	-0.497					No	16	4	No
18		17	4	11	-1.426	16	-0.394	14	-0.559					No	17	4	No
19		19	5	18	-1.809	1	-0.535	11	-0.618					No	20	4	No
20																	
		Item hold Dependency		Item Hold Low Fit Res		Most Easy Item		Low PC1 Loading						Disordered Threshold Count : 9 Ordered Threshold Count : 10			
				Fit-Res Order		Loc Order		Multidimensional									

Table 8.3.3 Local Dependencies Matrix in Texture 1 – Calibrated

Item	1	2	3	4	5	6	7	8	9	11	12	13	14	15	16	17	18	19	20
1																			
2	-0.313																		
3	-0.138	0.099																	
4	-0.118	0.14	0.041																
5	0.137	-0.233	-0.252	-0.207															
6	0.347	-0.214	-0.214	-0.225	0.284														
7	0.233	-0.217	-0.317	-0.136	0.19	0.239													
8	0.09	-0.095	-0.171	-0.248	0.173	0.335	0.221												
9	-0.115	-0.043	0.092	-0.096	-0.308	-0.211	-0.18	-0.048											
11	-0.253	0.168	0.278	0.124	-0.173	-0.354	-0.443	-0.224	0.119										
12	-0.16	0.011	0.16	0.043	-0.104	-0.138	-0.273	-0.218	0.013	0.219									
13	-0.219	0.033	0.015	0.075	-0.036	-0.219	-0.152	-0.22	-0.091	0.091	-0.127								
14	-0.219	0.216	0.016	0.099	-0.24	-0.472	-0.368	-0.323	0.001	0.198	0.112	0.171							
15	-0.125	-0.233	-0.005	-0.136	0.014	-0.097	-0.112	-0.153	0.009	0.023	0.092	-0.028	-0.003						
16	0.073	-0.246	-0.235	-0.266	0.107	0.203	0.197	0.216	0.04	-0.296	-0.354	-0.159	-0.194	-0.091					
17	0.057	-0.177	-0.223	-0.17	0.208	0.298	0.191	0.066	-0.15	-0.246	-0.177	-0.203	-0.293	-0.099	0.158				
18	-0.154	0.097	0.299	0.199	-0.208	-0.351	-0.274	-0.191	-0.05	0.235	0.127	0.136	0.15	0	-0.249	-0.328			
19	0.229	-0.252	-0.205	-0.237	0.261	0.114	0.394	0.101	-0.254	-0.453	-0.298	-0.08	-0.197	-0.127	0.325	0.211	-0.185		
20	-0.269	0.215	-0.083	0.073	-0.187	-0.227	-0.231	-0.178	-0.113	0.156	0.076	0.13	0.315	-0.041	-0.322	-0.257	0.032	-0.161	

Remarks : The highlighted item is associate with problematic dependencies in which the value was greater than cut-off value as below;
Average : 0.015 Cut-off : 0.184

Table 8.3.4 Summary Statistics of Non-Facet Analysis for Texture 2

	Location				**Fit Residuals**				**Item-trait interaction**						
	Item		**Persons**		**Item**		**Persons**		**Chi square**				**RMSEA**		
Analysis	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Value**	**df**	**p**	**Bonferroni's alpha**	**N**	**Value**	**Interpr**
Initial Data	0.000	0.247	0.803	0.989	-0.072	1.612	-0.577	1.968	97.4	40	0.001	0.0025	145	97.396	Inadequate fit
Calibrated Data 1	0.000	0.304	0.907	1.077	0.018	1.526	-0.591	1.988	99.4	40	0.001	0.0025	145	0.102	Inadequate fit
Calibrated Data 2	0.000	0.278	0.891	1.030	-0.040	1.096	-0.630	1.937	63.7	36	0.004	0.0025	144	0.073	Inadequate fit
Calibrated Data 3	0.000	0.328	1.000	1.059	0.071	1.140	-0.432	1.676	63.7	36	0.004	0.0025	136	0.075	Inadequate fit

Reliability

	PSI		**Alpha**		
Analysis	**With extrms**	**No extrms**	**With extrms**	**No extrms**	**Interpretation**
Initial Data	0.940	0.935	0.936	0.932	Single person measurement
Calibrated Data 1	0.939	0.934	0.934	0.930	Single person measurement
Calibrated Data 2	0.935	0.935	0.930	0.930	Single person measurement
Calibrated Data 3	0.935	0.935	0.929	0.929	Single person measurement

Unidimensionality

	Paired t-tests				Binomial Test	
N Significant tests	Sample	% PST	%LB95CI	Is	P-expected Proportion	0.05
38 (Initial)	145	26.20%	22.70%	Not acceptable	Sample Size	136
31 (Final Calibrations)	136	22.80%	19.10%	Not acceptable	Observed proportion	31
						0.227941
					Lower 95% CI - Proportion	0.191

Table 8.3.5 Unidimensionality Test on each item performance of Non-Facet Analysis – Texture 2

Local Dependency				Item Fit				Unidimensionality						Thresholds info			
Item	Item Hold Dependency	Item hold Independent		Item Hold High Fit Res		Most Difficult Item		High PC1 Loading		Positive Load		Negative Load		Item Hold Disordered?	Item	Max Score	Rescored
1	5,6,16,17,19	1	0	7	1.921	3	0.563	5	0.587	5	0.587	15	-0.098	Yes	1	3	Yes
2	3,4,12,13,20	2	0	19	1.751	15	0.533	1	0.583	1	0.583	4	-0.341	Yes	3	3	Yes
3	4,9,12,13,20	7	0	17	1.356	18	0.424	19	0.579	19	0.579	14	-0.368	Yes	4	3	Yes
4		8	0	4	1.117	14	0.253	17	0.554	17	0.554	12	-0.413	Yes	5	3	Yes
5	6,17,19	15	0	6	1.089	19	0.157	6	0.549	6	0.549	9	-0.453	Yes	6	3	Yes
6		3	1	1	0.812	16	0.137	7	0.414	7	0.414	2	-0.497	Yes	9	3	Yes
7	19	5	1	15	0.648	6	0.033	16	0.366	16	0.366	20	-0.52	Yes	12	3	Yes
8		9	1	5	0.593	12	0.031	8	0.245	8	0.245	3	-0.559	Yes	13	3	Yes
9	13,18,20	14	1	16	0.559	5	0.015	15	-0.098			18	-0.565	Yes	20	3	Yes
10	Item Deleted	16	1	18	-0.422	9	-0.051	4	-0.341			13	-0.572	No	2	4	No
11	Item Deleted	4	2	9	-0.474	2	-0.062	14	-0.368					No	7	4	No
12	18	6	2	13	-0.594	7	-0.088	12	-0.413					No	8	4	No
13	14,18,20	12	2	3	-0.669	8	-0.15	9	-0.453					No	14	4	No
14	18	13	3	8	-1.059	4	-0.184	2	-0.497					No	15	4	No
15		17	3	20	-1.148	17	-0.206	20	-0.52					No	16	4	No
16	17	18	3	14	-1.288	1	-0.217	3	-0.559					No	17	4	No
17		19	3	2	-1.32	20	-0.399	18	-0.565					No	18	4	No
18	20	20	5	12	-1.588	13	-0.789	13	-0.572					No	19	4	No
19																	
20																	
		Item hold Dependency		Item Hold Low Fit Res		Most Easy Item		Low PC1 Loading						Disordered Threshold Count : 9 Ordered Threshold Count : 9			
				Fit-Res Order		Loc Order		Multidimensional									

Table 8.3.6 Local Dependencies Matrix in Texture 2

Item	1	2	3	4	5	6	7	8	9	12	13	14	15	16	17	18	19	20
1																		
2	-0.304																	
3	-0.272	0.181																
4	-0.236	0.166	0.21															
5	0.248	-0.379	-0.266	-0.223														
6	0.335	-0.179	-0.406	-0.157	0.444													
7	-0.011	-0.137	-0.322	-0.037	0.002	0.082												
8	0.071	-0.254	-0.109	-0.062	0.138	-0.004	0.023											
9	-0.164	0.113	0.202	0.107	-0.226	-0.18	-0.255	-0.098										
12	-0.208	0.176	0.277	0.104	-0.281	-0.087	-0.202	-0.008	0.04									
13	-0.393	0.222	0.17	0.123	-0.275	-0.33	-0.14	-0.189	0.222	-0.021								
14	-0.19	-0.008	0.124	-0.105	-0.206	-0.277	-0.294	-0.085	0.005	0.031	0.268							
15	-0.017	-0.121	-0.034	-0.179	0.052	-0.192	-0.157	-0.016	0.032	-0.158	0.054	0.115						
16	0.171	-0.196	-0.131	-0.229	-0.01	0.069	0.075	-0.077	-0.211	-0.152	-0.297	-0.077	-0.232					
17	0.262	-0.323	-0.242	-0.21	0.195	0.125	0.11	0.056	-0.301	-0.268	-0.234	-0.076	-0.224	0.307				
18	-0.363	0.143	0.157	-0.005	-0.295	-0.299	-0.235	-0.242	0.187	0.241	0.221	0.168	-0.002	-0.166	-0.303			
19	0.187	-0.126	-0.33	-0.183	0.225	0.123	0.359	0.04	-0.265	-0.329	-0.272	-0.344	-0.073	0.066	0.143	-0.327		
20	-0.211	0.246	0.136	0.126	-0.272	-0.162	-0.288	-0.139	0.18	0.099	0.26	0.084	0.036	-0.122	-0.314	0.2	-0.366	

Remarks : The highlighted item is associate with problematic dependencies in which the value was greater than cut-off value as below;

Average : -0.056

Cut-off : 0.143

Table 8.3.7 Summary Statistics of Non-Facet Analysis for Texture 3

	Location				Fit Residuals				Item-trait interaction						
	Item		Persons		Item		Persons		Chi square				RMSEA		
Analysis	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Value**	**df**	**p**	**Bonferroni's alpha**	**N**	**Value**	**Interpr**
Initial Data	0.000	0.321	0.425	0.971	0.260	1.204	-0.544	2.065	80.2	40	0.001	0.0025	144	0.084	Inadequate fit
Calibrated Data 1	0.000	0.383	0.465	1.135	0.167	1.067	-0.621	2.170	69.4	40	0.002	0.0025	144	0.072	Inadequate fit
Calibrated Data 2	0.000	0.393	0.458	1.181	0.124	0.973	-0.618	2.074	64.2	38	0.005	0.0025	144	0.069	Inadequate fit
Calibrated Data 3	0.000	0.426	0.509	1.142	0.183	1.031	-0.422	1.720	66.0	38	0.003	0.025	136	0.074	Inadequate fit

Reliability

	PSI		Alpha		
Analysis	**With extrms**	**No extrms**	**With extrms**	**No extrms**	**Interpretation**
Initial Data	0.922	0.917	0.926	0.922	Single person measurement
Calibrated Data 1	0.919	0.914	0.920	0.916	Single person measurement
Calibrated Data 2	0.919	0.914	0.920	0.916	Single person measurement
Calibrated Data 3	0.909	0.909	0.910	0.910	Single person measurement

Unidimensionality

	Paired t-tests				Binomial Test	
N Significant tests	Sample	% PST	%LB95CI	Is	P-expected Proportion	0.05
46 (Initial)	145	31.70%	28..20%	Not acceptable	Sample Size	136
37 (Final Calibrations)	136	27.20%	23.50%	Not acceptable	Observed proportion	37
						0.272059
					Lower 95% CI - Proportion	0.235

Table 8.3.8 Unidimensionality Test on each item performance of Non-Facet Analysis – Texture 3

Local Dependency				Item Fit				Unidimensionality						Thresholds info			
Item	Item Hold Dependency	Item hold Independent		Item Hold High Fit Res		Most Difficult Item		High PC1 Loading		Positive Load		Negative Load		Item Hold Disordered?	Item	Max Score	Rescored
1	5,6,17,19	1	0	6	1.774	18	0.61	19	0.66	19	0.66	15	-0.051	Yes	1	3	Yes
2	10,11	2	0	5	1.66	15	0.546	17	0.655	17	0.655	4	-0.178	Yes	3	3	Yes
3	11,12,18	3	0	19	1.493	12	0.501	6	0.61	6	0.61	9	-0.185	Yes	4	3	Yes
4	14	4	0	1	1.231	14	0.482	5	0.599	5	0.599	10	-0.255	Yes	6	3	Yes
5	6,17,19	8	0	15	1.168	9	0.4	16	0.537	16	0.537	2	-0.277	Yes	8	3	Yes
6	16,17,19	9	0	2	0.928	2	0.371	1	0.425	1	0.425	20	-0.338	Yes	9	3	Yes
7	Item Deleted	15	0	17	0.648	10	0.29	8	0.323	8	0.323	13	-0.378	Yes	10	3	Yes
8	16,19	20	0	16	0.506	3	0.207	15	-0.051			18	-0.422	Yes	11	3	Yes
9		5	1	9	0.196	11	0.206	4	-0.178			3	-0.496	Yes	12	3	Yes
10		10	1	4	0.022	13	-0.045	9	-0.185			14	-0.543	Yes	14	3	Yes
11	12,14,18	13	1	10	-0.057	19	-0.153	10	-0.255			11	-0.6	Yes	18	3	Yes
12	13,14,18	6	2	12	-0.177	6	-0.236	2	-0.277			12	-0.604	Yes	19	3	Yes
13	14	11	2	14	-0.188	16	-0.242	20	-0.338					No	2	4	No
14		12	2	3	-0.303	5	-0.379	13	-0.378					No	5	4	No
15		16	2	13	-0.476	4	-0.385	18	-0.422					No	13	4	No
16	17,19	18	3	20	-0.69	17	-0.453	3	-0.496					No	15	4	No
17	19	14	4	8	-1.05	1	-0.46	14	-0.543					No	16	4	No
18		17	4	18	-1.498	20	-0.581	11	-0.6					No	17	4	No
19		19	6	11	-1.719	8	-0.679	12	-0.604					No	20	4	No
20																	
		Item hold Dependency		Item Hold Low Fit Res		Most Easy Item		Low PC1 Loading						Disordered Threshold Count : 12 Ordered Threshold Count : 7			
				Fit-Res Order		Loc Order		Multidimensional									

Table 8.3.9 Local Dependencies Matrix in Texture 3

Item	1	2	3	4	5	6	8	9	10	11	12	13	14	15	16	17	18	19	20
1																			
2	-0.221																		
3	-0.214	-0.042																	
4	-0.067	-0.01	-0.032																
5	0.304	-0.387	-0.069	-0.2															
6	0.167	-0.064	-0.305	-0.104	0.278														
8	-0.092	-0.022	-0.287	0.015	0.087	0.061													
9	-0.03	0.1	0.094	-0.005	-0.096	-0.305	-0.201												
10	-0.192	0.156	0.103	0.061	-0.099	-0.057	-0.198	0.058											
11	-0.161	0.177	0.253	0.022	-0.474	-0.261	-0.099	-0.086	0.118										
12	-0.259	-0.081	0.299	-0.037	-0.275	-0.442	-0.26	-0.022	-0.009	0.356									
13	-0.177	-0.137	0.002	0.097	-0.211	-0.283	-0.023	-0.016	-0.129	0.128	0.204								
14	-0.253	0.075	0.051	0.153	-0.342	-0.254	-0.14	-0.022	-0.044	0.222	0.23	0.318							
15	-0.187	0.036	-0.092	-0.175	-0.169	-0.174	-0.005	-0.024	-0.142	-0.082	0.054	0.106	0.034						
16	0.023	-0.159	-0.327	-0.17	0.071	0.227	0.289	-0.154	-0.227	-0.194	-0.249	-0.212	-0.246	0.074					
17	0.158	-0.292	-0.297	-0.144	0.369	0.35	0.02	-0.124	-0.139	-0.355	-0.249	-0.284	-0.356	-0.188	0.268				
18	-0.122	-0.057	0.354	-0.094	-0.235	-0.233	-0.176	-0.063	0.045	0.195	0.369	-0.012	0.133	-0.126	-0.271	-0.164			
19	0.208	-0.189	-0.285	-0.098	0.198	0.253	0.231	-0.138	-0.298	-0.417	-0.344	-0.23	-0.363	0.072	0.31	0.333	-0.181		
20	-0.141	0.058	0.077	0.099	-0.191	-0.211	-0.043	0.052	0.048	0.114	0.062	0.04	0.138	-0.142	-0.219	-0.265	-0.028	-0.245	

Remarks : The highlighted item is associate with problematic dependencies in which the value was greater than cut-off value as below;.

Average : -0.053

Cut-off : 0.146

Table 8.3.10 Summary Statistics of Non-Facet Analysis for Texture 4

	Location				Fit Residuals				Item-trait interaction							
	Item		Persons		Item		Persons		Chi square					RMSEA		
Analysis	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Value**	**df**	**p**	**Bonferroni's alpha**	**N**	**Value**	**Interpr**
Initial Data	0.000	0.332	0.522	1.130	0.365	1.520	-0.376	1.849	67.3	40	0.004	0.0025	145	0.069	Inadequate fit
Calibrated Data 1	0.000	0.364	0.523	1.153	0.227	1.283	-0.433	1.840	41.8	40	0.390	0.0025	144	0.018	good fit
Calibrated Data 2	0.000	0.364	0.523	1.153	0.227	1.283	-0.433	1.840	41.8	40	0.390	0.0025	144	0.018	good fit
Calibrated Data 3	0.000	0.362	0.542	1.174	0.278	1.279	-0.356	1.709	41.8	40	0.390	0.0025	141	0.018	good fit

Reliability

	PSI		Alpha		
Analysis	**With extrms**	**No extrms**	**With extrms**	**No extrms**	**Interpretation**
Initial Data	0.928	0.926	0.928	0.925	Single person measurement
Calibrated Data 1	0.928	0.926	0.928	0.925	Single person measurement
Calibrated Data 2	0.926	0.926	0.925	0.925	Single person measurement
Calibrated Data 3	0.927	0.927	0.927	0.927	Single person measurement

Unidimensionality

	Paired t-tests				Binomial Test	
N Significant tests	Sample	% PST	%LB95CI	Is	P-expected Proportion	0.05
43 (Initial)	145	29.70%	26.10%	Not acceptable	Sample Size	141
39 (Final Calibrations)	141	27.70%	24.10%	Not acceptable	Observed proportion	39
						0.279596
					Lower 95% CI - Proportion	0.241

Table 8.3.11 Unidimensionality Test on each item performance of Non-Facet Analysis – Texture 4

Local Dependency				Item Fit				Unidimensionality						Thresholds info			
Item	Item Hold Dependency	Item hold Independent		Item Hold High Fit Res		Most Difficult Item		High PC1 Loading		Positive Load		Negative Load		Item Hold Disordered?	Item	Max Score	Rescored
1	5,6,17,18	1	0	10	2.104	18	0.793	5	0.687	5	0.687	10	-0.1	Yes	1	3	Yes
2	14	2	0	1	1.961	15	0.564	6	0.601	6	0.601	15	-0.135	Yes	3	3	Yes
3	10,11,12	3	0	2	1.743	2	0.472	19	0.6	19	0.6	4	-0.165	Yes	6	3	Yes
4		4	0	7	1.563	14	0.404	7	0.551	7	0.551	20	-0.342	Yes	9	3	Yes
5	6,7,8,16,17,19	9	0	6	1.421	12	0.382	8	0.48	8	0.48	9	-0.354	Yes	10	3	Yes
6	8,16,17,19	13	0	15	1.18	11	0.048	17	0.471	17	0.471	13	-0.362	Yes	12	3	Yes
7	8,16,19	15	0	19	0.898	9	0.041	16	0.468	16	0.468	2	-0.371	Yes	15	3	Yes
8	16	5	1	4	0.756	20	-0.017	1	0.418	1	0.418	3	-0.415	Yes	19	3	Yes
9	18	7	1	16	0.704	13	-0.023	10	-0.1			11	-0.493	No	2	4	No
10		10	1	3	0.676	4	-0.052	15	-0.135			14	-0.503	No	4	4	No
11	12,14,18	11	1	5	0.591	5	-0.064	4	-0.165			12	-0.515	No	5	4	No
12	18	14	1	8	0.335	10	-0.088	20	-0.342			18	-0.527	No	7	4	No
13	18,20	20	1	14	-0.242	3	-0.145	9	-0.354					No	8	4	No
14		6	2	17	-0.276	7	-0.148	13	-0.362					No	11	4	No
15		12	2	13	-0.347	19	-0.182	2	-0.371					No	13	4	No
16	17,19	8	3	11	-1.051	8	-0.249	3	-0.415					No	14	4	No
17		17	4	20	-1.324	16	-0.292	11	-0.493					No	16	4	No
18		16	4	9	-1.347	6	-0.326	14	-0.503					No	17	4	No
19		19	4	12	-1.375	17	-0.424	12	-0.515					No	18	4	No
20		18	5	18	-2.42	1	-0.692	18	-0.527					No	20	4	No
		Item hold Dependency		Item Hold Low Fit Res		Most Easy Item		Low PC1 Loading						Disordered Threshold Count : 8 Ordered Threshold Count : 12			
				Fit-Res Order		Loc Order		Multidimensional									

Table 8.3.12 Local Dependencies Matrix in Texture 4

Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1																				
2	-0.219																			
3	-0.203	0.11																		
4	-0.057	0.035	-0.081																	
5	0.292	-0.218	-0.371	-0.116																
6	0.265	-0.24	-0.233	-0.058	0.293															
7	-0.006	-0.136	-0.167	-0.127	0.317	0.179														
8	-0.069	-0.087	-0.209	-0.265	0.308	0.224	0.22													
9	-0.008	0.017	0.073	-0.001	-0.343	-0.165	-0.214	-0.086												
10	0.047	-0.154	0.444	-0.188	-0.099	-0.194	-0.057	-0.183	0.06											
11	-0.26	-0.013	0.218	0.054	-0.306	-0.37	-0.281	-0.218	-0.01	0.095										
12	-0.149	0.043	0.191	0.052	-0.26	-0.274	-0.241	-0.234	0.099	-0.002	0.367									
13	-0.108	0.052	-0.011	0.058	-0.291	-0.201	-0.207	-0.214	0.077	-0.228	-0.012	0.063								
14	-0.296	0.234	-0.052	0.011	-0.256	-0.276	-0.238	-0.225	0.031	-0.183	0.116	0.068	0.262							
15	0.014	-0.082	-0.051	-0.056	-0.12	-0.074	-0.253	-0.082	0.01	-0.057	0.011	-0.093	0.121	0.078						
16	0.018	-0.294	-0.19	-0.154	0.145	0.122	0.248	0.244	-0.172	-0.088	-0.203	-0.203	-0.165	-0.225	-0.132					
17	0.161	-0.236	-0.16	-0.115	0.196	0.325	0.113	0.057	-0.259	-0.027	-0.147	-0.275	-0.106	-0.21	-0.008	0.113				
18	-0.305	0.057	0.03	0.012	-0.337	-0.274	-0.316	-0.271	0.166	-0.098	0.147	0.241	0.119	0.283	0.097	-0.104	-0.223			
19	0.164	-0.272	-0.289	-0.038	0.307	0.196	0.275	0.162	-0.24	-0.077	-0.229	-0.365	-0.194	-0.341	-0.076	0.308	0.15	-0.247		
20	-0.21	0.083	0.006	-0.026	-0.235	-0.194	-0.221	-0.105	0.121	-0.001	0.031	0.064	0.104	0.146	-0.055	-0.155	-0.282	0.095	-0.081	

Remarks : The highlighted item is associate with problematic dependencies in which the value was greater than cut-off value as below;

Average : -0.051

Cut-off : 0.148

Table 8.3.13 Summary Statistics of Non-Facet Analysis for Texture 5

	Location				**Fit Residuals**				**Item-trait interaction**						
	Item		**Persons**		**Item**		**Persons**		**Chi square**				**RMSEA**		
Analysis	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Value**	**df**	**p**	**Bonferroni's alpha**	**N**	**Value**	**Interpr**
Initial Data	0.000	0.242	0.632	1.059	0.433	1.579	-0.479	2.013	78.9	40	0.001	0.0025	145	0.082	Inadequate fit
Calibrated Data 1	0.000	0.279	0.672	1.166	0.368	1.637	-0.043	1.987	88.2	40	0.001	0.0025	145	0.091	Inadequate fit
Calibrated Data 2	0.000	0.272	0.735	1.245	0.234	1.622	-0.440	1.854	89.2	36	0.001	0.0025	145	0.101	Inadequate fit
Calibrated Data 3	0.000	0.264	0.720	1.215	0.320	1.598	-0.324	1.616	93.1	36	0.001	0.0025	139	0.107	Inadequate fit
Calibrated Data 4	0.000	0.268	0.753	1.274	0.323	1.638	-0.313	1.571	79.5	34	0.001	0.0025	139	0.098	Inadequate fit
Calibrated Data 5	0.000	0.276	0.782	1.334	0.286	1.526	-0.340	1.563	65.7	32	0.001	0.0025	139	0.087	Inadequate fit
Calibrated Data 6	0.000	0.273	0.823	1.389	0.290	1.662	-0.300	1.479	63.4	30	0.001	0.0025	139	0.090	Inadequate fit
Calibrated Data 7	0.000	0.313	0.868	1.407	0.274	1.774	-0.257	1.367	58.3	26	0.001	0.0025	137	0.096	Inadequate fit

Reliability

	PSI		**Alpha**		
Analysis	**With extrms**	**No extrms**	**With extrms**	**No extrms**	**Interpretation**
Initial Data	0.923	0.923	0.932	0.915	Single person measurement
Calibrated Data 1	0.924	0.924	0.927	0.927	Single person measurement
Calibrated Data 2	0.921	0.921	0.924	0.924	Single person measurement
Calibrated Data 3	0.920	0.920	0.922	0.922	Single person measurement
Calibrated Data 4	0.921	0.921	0.923	0.923	Single person measurement
Calibrated Data 5	0.919	0.919	0.922	0.922	Single person measurement
Calibrated Data 6	0.919	0.919	0.922	0.922	Single person measurement
Calibrated Data 7	0.915	0.906	0.919	0.909	Single person measurement

Unidimensionality

	Paired t-tests				Binomial Test	
N Significant tests	Sample	% PST	%LB95CI	Is	P-expected Proportion	0.05
39 (Initial)	145	26.90%	23.30%	Not acceptable	Sample Size	141
28 (Final Calibrations)	137	20.40%	16.80%	Not acceptable	Observed proportion	39
						0.20438
					Lower 95% CI - Proportion	0.168

Table 8.3.14 Unidimensionality Test on each item performance of Non-Facet Analysis – Texture 5

Local Dependency				Item Fit				Unidimensionality						Thresholds info			
Item	Item Hold Dependency	Item hold Independent		Item Hold High Fit Res		Most Difficult Item		High PC1 Loading		Positive Load		Negative Load		Item Hold Disordered?	Item	Max Score	Rescored
1	Item Deleted	2	0	19	3.762	9	0.488	7	0.795	7	0.795	4	-0.119	Yes	3	3	Yes
2	3	4	0	7	3.461	18	0.367	19	0.722	19	0.722	14	-0.122	Yes	4	3	Yes
3	9,12	7	0	2	1.556	19	0.304	8	0.493	8	0.493	13	-0.164	Yes	9	3	Yes
4		11	0	14	1.011	3	0.265	20	0.195	20	0.195	2	-0.258	Yes	11	3	Yes
5	Item Deleted	13	0	9	0.66	2	0.155	4	-0.119			12	-0.399	Yes	12	3	Yes
6	Item Deleted	14	0	3	-0.127	8	0.115	14	-0.122			11	-0.416	Yes	13	3	Yes
7	8,9	20	0	12	-0.294	7	0.082	13	-0.164			9	-0.443	Yes	18	3	Yes
8	19	3	1	20	-0.368	13	-0.145	2	-0.258			18	-0.518	No	2	4	No
9	12	8	1	8	-0.541	14	-0.217	12	-0.399			3	-0.547	No	7	4	No
10	Item Deleted	19	1	4	-0.865	20	-0.222	11	-0.416					No	8	4	No
11	18	9	2	13	-1.402	11	-0.344	9	-0.443					No	14	4	No
12	18	12	2	11	-1.481	4	-0.419	18	-0.518					No	19	4	No
13	18	18	3	18	-1.809	12	-0.428	3	-0.547					No	20	4	No
14																	
15	Item Deleted																
16	Item Deleted																
17	Item Deleted																
18																	
19																	
20																	
		Item hold Dependency		Item Hold Low Fit Res		Most Easy Item		Low PC1 Loading						Disordered Threshold Count : 7 Ordered Threshold Count : 6			
				Fit-Res Order		Loc Order		Multidimensional									

Table 8.3.15 Local Dependencies Matrix in Texture 5

Item	2	3	4	7	8	9	11	12	13	14	18	19	20
2													
3	0.19												
4	-0.104	-0.059											
7	-0.21	-0.314	-0.165										
8	-0.361	-0.356	-0.053	0.178									
9	-0.132	0.163	0.062	-0.336	-0.103								
11	-0.019	0.108	-0.165	-0.338	-0.085	0.021							
12	-0.083	0.219	-0.11	-0.187	-0.242	0.164	0.07						
13	-0.08	-0.097	-0.031	-0.267	0.117	-0.009	0.085	-0.164					
14	-0.143	-0.103	0.105	-0.193	-0.104	-0.081	-0.044	-0.117	-0.04				
18	-0.038	0.032	0.043	-0.377	-0.199	0.028	0.202	0.154	0.162	0.018			
19	-0.188	-0.264	-0.223	0.493	0.129	-0.359	-0.22	-0.192	-0.175	-0.248	-0.315		
20	-0.019	-0.273	0.038	-0.003	0.063	-0.076	-0.167	-0.136	-0.114	0.019	-0.162	-0.157	

Remarks : The highlighted item is associate with problematic dependencies in which the value was greater than cut-off value as below;.

Average : -0.076

Cut-off : 0.124

Table 8.3.16 Summary Statistics of Non-Facet Analysis for Texture 6

	Location				Fit Residuals				Item-trait interaction						
	Item		Persons		Item		Persons		Chi square				RMSEA		
Analysis	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Value	df	p	Bonferroni's alpha	N	Value	Interpr
Initial Data	0.000	0.571	-0.065	1.094	0.460	1.603	-0.340	1.771	91.9	40	0.001	0.0025	144	0.095	Inadequate fit
Calibrated Data 1	0.000	0.696	-0.100	1.329	0.198	1.572	-0.425	1.760	93.3	40	0.001	0.0025	144	0.097	Inadequate fit
Calibrated Data 2	0.000	0.696	-0.100	1.329	0.198	1.572	-0.425	1.760	93.3	40	0.001	0.0025	144	0.097	Inadequate fit
Calibrated Data 3	0.000	0.747	-0.998	1.398	0.422	1.484	-0.466	1.675	66.3	36	0.001	0.0025	144	0.077	Inadequate fit
Calibrated Data 4	0.000	0.789	-0.091	1.459	0.054	1.522	-0.091	1.459	67.3	34	0.001	0.0025	144	0.083	Inadequate fit
Calibrated Data 5	0.000	0.782	-0.101	1.485	0.074	1.463	-0.413	1.579	59.5	34	0.004	0.0025	144	0.072	Inadequate fit

Reliability

	PSI		Alpha		
Analysis	**With extrms**	**No extrms**	**With extrms**	**No extrms**	**Interpretation**
Initial Data	0.937	0.937	0.932	0.929	Single person measurement
Calibrated Data 1	0.932	0.929	0.928	0.924	Single person measurement
Calibrated Data 2	0.929	0.929	0.924	0.924	Single person measurement
Calibrated Data 3	0.924	0.924	0.920	0.920	Single person measurement
Calibrated Data 4	0.924	0.924	0.920	0.920	Single person measurement
Calibrated Data 5	0.926	0.926	0.922	0.922	Single person measurement

Unidimensionality

	Paired t-tests				Binomial Test	
N Significant tests	Sample	% PST	%LB95CI	Is	P-expected Proportion	0.05
39 (Initial)	145	26.80%	23.30%	Not acceptable	Sample Size	141
31 (Final Calibrations)	141	21.98%	18.40%	Not acceptable	Observed proportion	31 0.2219858
					Lower 95% CI - Proportion	0.184

Table 8.3.17 Unidimensionality Test on each item performance of Non-Facet Analysis – Texture 6

Local Dependency				Item Fit				Unidimensionality						Thresholds info			
Item	Item Hold Dependency	Item hold Independent		Item Hold High Fit Res		Most Difficult Item		High PC1 Loading		Positive Load		Negative Load		Item Hold Disordered?	Item	Max Score	Rescored
1	5,6,16,17	1	0	1	2.433	6	1.319	5	0.68	5	0.68	4	-0.094	Yes	1	3	Yes
2	3,20	2	0	5	2.233	1	1.277	6	0.647	6	0.647	10	-0.097	Yes	3	3	Yes
3	9,18	3	0	6	2.025	5	1.201	8	0.608	8	0.608	15	-0.158	Yes	4	3	Yes
4		4	0	3	1.955	16	0.552	1	0.586	1	0.586	9	-0.244	Yes	5	3	Yes
5	6,7,16,17	9	0	17	1.336	8	0.385	17	0.562	17	0.562	20	-0.332	Yes	6	3	Yes
6	8,16,17	10	0	10	0.486	15	0.301	16	0.4	16	0.4	3	-0.342	Yes	9	3	Yes
7	Item Deleted	15	0	2	-0.04	17	0.204	4	-0.094			11	-0.445	Yes	10	3	Yes
8	16,17	5	1	16	-0.105	10	0.154	10	-0.097			13	-0.469	Yes	11	3	Yes
9	18	8	1	8	-0.126	9	-0.109	15	-0.158			2	-0.473	Yes	14	3	Yes
10	11	11	1	9	-0.262	18	-0.147	9	-0.244			14	-0.521	Yes	15	3	Yes
11	13,18,20	13	1	15	-0.384	11	-0.339	20	-0.332			18	-0.58	Yes	16	3	Yes
12	Item Deleted	14	1	4	-0.399	4	-0.478	3	-0.342					Yes	17	3	Yes
13	14	6	2	20	-1.005	20	-0.632	11	-0.445					Yes	18	3	Yes
14	18,20	16	4	14	-1.325	3	-0.734	13	-0.469					Yes	20	3	Yes
15	20	17	4	13	-1.818	14	-0.769	2	-0.473					No	2	4	No
16		18	4	18	-1.86	13	-0.923	14	-0.521					No	8	4	No
17		20	4	11	-1.889	2	-1.262	18	-0.58					No	13	4	No
18																	
19	Item Deleted																
20																	
		Item hold Dependency		Item Hold Low Fit Res		Most Easy Item		Low PC1 Loading						Disordered Threshold Count : 14 Ordered Threshold Count : 3			
				Fit-Res Order		Loc Order		Multidimensional									

Table 8.3.18 Local Dependencies Matrix in Texture 6

Item	1	2	3	4	5	6	8	9	10	11	13	14	15	16	17	18	20
1																	
2	-0.332																
3	-0.159	0.15															
4	-0.159	0.011	-0.015														
5	0.201	-0.305	-0.313	-0.027													
6	0.343	-0.395	-0.119	-0.091	0.262												
8	0.106	-0.258	-0.228	-0.018	0.423	0.247											
9	0.011	-0.043	0.156	-0.126	-0.306	-0.106	-0.282										
10	-0.01	-0.035	0.109	-0.162	-0.117	-0.06	-0.165	-0.039									
11	-0.165	0.018	0.047	0.011	-0.241	-0.29	-0.259	-0.097	0.14								
13	-0.281	0.1	-0.133	-0.046	-0.292	-0.348	-0.249	0.104	-0.078	0.167							
14	-0.334	0.098	0.033	0.017	-0.319	-0.351	-0.263	0.122	-0.105	-0.004	0.318						
15	-0.184	0.079	-0.166	-0.112	-0.079	-0.192	0.032	-0.073	-0.156	0.063	0.02	0.075					
16	0.16	-0.208	-0.294	-0.083	0.136	0.214	0.166	-0.15	-0.178	-0.034	-0.097	-0.266	-0.118				
17	0.243	-0.165	-0.209	-0.169	0.317	0.155	0.213	-0.088	-0.073	-0.469	-0.23	-0.154	-0.067	0.084			
18	-0.32	0.083	0.187	-0.002	-0.397	-0.314	-0.376	0.174	0.082	0.257	0.092	0.138	-0.003	-0.115	-0.347		
20	-0.333	0.177	-0.08	0.035	-0.099	-0.178	-0.14	-0.244	-0.212	0.134	0.068	0.21	0.165	-0.129	-0.2	0.078	

Remarks : The highlighted item is associate with problematic dependencies in which the value was greater than cut-off value as below;.

Average : -0.059

Cut-off : 0.140

Table 8.3.19 Summary Statistics of Non-Facet Analysis for Texture 7

	Location				**Fit Residuals**				**Item-trait interaction**						
	Item		**Persons**		**Item**		**Persons**		**Chi square**				**RMSEA**		
Analysis	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Value**	**df**	**p**	**Bonferroni's alpha**	**N**	**Value**	**Interpr**
Initial Data	0.000	0.262	0.611	1.110	0.481	1.760	-0.353	1.788	84.6	40	0.001	0.0025	144	0.088	Inadequate fit
Calibrated Data 1	0.000	0.323	0.709	1.285	0.232	1.505	-0.422	1.836	59.5	40	0.024	0.0025	144	0.058	Inadequate fit
Calibrated Data 2	0.000	0.323	0.709	1.285	0.232	1.505	-0.422	1.836	59.5	40	0.024	0.0025	144	0.058	Inadequate fit
Calibrated Data 3	0.000	0.359	0.803	1.344	0.266	1.591	-0.286	1.626	69.4	40	0.003	0.0025	139	0.073	Inadequate fit
Calibrated Data 4	0.000	0.321	0.858	1.402	0.283	1.102	-0.294	1.521	38.8	34	0.261	0.0025	139	0.032	Inadequate fit

Reliability

	PSI		**Alpha**		
Analysis	**With extrms**	**No extrms**	**With extrms**	**No extrms**	**Interpretation**
Initial Data	0.933	0.933	0.940	0.938	Single person measurement
Calibrated Data 1	0.933	0.933	0.937	0.935	Single person measurement
Calibrated Data 2	0.933	0.926	0.935	0.935	Single person measurement
Calibrated Data 3	0.936	0.936	0.938	0.938	Single person measurement
Calibrated Data 4	0.927	0.927	0.930	0.930	Single person measurement
Calibrated Data 5	0.933	0.933	0.940	0.938	Single person measurement

Unidimensionality

	Paired t-tests				Binomial Test	
N Significant tests	Sample	% PST	%LB95CI	Is	P-expected Proportion	0.05
46 (Initial)	145	31.70%	28.20%	Not acceptable	Sample Size	139
43 (Final Calibrations)	137	30.90%	27.30%	Not acceptable	Observed proportion	43
						0.309353
					Lower 95% CI - Proportion	0.273

Table 8.3.20 Unidimensionality Test on each item performance of Non-Facet Analysis – Texture 7

Local Dependency				Item Fit				Unidimensionality						Thresholds info			
Item	Item Hold Dependency	Item hold Independent		Item Hold High Fit Res		Most Difficult Item		High PC1 Loading		Positive Load		Negative Load		Item Hold Disordered?	Item	Max Score	Rescored
1	6,9	1	0	6	1.879	14	0.62	18	0.614	18	0.614	8	-0.171	Yes	1	3	Yes
2	Item Deleted	3	0	14	1.609	5	0.436	14	0.508	14	0.508	17	-0.34	Yes	3	3	Yes
3	4,9,10,12,18	5	0	10	1.489	18	0.399	13	0.5	13	0.5	1	-0.351	Yes	4	3	Yes
4		13	0	1	1.475	6	0.25	3	0.499	3	0.499	16	-0.467	Yes	5	3	Yes
5	6,7,8	4	1	12	1.35	12	0.243	12	0.456	12	0.456	7	-0.526	Yes	6	3	Yes
6	17	7	1	19	1.265	19	0.081	20	0.444	20	0.444	19	-0.529	Yes	9	3	Yes
7	16,19	8	1	7	1.223	16	0.038	9	0.335	9	0.335	6	-0.568	Yes	10	3	Yes
8	16	12	1	5	0.656	20	0.023	10	0.218	10	0.218	5	-0.577	Yes	12	3	Yes
9	10	14	1	16	-0.141	13	-0.03	4	0.131	4	0.131			Yes	17	3	Yes
10		17	1	13	-0.347	9	-0.086	8	-0.171					Yes	18	3	Yes
11	Item Deleted	6	2	4	-0.554	17	-0.093	17	-0.34					No	7	4	No
12	18,20	9	2	17	-0.626	7	-0.126	1	-0.351					No	8	4	No
13	14,18	10	2	20	-0.662	10	-0.207	16	-0.467					No	13	4	No
14	18,20	16	2	18	-0.864	1	-0.264	7	-0.526					No	14	4	No
15	Item Deleted	19	2	9	-0.946	4	-0.305	19	-0.529					No	16	4	No
16	19	20	3	3	-0.966	8	-0.307	6	-0.568					No	19	4	No
17		18	4	8	-1.032	3	-0.673	5	-0.577					No	20	4	No
18																	
19																	
20	20																
		Item hold Dependency		Item Hold Low Fit Res		Most Easy Item		Low PC1 Loading						Disordered Threshold Count : 10 Ordered Threshold Count : 7			
				Fit-Res Order		Loc Order		Multidimensional									

Table 8.3.21 Local Dependencies Matrix in Texture 7

Item	1	3	4	5	6	7	8	9	10	12	13	14	16	17	18	19	20
1																	
3	-0.146																
4	-0.111	0.136															
5	0.122	-0.271	-0.021														
6	0.403	-0.243	-0.23	0.311													
7	-0.041	-0.215	-0.031	0.174	0.123												
8	-0.266	-0.252	-0.016	0.125	-0.167	0.088											
9	0.131	0.218	-0.11	-0.149	-0.12	-0.334	-0.231										
10	0.025	0.272	0.015	-0.13	-0.121	-0.315	-0.186	0.135									
12	-0.179	0.156	0.074	-0.318	-0.165	-0.241	-0.122	-0.029	-0.062								
13	-0.134	0.086	0.022	-0.259	-0.274	-0.313	-0.123	0.072	0.076	0.12							
14	-0.267	-0.039	-0.132	-0.293	-0.314	-0.121	0.004	0.107	-0.118	0.026	0.217						
16	-0.009	-0.241	-0.072	0.079	-0.049	0.149	0.157	-0.266	-0.148	-0.165	-0.213	-0.299					
17	0.048	-0.16	-0.028	0.085	0.24	0.017	0.023	-0.108	-0.013	-0.096	-0.235	-0.222	0.12				
18	-0.19	0.228	0.041	-0.3	-0.337	-0.369	-0.097	0.089	0.023	0.21	0.139	0.22	-0.222	-0.179			
19	0.024	-0.261	-0.153	0.121	0.104	0.257	0.004	-0.153	-0.089	-0.31	-0.233	-0.283	0.285	-0.091	-0.278		
20	-0.339	0.03	-0.102	-0.242	-0.23	-0.092	0.05	-0.029	-0.166	0.16	0.065	0.291	-0.215	-0.204	0.22	-0.243	

Remarks : The highlighted item is associate with problematic dependencies in which the value was greater than cut-off value as below;.

Average : -0.060

Cut-off : 0.13

Appendix A – Subtest 2

Table 8.3.22: Second Unidimensionality T-test for texture 1

Unidimensionality

Paired t-tests

Subset form PC1 loading	Sample *n*	Observed *n*	% PST	%LB95CI	Is
Positive-Positive subset	138	10	7.24%	0.036%	Acceptable
Negative-Negative subset	145	11	7.58%	0.040%	Acceptable
Positive-Negative subset	145	12	8.27%	0.047%	Acceptable

	Item	PC1	Binomial Test	
1	6	0.724	P-expected Proportion	0.05
2	5	0.460	Sample Size *n*	145
3	1	0.382	Observed proportion *n*	12
4	17	0.076	Proportion of significant test	0.082759
5	8	0.055	q	0.95
6	7	-0.167	SD	2.62
7	13	-0.410	X	7.25
8	19	-0.442	t 0.05	1.96
9	16	-0.534	Lower 95% CI-Proportion	0.047

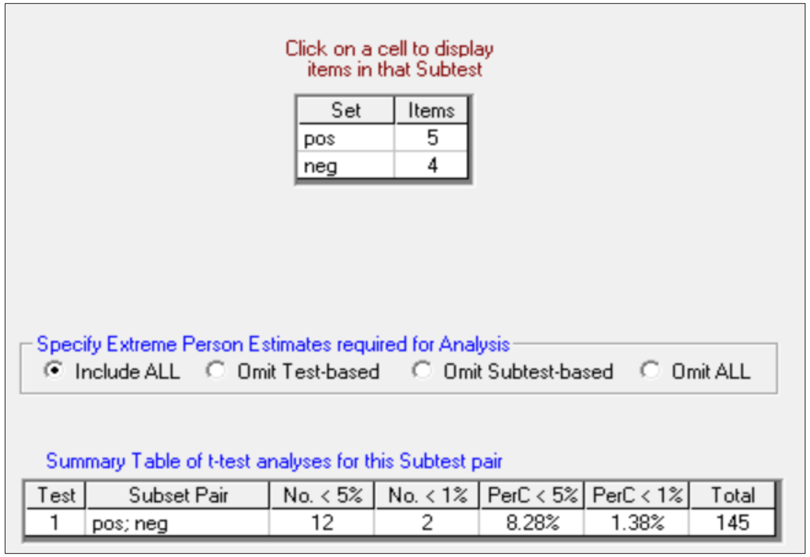


Table 8.3.23: Second Unidimensionality T-test for texture 2

Unidimensionality

Paired t-tests

Subset form PC1 loading	Sample *n*	Observed *n*	% PST	%LB95CI	Is
Positive-Positive subset	145	38	22.7%	0.262%	Not Acceptable
Negative-Negative subset	139	15	10.7%	0.072%	Not Acceptable
Positive-Negative subset	138	10	7.24%	0.036%	Acceptable

	Item	**PC1**	**Binomial Test**	
1	13	0.681	P-expected Proportion	0.05
2	14	0.613	Sample Size *n*	138
3	18	0.302	Observed proportion *n*	10
4	20	-0.064	Proportion of significant test	0.072464
5	9	-0.082	q	0.95
6	2	-0.258	SD	2.56
7	3	-0.354	X	6.90
8	4	-0.470	t 0.05	1.96
9	12	-0.476	Lower 95% CI-Proportion	0.036

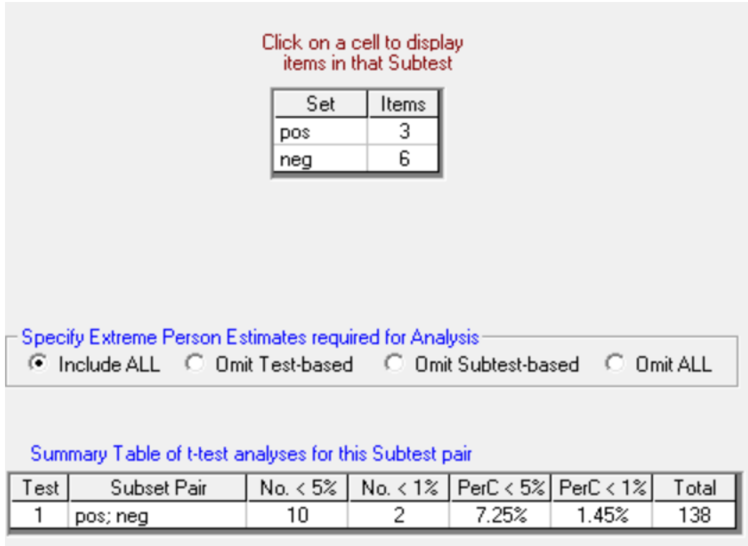


Table 8.3.24: Second Unidimensionality T-test for texture 3

Unidimensionality

Paired t-tests

Subset form PC1 loading	Sample *n*	Observed *n*	% PST	%LB95CI	Is
Positive-Positive subset	145	15	10.3%	0.068%	Not Acceptable
Negative-Negative subset	138	10	7.24%	0.036%	Acceptable
Positive-Negative subset	145	12	8.27%	0.047%	Acceptable

	Item	**PC1**	**Binomial Test**	
1	13	0.676	P-expected Proportion	0.05
2	14	0.545	Sample Size *n*	145
3	4	0.466	Observed proportion *n*	12
4	20	0.325	Proportion of significant test	0.082759
5	12	0.035	q	0.95
6	9	-0.087	SD	2.62
7	11	-0.287	X	7.25
8	2	-0.293	t 0.05	1.96
9	10	-0.421	Lower 95% CI-Proportion	0.047
10	3	-0.473		
11	18	-0.501		

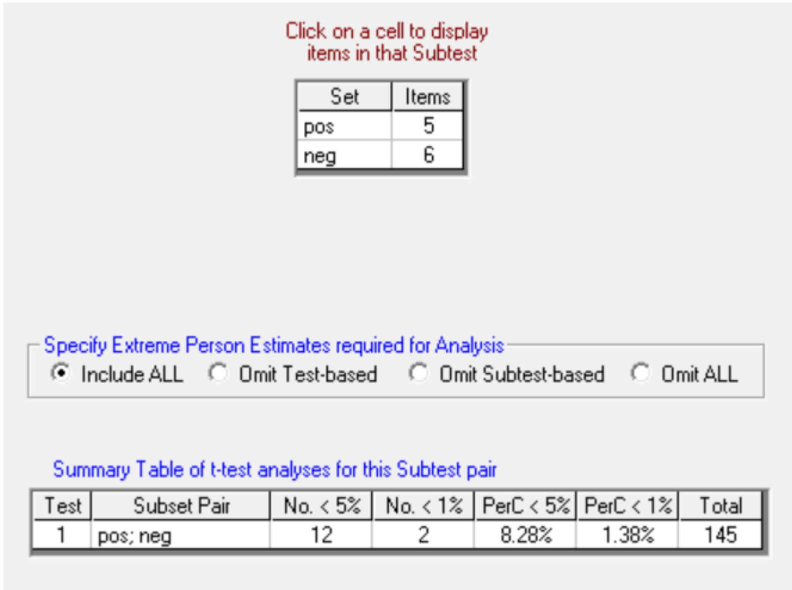


Table 8.3.25: Second Unidimensionality T-test for texture 4

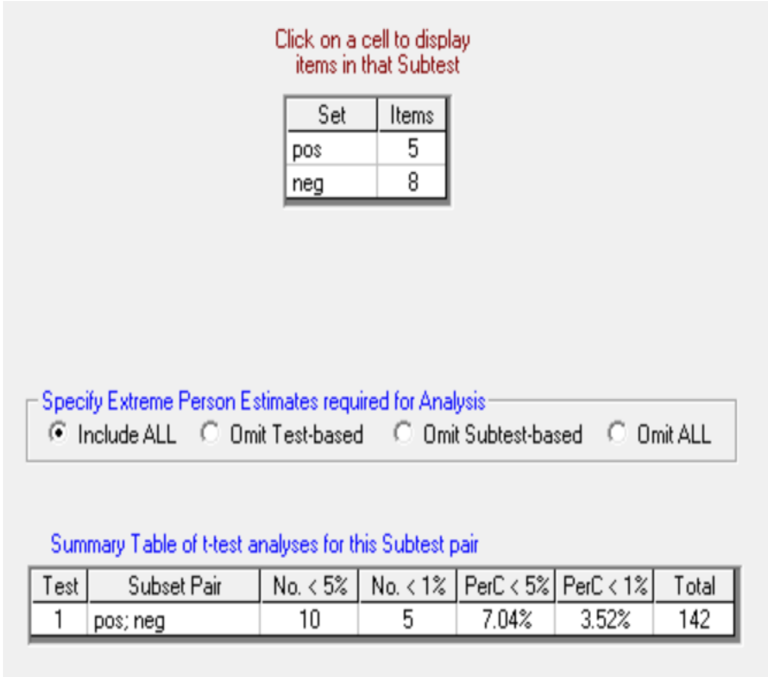
Unidimensionality

Paired t-tests

Subset form PC1 loading	Sample *n*	Observed *n*	% PST	%LB95CI	Is
Positive-Positive subset	145	14	9.65%	0.061%	Not Acceptable
Negative-Negative subset	145	10	6.89%	0.033%	Acceptable
Positive-Negative subset	145	10	6.89%	0.033%	Acceptable

	Item	**PC1**
1	10	0.719
2	3	0.617
3	11	0.320
4	12	0.312
5	1	0.205
6	9	-0.034
7	15	-0.155
8	20	-0.215
9	2	-0.226
10	4	-0.231
11	18	-0.329
12	13	-0.424
13	14	-0.560

Binomial Test	
P-expected Proportion	0.05
Sample Size *n*	145
Observed proportion *n*	10
Proportion of significant test	0.068966
q	0.95
SD	2.62
X	7.25
t 0.05	1.96
Lower 95% CI-Proportion	0.033



Click on a cell to display items in that Subtest

Set	Items
pos	5
neg	8

Specify Extreme Person Estimates required for Analysis

Include ALL | Omit Test-based | Omit Subtest-based | Omit ALL

Summary Table of t-test analyses for this Subtest pair

Test	Subset Pair	No. < 5%	No. < 1%	PerC < 5%	PerC < 1%	Total
1	pos; neg	10	5	7.04%	3.52%	142

Table 8.3.26: Second Unidimensionality T-test for texture 5

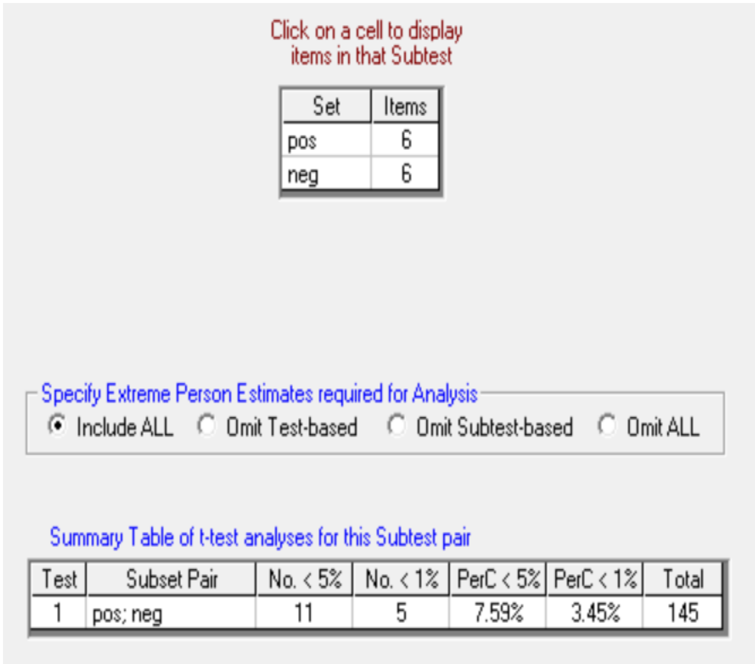
Unidimensionality

Paired t-tests

Subset form PC1 loading	Sample *n*	Observed *n*	% PST	%LB95CI	Is
Positive-Positive subset	145	10	6.89%	0.033%	Acceptable
Negative-Negative subset	145	11	7.58%	0.040%	Acceptable
Positive-Negative subset	145	11	7.58%	0.040%	Acceptable

	Item	PC1
1	3	0.666
2	12	0.529
3	11	0.365
4	9	0.246
5	2	0.232
6	18	0.214
7	10	-0.010
8	15	-0.154
9	13	-0.160
10	4	-0.313
11	14	-0.394
12	20	-0.525
13	8	-0.558

Binomial Test	
P-expected Proportion	0.05
Sample Size *n*	145
Observed proportion *n*	11
Proportion of significant test	0.075862
q	0.95
SD	2.62
X	7.25
t 0.05	1.96
Lower 95% CI-Proportion	0.040


Click on a cell to display items in that Subtest

Set	Items
pos	6
neg	6

Specify Extreme Person Estimates required for Analysis

Include ALL Omit Test-based Omit Subtest-based Omit ALL

Summary Table of t-test analyses for this Subtest pair

Test	Subset Pair	No. < 5%	No. < 1%	PerC < 5%	PerC < 1%	Total
1	pos; neg	11	5	7.59%	3.45%	145

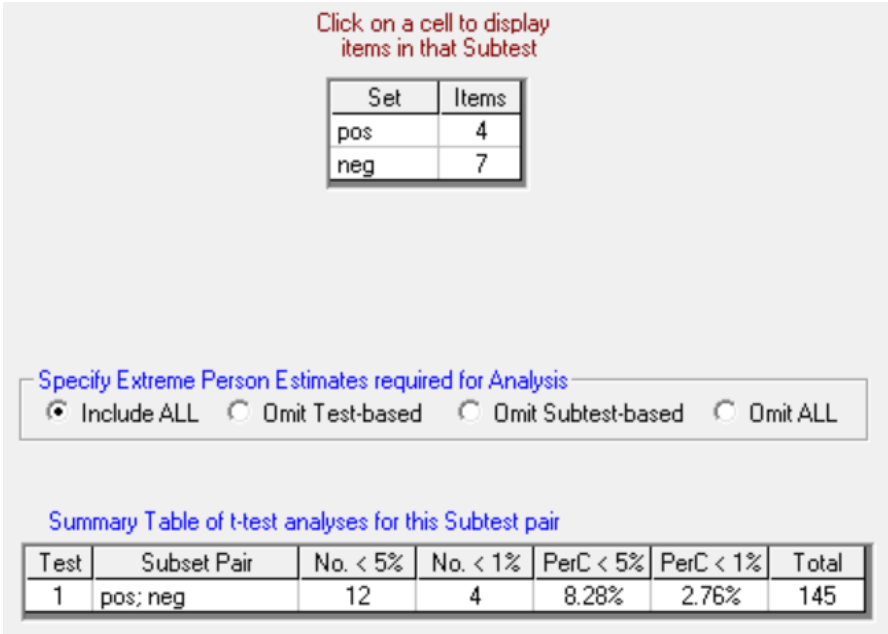
Table 8.3.27: Second Unidimensionality T-test for texture 6

Unidimensionality

Paired t-tests

Subset form PC1 loading	Sample *n*	Observed *n*	% PST	%LB95CI	Is
Positive-Positive subset	145	12	8.27%	0.047%	Acceptable
Negative-Negative subset	145	19	13.1%	0.096%	Not Acceptable
Positive-Negative subset	145	12	8.27%	0.047%	Acceptable

	Item	**PC1**	**Binomial Test**	
1	14	0.576	P-expected Proportion	0.05
2	20	0.545	Sample Size *n*	145
3	13	0.493	Observed proportion *n*	12
4	15	0.403	Proportion of significant test	0.082759
5	2	0.226	q	0.95
6	4	-0.004	SD	2.62
7	9	-0.077	X	7.25
8	18	-0.167	t 0.05	1.96
9	11	-0.289	Lower 95% CI-Proportion	0.047
10	12	-0.400		
11	10	-0.441		
12	3	-0.565		


Click on a cell to display items in that Subtest

Set	Items
pos	4
neg	7

Specify Extreme Person Estimates required for Analysis

Include ALL | Omit Test-based | Omit Subtest-based | Omit ALL

Summary Table of t-test analyses for this Subtest pair

Test	Subset Pair	No. < 5%	No. < 1%	PerC < 5%	PerC < 1%	Total
1	pos; neg	12	4	8.28%	2.76%	145

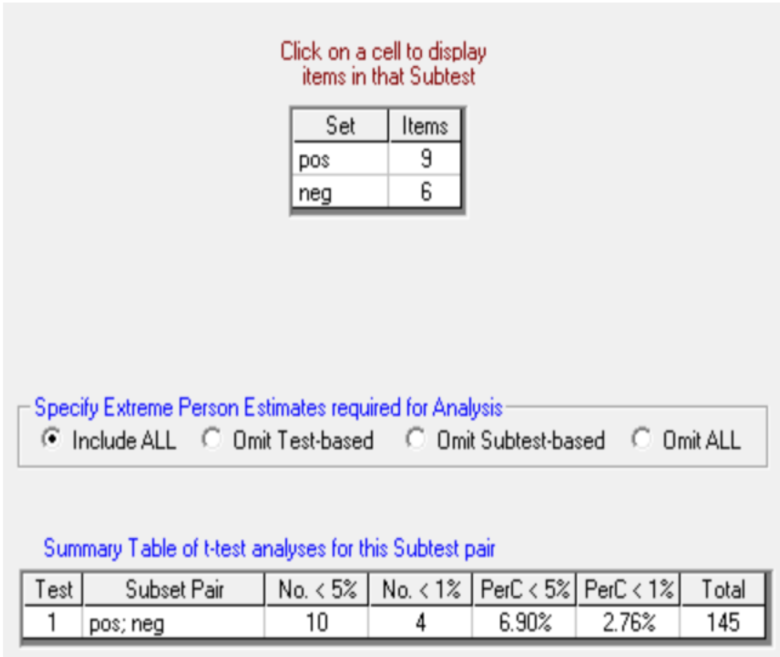
Table 8.3.28: Second Unidimensionality T-test for texture 7

Unidimensionality

Paired t-tests

Subset form PC1 loading	Sample *n*	Observed *n*	% PST	%LB95CI	Is
Positive-Positive subset	145	9	6.20%	0.027%	Acceptable
Negative-Negative subset	145	11	7.58%	0.040%	Not Acceptable
Positive-Negative subset	145	10	6.89%	0.033%	Acceptable

	Item	**PC1**	**Binomial Test**	
1	3	0.605	P-expected Proportion	0.05
2	10	0.516	Sample Size *n*	145
3	9	0.423	Observed proportion *n*	10
4	11	0.411	Proportion of significant test	0.068966
5	12	0.305	q	0.95
6	1	0.236	SD	2.62
7	4	0.174	X	7.25
8	18	0.136	t 0.05	1.96
9	13	0.012	Lower 95% CI-Proportion	0.033
10	16	-0.193		
11	2	-0.229		
12	15	-0.354		
13	14	-0.429		
14	20	-0.461		



Appendix B

Performance, Governance and Operations
Research & Innovation Service
Charles Thackrah Building
101 Clarendon Road
Leeds LS2 9LJ Tel: 0113 343 4873
Email: ResearchEthics@leeds.ac.uk

UNIVERSITY OF LEEDS

Farzilnizam Ahmad
School of Mechanical Engineering
University of Leeds
Leeds, LS2 9JT

MaPS and Engineering joint Faculty Research Ethics Committee (MEEC FREC)
University of Leeds

12 May 2015

Dear Farzilnizam

Title of study	**Affective Design of Textures for Vehicle Interiors**
Ethics reference	**MEEC 14-025**

I am pleased to inform you that the application listed above has been reviewed by the MaPS and Engineering joint Faculty Research Ethics Committee (MEEC FREC) and I can confirm a favourable ethical opinion as of the date of this letter. The following documentation was considered:

Document	*Version*	*Date*
MEEC 14-025 Farzil_ethics_application_March_2015.pdf	1	25/03/15
MEEC 14-025 Farzilnizam_Ahmad_ethics_application_signatures.pdf	1	25/03/15
MEEC 14-025 volunteers needed for research.txt	1	25/03/15
MEEC 14-025 farzil_info_consent_form.doc	1	25/03/15
MEEC 14-025 Farzil_consent_form.doc	1	25/03/15
MEEC 14-025 Farzil_H&S_Risk_Assessment.docx	1	25/03/15

Committee members made the following comments about the application:

- A4) You needed to indicate this project is part of a PhD qualification.
- A8) Based on answers to C7 it is likely you will be using participants known to them so this should have been mentioned.
- C15) Presumably should read “No”.
- C16) “Yes” needed to be ticked.
- C20) Any hard copies of consent forms would need to be stored in a secure place (eg a locked cupboard) on University of Leeds premises.

- The information sheet and consent form should be on headed paper and include version numbers and your ethics reference and date of approval.

Please notify the committee if you intend to make any amendments to the original research as submitted at date of this approval, including changes to recruitment methodology. All changes must receive ethical approval prior to implementation. The amendment form is available at http://ris.leeds.ac.uk/EthicsAmendment.

Please note: You are expected to keep a record of all your approved documentation, as well as documents such as sample consent forms, and other documents relating to the study. This should be kept in your study file, which should be readily available for audit purposes. You will be given a two week notice period if your project is to be audited. There is a checklist listing examples of documents to be kept which is available at http://ris.leeds.ac.uk/EthicsAudits.

We welcome feedback on your experience of the ethical review process and suggestions for improvement. Please email any comments to ResearchEthics@leeds.ac.uk.

Yours sincerely

Jennifer Blaikie
Senior Research Ethics Administrator, Research & Innovation Service
On behalf of Professor Gary Williamson, Chair, MEEC FREC

CC: Student’s supervisor(s)

Performance, Governance and Operations
Research & Innovation Service
Charles Thackrah Building
101 Clarendon Road
Leeds LS2 9LJ Tel: 0113 343 4873
Email: ResearchEthics@leeds.ac.uk

UNIVERSITY OF LEEDS

Farzilnizam Ahmad
Mechanical Engineering
University of Leeds
Leeds, LS2 9JT

MaPS and Engineering joint Faculty Research Ethics Committee (MEEC FREC)

University of Leeds

20 June 2019

Dear Farzilnizam

Title of study	**Measurement of affective responses from paired comparisons of stimuli**
Ethics reference	**MEEC 15-027**

I am pleased to inform you that the application listed above has been reviewed by the MaPS and Engineering joint Faculty Research Ethics Committee (MEEC FREC) and following receipt of your response to the Committee's initial comments, I can confirm a favourable ethical opinion as of the date of this letter. The following documentation was considered:

Document	*Version*	*Date*
MEEC 15-027 Farzil_Paired comparisons Ethical_Review_Form_version_2.doc	2	16/02/16
MEEC 15-027 volunteers needed for research.txt	2	16/02/16
MEEC 15-027 Paired_comparisons_info_consent_form_version_2.doc	2	16/02/16
MEEC 15-027 Paired_comparisons_consent_form.doc	2	16/02/16

Please notify the committee if you intend to make any amendments to the original application as submitted at date of this approval as all changes must receive ethical approval prior to implementation. The amendment form is available at http://ris.leeds.ac.uk/EthicsAmendment.

Please note: You are expected to keep a record of all your approved documentation. You will be given a two week notice period if your project is to be audited. There is a checklist listing examples of documents to be kept which is available at http://ris.leeds.ac.uk/EthicsAudits.

We welcome feedback on your experience of the ethical review process and suggestions for improvement. Please email any comments to ResearchEthics@leeds.ac.uk.

Yours sincerely

Jennifer Blaikie

Senior Research Ethics Administrator, Research & Innovation Service
On behalf of Professor Gary Williamson, Chair, MEEC FREC

CC: Student's supervisor(s)

Performance, Governance and Operations
Research & Innovation Service
Charles Thackrah Building
101 Clarendon Road
Leeds LS2 9LJ Tel: 0113 343 4873
Email: ResearchEthics@leeds.ac.uk

UNIVERSITY OF LEEDS

Farzilnizam Ahmad
School of Mechanical Engineering
University of Leeds
Leeds, LS2 9JT

MaPS and Engineering joint Faculty Research Ethics Committee (MEEC FREC)

University of Leeds

16 October 2017

Dear Farzilnizam

Title of study	**Measuring Affective Response of Vehicle Interior Texture Using Pair Comparison Method in Rasch Model.**
Ethics reference	**MEEC 16-050**

I am pleased to inform you that the application listed above has been reviewed by the MaPS and Engineering joint Faculty Research Ethics Committee (MEEC FREC) and following receipt of your response to the Committee's initial comments, I can confirm a favourable ethical opinion as of the date of this letter. The following documentation was considered:

Document	Version	Date
MEEC 16-050 response 1.txt	1	11/09/17
7) MEEC 16-050 Version 2 Nested PC Participant Information Sheet.doc	2	11/09/17
8) MEEC 16-050 Version 2 Nested PC Consent_form.doc	2	11/09/17
MEEC 16-050 1) Nested PC Ethical Approval Form 19July 17.docx	1	19/07/17
MEEC 16-050 2) Nested Pair Comparison H&S_Risk_Assessment.docx	1	19/07/17
MEEC 16-050 Poster NESTED PC 010817.jpg	1	19/07/17
MEEC 16-050 Nested PC Ethical Approval 2017.pdf	1	19/07/17
MEEC 16-050 Nested PC Risk Assessment2017.pdf	1	19/07/17

Please notify the committee if you intend to make any amendments to the information in your ethics application as submitted at date of this approval as all changes must receive ethical approval prior to implementation. The amendment form is available at http://ris.leeds.ac.uk/EthicsAmendment.

Please note: You are expected to keep a record of all your approved documentation and other documents relating to the study, including any risk assessments. This should be kept in your study file, which should be readily available for audit purposes. You will be given a two week notice period if your project is to be audited. There is a checklist listing examples of documents to be kept which is available at http://ris.leeds.ac.uk/EthicsAudits.

We welcome feedback on your experience of the ethical review process and suggestions for improvement. Please email any comments to ResearchEthics@leeds.ac.uk.

Yours sincerely

Jennifer Blaikie

Senior Research Ethics Administrator, the Secretariat
On behalf of Dr Dawn Groves, Chair, MEEC FREC

CC: Student's supervisor(s)

Appendix C

UNIVERSITY OF LEEDS

Semantic Differential Protocol

(MEEC 14-025)

Introduction
Hi everyone..

Let me introduce myself. My name is Farzilnizam. I am a PhD student in the Affective Engineering Research Group, Institute of Design Robotics & Optimisation, School of Mechanical Engineering. I will be conducting this research session today.

The objective of this research is to find out how people feel about the quality interior vehicle trims design specially related to texture design. How people perceive about high quality or low quality, comfort and discomfort as well others things related to the subjective feeling of interior vehicle trims. This study will also investigate and identify new possibilities on how the interior vehicle trim can be improved. We have made special texture samples that are represent the interior vehicle trims.

This research concerns the improvement of products development process by address the quality interior trim of production vehicles with the aim to investigate and identify new possibilities on how the interior vehicle trim can be improve.

In front of you, you will find a collection of texture samples that have been numbered 1 to 7. In a moment, we will ask you to touch the texture sample and completed questionnaires to say how you feel to touch the texture samples. When you touch the texture sample, please use your hand that you write with and touch the texture sample in way that seems most normal and natural to you. Like this.
(Demonstration of touching surface).

Please take a moment now to touch each texture samples. As you do so, please do not say anything or make any remarks about the texture samples to other participants. Make sure you touch each of the seven texture samples. As you touch the texture sample, remember that we are thinking about how they would feel if they were used to design vehicle interiors.
(Show participants to the posters and wait patiently until everyone is ready to continue)

Is everybody happy that they can touch the texture samples and that for the most part they each feel slightly different?
What we would like you to do now is the following. In a moment, we will give each of you a set of questionnaires. At the top of each questionnaire is the numbers of the texture sample that the questionnaire refers to.

(Hold up an example semantic questionnaire.. show “Use this questionnaire for SAMPLE 1,2,3..7”)

For each questionnaire, we want you to take the texture sample and touch in the way we showed you before.
(Repeat demonstration of touching surface)

We then ask you to complete the questionnaire for that texture sample. You complete the questionnaire like this.
(Hold up an example semantic questionnaire – Instruction to tick the semantic box)

The questionnaire is made up of pair words of opposite meaning that are separated on a seven point scale. For each texture sample, we would like you to decide which of the two words best describes how you feel about the texture sample if it were used to designing the vehicle interiors. If you strongly agree that one of the words describes how you feel, then make a mark, a tick or a cross in the box closest to the word, if you agree that the word describe how you feel, then mark the next box along, and if you partly agree that the word describes how you feel then mark the box next to that one. If you are indifferent or cannot say which word best describes your feelings, then mark the box in the middle. I have bold the middle box for you to recognise easily.

(Demonstrate to tick the semantic box)
Do this for each pair of words. The order of the words and which side of the scale they appear on is random order, so you will have to pay attention to each sheet. Additionally, at the bottom of each sheet you are asked to indicate whether you like or dislike the texture sample on a three point scale. (Each words are not in order in every page)

Please work quickly through the questionnaires. Do not dwell (stay) on any of the words and avoid becoming too analytical. We want your initial, instinctive reactions to each texture samples.

Feel free to touch the texture samples as often as you need whilst you fill in the questionnaire. Once you have finished the questionnaire for each texture sample, return the texture sample the middle of the table.

We ask you to complete the questionnaires in the order that they are presented to you because the order of the sheets is special to you. If you need the texture sample and someone else is using it, please wait for it to become available.

As you complete the questionnaires, we ask you not to say anything or make any remarks about the texture samples.
It should take you about 8 minutes to complete each page or the sample of questionnaires.

Are there any questions?
(wait ten seconds)

As reminder, we want to find out how you feel about these texture surfaces if they were to be used to designing the vehicle interiors, and we want you to work quickly through the questionnaires, not thinking too much about them but giving your instinctive reactions. If you have any questions during the session, please do not hesitate to ask. I will now give out the questionnaires. Feel free to start completing them as soon as you received them.

UNIVERSITY OF LEEDS

Likert Scale Protocol

(MEEC 14-025)

Introduction
Hi everyone..

Let me introduce myself. My name is Farzilnizam. I am a PhD student in the Affective Engineering Research Group, Institute of Design Robotics & Optimisation, School of Mechanical Engineering. I will be conducting this research session today.

The objective of this research is to find out how people feel about the quality interior vehicle trims design specially related to texture design. How people perceive about high quality or low quality, comfort and discomfort as well others things related to the subjective feeling of interior vehicle trims. This study will also investigate and identify new possibilities on how the interior vehicle trim can be improved. We have made special texture samples that are represent the interior vehicle trims.

This research concerns the improvement of products development process by address the quality interior trim of production vehicles with the aim to investigate and identify new possibilities on how the interior vehicle trim can be improve.

In front of you, you will find a collection of texture samples that have been numbered 1 to 7. In a moment, we will ask you to touch the texture sample and completed questionnaires to say how you feel to touch the texture samples. When you touch the texture sample, please use your hand that you write with and touch the texture sample in way that seems most normal and natural to you. Like this.
(Demonstration of touching surface).

Please take a moment now to touch each texture samples. As you do so, please do not say anything or make any remarks about the texture samples to other participants. Make sure you touch each of the seven texture samples. As you touch the texture sample, remember that we are thinking about how they would feel if they were used to design vehicle interiors.

(Show participants to the posters and wait patiently until everyone is ready to continue)
Is everybody happy that they can touch the texture samples and that for the most part they each feel slightly different?

What we would like you to do now is the following. In a moment, we will give each of you a set of questionnaires. At the top of each questionnaire is the numbers of the texture sample that the questionnaire refers to.
(Hold up an example LS questionnaire.. show "Use this questionnaire for SAMPLE 1,2,3..7")

For each questionnaire, we want you to take the texture sample and touch in the way we showed you before.
(Repeat demonstration of touching surface)

We then ask you to complete the questionnaire for that texture sample. You complete the questionnaire like this.
(Hold up an example LS questionnaire – Instruction to tick the LS box)

The questionnaire is made up of statements on a five point LS scale. For each texture sample, we would like you to decide which of the LS point best describes how you feel about the texture sample if it were used to designing the vehicle interiors.

The LS scale consist 5 point value that describe *Strongly Agree, Agree, Neutral, Disagree and Strongly Disagree.*

If you S*trongly Agree* on the statement indicate, then make a mark, a tick or a cross in the circle below to the words *"Strongly Agree".* Do the same things for *agree, disagree and strongly disagree*. If you are indifferent or cannot say which word best describes your feelings, then mark the circle in below the words *"Neutral".*
(Demonstrate to tick the LS questionnaire box)

Do this for each statement. The order of the statement appear on is random order, so you will have to pay attention to each sheet. (Statements are not in order in every page)

Please work quickly through the questionnaires. Do not dwell (stay) on any of the words and avoid becoming too analytical. We want your initial, instinctive reactions to each texture samples.

Feel free to touch the texture samples as often as you need whilst you fill in the questionnaire. Once you have finished the questionnaire for each texture sample, return the texture sample the middle of the table.

We ask you to complete the questionnaires in the order that they are presented to you because the order of the sheets is special to you. If you need the texture sample and someone else is using it, please wait for it to become available.

As you complete the questionnaires, we ask you not to say anything or make any remarks about the texture samples.
It should take you about 8 minutes to complete each page or the sample of questionnaires.

Are there any questions?
(wait ten seconds)

As reminder, we want to find out how you feel about these texture surfaces if they were to be used to designing the vehicle interiors, and we want you to work quickly through the questionnaires, not thinking too much about them but giving your instinctive reactions. If you have any questions during the session, please do not hesitate to ask.
I will now give out the questionnaires. Feel free to start completing them as soon as you received them.

UNIVERSITY OF LEEDS

Pair Comparison Protocol - Confectioneries
(MEEC 15-027)

Introduction
Hi everyone..

Let me introduce myself. My name is Farzilnizam. I am a PhD student in the Affective Engineering Research Group, Institute of Design Robotics & Optimisation, School of Mechanical Engineering. I will be conducting this research session today.

The objective of this research is to find out the specialness of chocolate. The aim of this new research is improve how we can measure people's emotional engagement with products.

In front of you, you will find a collection of chocolates samples that have been labelled 1 to 4. This chocolate are physical representation that correspond to visual sample (stimuli) in computer screen when you were asked to answer the questionnaire later.

Health & Safety
As recommendation by The Food Standards Agency, UK, you are not require to eat or taste the chocolates. The chocolates may have contain nut, gluten, soy, milk, alcohol and other ingredients alert. This means the product is a possible health risk for anyone with an allergy.

Brief introduction
In a moment, we will ask you to observe the sample of chocolates and completed questionnaires pretty soon. Feel free to touch the sample however, please do not squeeze it.

While you observe the chocolates, I would like to show you short video about chocolates and simple introduction briefing about our research context today. (To evoke participants' interest about topic or research context)

Are you happy to start the experimental survey now?
(wait ten seconds)

Brief about survey instruments using PC

What we would like you to do now is the following. This is online questionnaire, no paper and pencils. In a moment, you will asked to complete the questionnaire. The questionnaire is made up of pair of stimuli (chocolate sample). (Show to computer screen)

When you responding to the question, you need to simply click your answer whether your answer are correspond to the chocolates on your left of right by clicking the pictures like this. To move to next question, please click next button as shown in the computer screen.

(Demonstration of answering the questionnaire using the PC)

For each question, we would like you to decide which one of the two chocolates best describes your answer. If you decide chocolate on your left, please click box on your left and if you decide chocolates on your right, please click the box on your right.

You need to answer it before you can move to the next question. In other words you cannot leave the question blank (without answering) the computer program will not proceed to the next question.

The order of the question and sample appeared in random order, so you will have to pay attention to each of questionnaire. You might see the sample or questionnaire seems repeated but there are not similar what you have answered previously.

There are 72 questions in total, therefore please work quickly through the questionnaires. Do not dwell on any of the question and avoid becoming too analytical. We want your initial, instinctive reactions to each question.

Feel free to observe and touch the physical samples next to you as often as you need whilst you fill in the online questionnaire.
Are there any questions?

(wait ten seconds)

START

As friendly reminder, we want to find out your perception about the specialness of chocolates. We want you to work quickly through the questionnaires, not thinking too much about them but giving your instinctive reactions.

If you have any questions during the session, please do not hesitate to ask. You can start completing the online survey now, good luck and all the best.

Thank you

UNIVERSITY OF LEEDS

Pair Comparison Protocol - Textures
(MEEC 16-050)

Introduction
Hi everyone..

Let me introduce myself. My name is Farzilnizam. I am a Ph.D. student in the Affective Engineering Research Group, Institute of Design Robotics & Optimisation, School of Mechanical Engineering. I will be conducting this research session today.

The objective of this research is to find out how people feel about the quality interior vehicle trims design especially related to texture design. How people perceive about high quality or low quality, comfort and discomfort as well others things related to the subjective feeling of interior vehicle trims.
This study will also investigate and identify new possibilities on how the interior vehicle trim can be improved. We have made special texture samples that represent the interior vehicle trims.

This research concerns the improvement of products development process by address the quality interior trim of production vehicles with the aim to investigate and identify new possibilities on how the interior vehicle trim can be improved.
In front of you, you will find a collection of seven (7) pieces of texture samples that have been labelled as you can see in front of you. In a moment, we will ask you to touch the texture sample and completed questionnaires to say how you feel to touch the texture samples.

When you touch the texture sample, please use your hand that you write with and touch the texture sample in a way that seems most normal and natural to you. Like this.
(Demonstration of touching surface).

Please take a moment now to touch each texture samples. As you do so, please do not say anything or make any remarks about the texture samples to other participants. Make sure you touch each of the seven texture samples. As you touch the texture sample, remember that we are thinking about how they would feel if they were used to design vehicle interiors.
(Show participants to the posters and wait patiently until everyone is ready to continue)

Is everybody happy that they can touch the texture samples and that for the most part, they each feel slightly different?
While you observe the specimen, I would like to brief you how you can complete the research survey today. Simple and easy.
Are you happy to start the experimental survey now? (wait ten seconds)

Brief about survey instruments using PC

What we would like you to do now is the following. This is an online questionnaire, no paper, and pencils. In a moment, you will be asked to complete the questionnaire. The questionnaire is made up of pair of stimuli (texture sample) (Show to computer screen)

When you responding to the question, you need to simply click your answer whether your answer corresponds to the texture on your left or right by clicking the pictures like this. To move to next question, please click next button as shown on the computer screen.

For each questionnaire, we want you to take the texture sample (as shown in computer screen) and touch in the way we showed you before and rate them by clicking your answer on a computer screen using a computer mouse.

(Demonstration of answering the questionnaire using the PC)

We would like you to decide which one of the two textures corresponds best describes by the statement that appears on the screen. If you decide texture on your left, please click a label on your left and if you decide texture on your right, please click the label on your right.

There are **189 comparisons** in total, therefore please work quickly through the questionnaires. Do not dwell on any of the questions and avoid becoming too analytical. We want your initial, instinctive reactions to each question.

Each question illustrates two pairwise texture label (Show to computer screen). **You must observe and touch** the physical samples exactly according to the label that shows on the computer screen only.

You need to answer it before you can move to the next question. In other words, you cannot leave the question blank (without answering) the computer program will not proceed to the next question.

The order of the question and sample appeared in random order, so you will have to pay attention to each of questionnaire. You might see the sample or questionnaire seems repeated but there are not similar what you have answered previously.

Are there any questions?

(wait ten seconds)

START
As a reminder, we want to find out how you feel about these texture surfaces if they were to be used to designing the vehicle interiors, and we want you to work quickly through the questionnaires, not thinking too much about them but giving your instinctive reactions. If you have any questions during the session, please do not hesitate to ask. If you have any questions during the session, please do not hesitate to ask. You can start completing the online survey now, good luck and all the best.

Thank you

Appendix D

Focus Group Questionnaires

Task 1 : General Questions related to short video

In task one, participants were presented two short videos. Participants need to pay your attention to this video to develop your understanding about the research context.

No	Questions
1	What are you hearing people's say about vehicle quality in general, whether good or bad?
2	How it does influence you in making vehicle purchase decision in future?
3	What are the level of perceived quality in order you to accept or reject?

Task 2 : How consumers observe perceived quality – vehicle as subject

How consumers define general understanding about "Low Perceived Quality" & "High Perceived Quality" by visual perception on first eye contact.

In task two, participants were presented with four pair of pictures to response. The pictures consist of two exteriors vehicles and two interiors vehicles which was labelled as (X), (Y), (R) & (S).

No	Questions
1	Tell me between this two vehicles (picture X and Y) which one looks Low Perceived Quality? Would you explain further?
2	Tell me between this two vehicles (picture X and Y) which one looks High Perceived Quality? Would you explain further?
3	Tell me between this two dashboard (pictures R and S) which one looks High Perceived Quality Would you explain further?
4	Tell me between this two dashboard (pictures R and S) which one looks Low Perceived Quality Would you explain further?

Task 3 : Driving Pleasure Experiences – sharing driving experiences and memories

In task three, participants were presented example of driving scenarios on the posters to stimulate adjectives words. Participants were ask to share their driving experience, good or bad stories and memories with their friends, partner, families, kids or anyone.

Example : *"I fetched my girlfriend to watch a movies last weekend, I'd make her surprised, when I drove my new mini cooper. She's really love it, she told me, my car looks great and fun to drive, the seats were really comfy.*

Moderator will stimulate the one of the questions to probes discussion in greater scope as following;

I notice you didn't mention anything about ______________ can you talk about this for me?

1. What are the atmosphere or mood like?
2. What are memorable things you still remember about his/her car when you saw it at first time?
   Example nice color, big sports rims, ribbons on the door handle..
3. How do you find the interiors of the vehicles?
4. How do you find the touch and feeling of the interior?
5. What will you expect more from the car that gives positive driving experience, or pleasant to drive for you?
   Example nice leather instead normal fabrics
6. Do you think the car looks good for different occasion for example, weekdays travelling or long distance travelling
7. Can you tell me a bit more, the negative things or perceptions about the car? Why you don't like ________________? (specific items)

Task 4 : Physical vehicle interior textures (Texture plaques)

In task four, participants were presented seven texture plaques which were labelled as Z1, Y2, X3, W4, V5, U6 and T7 to minimised bias effect.

For this topic, we have numbers of physical texture slabs in your desk as stimuli. Meanwhile, for this section required your understanding to touch and feel the physical textures by using your own internal touch and visual perceptions to judge best textures feelings.

No	Questions
1	Can you tell me which texture slabs is nice to you? Would you explain further?
2	Can you tell me which texture slabs is not nice to you? Would you explain further?
3	Can you tell me which texture slabs is nice to touch? Would you explain further?
4	Can you tell me which texture slabs is not nice to touch? Would you explain further?
5	Can you tell me which texture will make you don't buy the car? Would you explain further?
6	Can you tell me which texture will make your decision to buy the cars? Would you explain further?
7	Can you tell me which texture slabs is very affective to describe Expensive or Premium Looks

	Can you give some justification?
8	Can you tell me which texture slabs is very affective to describe Plasticky Looks (toys-like)” Can you give some justification?
9	Can you tell me which texture slabs is very affective to describe Soft Feel Looks? Can you give some justification?
10	Can you tell me which texture slabs is very affective to describe Best Haptic feelings or best touch & feel? Can you give some justification?

Moderator will stimulate the one of the questions to probes discussion in greater scope as following;

I notice you didn’t mention anything about _____________ can you talk about this for me?

1. Aesthetic point of view – (impressive, high craftsmanship, graceful, simple, neat)
2. Physical Characters – (Big, thick, broad, sharp, smooth, round, geometrics, organic, flat, bumpy, etc)
3. Sensational point of view – (elegant, agile, balance, natural alike, warm, sleek, wet, dry, matt, rich, balance, strong, modern, dynamic, nice contour, elegant, classic, sporty, avant-garde, masculine) etc)
4. Operational point of view – (ergonomics, comfort, easy clean, rigid, slippery, loose, sleek, high grip etc)

End of focus group sessions

Before the sessions end, participants were encourage to give their thoughts, opinion, comments or ideas about the discussions which may probably not cover during the activity.

Moderator wrap up and conclude the focus group session with courtesy thanks and hand out the appreciation token to all the participants for joining the focus group.

The focus group dismissed

Appendix E

Recording transcripts from the focus group

Section 1 : Introduction – Develop the relationship

Participant	Phrases about product experience	Adjective words
1	*I like for wheel drive, I like Jip specifically because the wheel base is not too large, is not too small. Powerful engine, gear box, quite comfortable suspension for off road driving.*	1. Like 2. Not too large 3. Not too small 4. Powerful 5. Comfortable 6. Driving
2	*When comes to cars, I have a big family, so I prefer something spacious that can accommodate 7 peoples. I like to drive 7 seater SUV's (single utility vehicle). I used to drive smaller cars like Mercedes when my family was smaller. Currently in UK I'm driving Citroen Picasso, one because of the price of course being student we have to reduce all the expenses. Because of the 1.6 litter diesel engine everything the cost is cut off. The most important factor is the space. If given the chances I've drive solo, I would prefer something sporty, striking sporty. Journey to here it's depend on the parking space in the University available during school holidays, I will drive otherwise I'll travel by bus.*	1. Spacious 2. Like to drive 3. Smaller 4. Reduce 5. Cut off 6. Important 7. Space 8. Sporty 9. Striking sporty
3	*I use to drive pickup from Chevrolet. I use to work at big farm. I really like small cars. I don't like pickup or big cars. I prefer city cars don't know why, maybe because I'm very short. I feel more comfortable that kind of cars. I like Chevy (Chevrolet), I like Honda, I love Mitsubishi.*	1. Small cars 2. Don't like 3. Big cars 4. Don't know 5. Very short 6. More comfortable 7. I like 8. I love
4	*Related to cars, I use to drive Ford Fiesta, I drove every day to go to class, pick up some friends also to little village of my grandparents. I really to drove under country side its make me feel free. I think my car cool for both to drive in the city but as well between city and small town. Although is not very powerful engine, but if feel confident and it's really comfortable. I prefer normal regular cars. My uncle use to give me Toyota Hilux to go off road and it's amazing. You feel like you're in the tank it's really good.*	1. Feel free 2. Cool 3. Small town 4. Not very powerful 5. Comfortable 6. Regular cars 7. Amazing 8. Feel like 9. Really good

5	*Regards to car I use to drive Toyota Corolla 2005 model. It's very was quite good. But my striking things about the interior was the experience I had when I first time I enter 2005 or 2006 BMW 3 series. The interior was quite cool, when I touch the door it gave me this feel at launch me suit. It looks like a cockpit, I was classy. Later on in 2009 when my uncle bought the new Toyota Avanza I have the privilege to taking of all the packing. I like about the car is metallic color. I like shining, very-very shining feel. I love very bright fog head lamp, especially at night very brilliant.*	1. Quite good 2. Striking 3. Quite cool 4. Feel at launch me suit. 5. Cockpit 6. Classy 7. Privilege 8. Metallic color 9. Shining 10. Very-very shining feel 11. Love 12. Very bright 13. Very brilliant
6	*Regarding to cars, 10 to 7 years ago I was working regularly as Burley park, I don't really remember much but one thing make me absolutely shocking of entering Bentley Continental GT with red interior. Red leather on steering wheel, I was so amazing, the color, the texture, the feelings it was so amazing. I never had an experience until I enter the car interior. Unfortunately I can effort. I really like city cars, I feel comfortable the cars from B and C segment.*	1. Absolutely shocking 2. Red interior 3. Red steering wheel. 4. So amazing 5. The color 6. The texture 7. Really like 8. Feel comfortable

<u>Section 3 : Product experiences questions</u>

Q2 : Can you tell me, what is your first impression about these cars? Which one looks nice and attracts your eyes (X) or (Y)?

Quote by	Phrases about product experience	Adjective words
1	*The while pillar (B pillar) for X is a distraction whereas Y is cool.*	1. Distraction 2. Cool
2	*I agreed with the cool but I think it's really much subjective. The Y seems lack of some sort of column (B pillar) whereas the X is more steady and more safe. The feeling that I get the X is safer cars. If I will to sell, one of these 2 cars, potential customers walk-in in a grey suit in mid-50's or 40's and working with the company, I will showing the X.*	1. Cool 2. Lack 3. More steady 4. More safe 5. Grey suit
3	*It was clear the rim looks nice. I was remembering my friend he was about to retired, he still wears his suit but he bought a sports cars. Young insight, because he want to enjoy the moment.*	1. Rim looks nice 2. Sports cars 3. Young insight 4. Enjoy
4	*It's all depending on personal preferences actually, some people may for sedan, some people go for convertible. We are at certain age where we accept stylish.*	1. Personal preferences 2. Convertible 3. Stylish

Q4 : Can you tell me, between both interior vehicles, which one looks nice and attracts your eye (R) or (S) at first sight?

Participant	Phrases about product experience	Adjective words
4	*S because red colour are so prominent. The screen makes more colourful. R will lack of visual stimulations, simplicity..*	1. Prominent 2. Colourful 3. Simplicity
5	*I would say S is too messy, too many buttons. I just want to have basic cars. But if you look at the photo S, If I will to feel the confident with my driving, because my perception on the steering wheel, the grip is firm.*	1. Too messy 2. Many buttons 3. Basic cars 4. Confident 5. Grip is firm
6	*The trend few years ago to put many buttons as much as possible. The right things to have not many buttons but off course with metallic color, texture, this glowing and lights. The R is looks the cars from 90's.*	1. Many, much 2. Right things 3. Metallic colour 4. Glowing, lights 5. 90's
2	*The silver detail that separate from the plastic is really cool. I want everything on the steering wheel, that sense you have everything and control. It's better to have group from female, because the opinion is totally different from us.*	1. Silver details 2. Really cool 3. Control 4. Better
1	*I have a friends say, they would never buy a cars that doesn't have the flashy dashboard. R is very dull.*	1. Flashy 2. Very dull

Q6 : What do you think about cars advertisements in magazines, websites and TV's advertisements, does it influence and convince you in making a good purchase decision?

Participant	Phrases about product experience	Adjective words
1	*I like the car but I don't like the exploitation.*	Adjective Words - Exploitation - Masculine - Power - Respective - Nailing - Big cars - Famine factors - Carefully - Wrong - Power - Changing - Very poor - Performance
5	*What is tried to convey is masculine power, if you want to be respective. Human is nailing to the machine.*	
6	*Human nailing to the machine or women only nail to man who has big cars. If I like to be somebody, I don't need car to show myself.*	
2	*if this is famine factors in sells, but it should be use carefully, in this situation is completely wrong.*	
4	*James bond, they don't use women as the tools. The women with brain women with power, shows James bond, the perception is changing.*	
5	*Very poor performance of advertising.*	

Q7 : Do you watch favourite TVs shows about cars? How might they benefit you in changing your perceptions about interior quality motor vehicles?

Participant	Phrases about product experience	Adjective words
4	*Top Gear's has really shows the strength and performance of various Cars. I think they would really influence my decision of purchase. I like the way the car spin, they really shown what the cars can do in reality.*	- Strength - Performance - Influence - Spin - Can do - Reality - Stressing - To the limits - Taste - Touch - Power - Extreme - Useful - Easily - Agree - Awareness - Interest - Test - Basic - Good
5	*When the stressing the cars to the limits, but no about the interior because the process of purchasing the cars is the following, first you have to taste it get inside of the cars and you touch by yourself. But the opinion of the power of the cars or driving in extreme conditions of course is useful.*	
6	*But sometimes they do comment on the interior*	
2	*Yes, but the comment you can't check by yourself easily*	
1	*I agree with you but sometimes, this kind of free advertisement for Cars, can basically attires you create awareness about people, create Interest for you to eventually go to showroom and test the car. Without such of advertisement you may not have the basic ideas whether the cars is good or not.*	

Q7 : Do you watch favourite TVs shows about cars? How might they benefit you in changing your perceptions about interior quality motor vehicles?

Participant	Phrases about product experience	Adjective words
4	*Eventually when you buy a car is a very big purchase, you spent a lot of amount. You see the offer but you always sit down with your paper and then you start make the figures. It's so attractive (promotion) especially those general workers they not have higher qualification but it's doesn't works most of people.*	- Big purchase - A lot - Amount - Attractive - Qualification - Doesn't works - Price - Value - Savings - Look
5	*Buying cars is not like buying cloths. End of the day, is a price matter and trade-in value and whether the cars is fuel savings. The current Oils price conditions people will look something.*	

Q9 :What is your dream perceived level of quality do you like to be for your affordable cars and why do you think is important during your purchase decision?

Participant	Phrases about product experience	Adjective words

4	*if the interior is nice, it will help you to be comfortable inside the car.*	- Nice - Comfortable

Q10 : Can you tell us, about your friend's or family's cars, how do you experience them? What is special about the vehicles that gives you pleasant driving experience?

Participant	Phrases about product experience	Adjective words
2	*My mum has a Nissan Exterra, I was on motorway, I don't even know the performance of the cars. I don't drive very fast, I found myself in motorway is was like very free, I feel like was in the Jet. It was cool. The alloy (rims) is quite robust. The Toyota Camry in 2012, the feel of Fabrics or leather, its felt velvety tiny force that's really comfortable.*	- Very fast - Very free - Cool - Robust - Velvety - Comfortable

Q11 : Every car has design with special interior dashboard or instrument panel to give different driving experiences. The design can affect your driving workload and your attention while driving. How do you observe this issue when deciding to buy the car?

Participant	Phrases about product experience	Adjective words
4	*I prefer buttons, those kind of things (touch screen) its required your attentions. The classical knobs, you will feel you're turning the knob, so enough reaction feedback from what you're doing but you still have eyes on the road (focus).*	- Attentions - Classical knobs - Turning - Feedback - Focus - Important - Comfort - Easily
5	*The steering wheel is very important nowadays, I don't want anything else, I won't buy the car without anything (controls button) on the steering wheel.*	
6	*I have Peugeot with the radio control on steering wheel, once started to using that, I realised how comfort is, you can adjust the volume easily.*	

Q12 : Which type of fuel efficient cars do you think are best for you and you may seriously consider buying? What do you consider in this decision?

Participant	Phrases about product experience	Adjective words
3	*I would say I the car would be drive for an hour per day, I will into the Account of fuel efficiency. However I live in small town and drive daily for 5 minutes, I prefer 3 litter engine quite powerful engine that consumes a lot of petrol like water but I don't care because I don't drive that much. When I do driving, I do want to feel the power. Electric I would say no, because the supporting the technology seems do not much easy to obtained. Perhaps I will go for hybrid because if feel safe, if you run out of electricity you still petrol.*	- Small - Powerful engine - Consumes - Don't care - Don't drive - Much - Feel the power - Not much easy - Feel safe

Q13: Do you agree that interior design with rich touch and feel materials would affect your driving pleasure? How will the rich touch make you happy?

Participant	Phrases about product experience	Adjective words

4	*May you have to spend £1k - £2k for that, but is worth it.*	- Worth-it - Elegant - Red feel - Touch - Don't often - Visually convey - Sensational - Softness - More important
5	*You can see how elegant the red feel of the seat.*	
6	*I would say the steering wheel, the shift stick and the door, it doesn't matter how the feel when touch, for the dashboard, you don't often to touch them, but when you look at it, it should visually convey a sensational of softness.*	
1	*Visually of the dashboard is probably more important than touch.*	

Q14 : Do you agree that, rich touch and feel materials would affect your driving behaviour? How can this affect and help your driving workload and your attention while driving?

Participant	Phrases about product experience	Adjective words
4	*When you're comfortable, you can spend more time on driving.*	- Comfortable - Slightly affect - Mood - Not relaxing - Quite soft
5	*I would say it will slightly affect your mood. The red leather is not relaxing Colour, but the feelings from here looks the car seat quite soft. It does affect my experience.*	

Q15 : How long do you normally own a car? What makes you keep your car for a certain time?

Participant	Phrases about product experience	Adjective words
4	*as long we have relationship with the cars*	- Relationship - Big
5	*The steering wheel is very important nowadays, I Most of people change car, my brother change his car every 3 years.*	
6	*it's all depend how big is your pocket.*	

Q16 : What do you want a car to say about you because you own it?

Participant	Phrases about product experience	Adjective words
4	*I wouldn't buy car like this.*	- Prestige's - Bold - Performance - Appearance - Perform - Comfortable - Not comfy - Dress like supercars
5	*I would never buy car without black, black is prestige's. I like bold statement.*	
6	*I don't really care what they thinking about me, I care more about performance than the appearance. The type which is built to perform and to be comfortable. It's not comfy but looks like a race car.*	
1	*I really like normal car dress like supercars.*	
3	*Anyway you no need your car to say about you but eventually the car you choose is telling something about you.*	

Appendix F

SD and LS Questionnaires

Respondent 99

Use this questionnaire for SAMPLE 1

I would describe this material as feeling ...

revitalising	○ ○ ○ ○ ○ ○ ○	not revitalising
not plasticky	○ ○ ○ ○ ○ ○ ○	plasticky
stimulating	○ ○ ○ ○ ○ ○ ○	not stimulating
not useful	○ ○ ○ ○ ○ ○ ○	useful
edgy	○ ○ ○ ○ ○ ○ ○	not edgy
not grippy	○ ○ ○ ○ ○ ○ ○	grippy
not abrasive	○ ○ ○ ○ ○ ○ ○	abrasive
not sporty	○ ○ ○ ○ ○ ○ ○	sporty
effortless	○ ○ ○ ○ ○ ○ ○	not effortless
not eye-catching	○ ○ ○ ○ ○ ○ ○	eye-catching
velvety	○ ○ ○ ○ ○ ○ ○	not velvety
not comfortable	○ ○ ○ ○ ○ ○ ○	comfortable
not bumpy	○ ○ ○ ○ ○ ○ ○	bumpy
rugged	○ ○ ○ ○ ○ ○ ○	not rugged
stunning	○ ○ ○ ○ ○ ○ ○	not stunning
safe	○ ○ ○ ○ ○ ○ ○	not safe
enjoyable	○ ○ ○ ○ ○ ○ ○	not enjoyable
easy-handling	○ ○ ○ ○ ○ ○ ○	not easy-handling
nice quality	○ ○ ○ ○ ○ ○ ○	not nice quality
timeless design	○ ○ ○ ○ ○ ○ ○	not timeless design

My overall feeling for this material is

like	○ ○ ○	dislike

Use this questionnaire for Sample 1

125

Tick one box against each statement to indicate the extent to which you agree or disagree that it describes the texture.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
The feel of this texture on my steering wheel or switches would help me keep my eyes on the road without distraction.	O	O	O	O	O
If I gripped a steering wheel which had this texture, it would not be too slippery.	O	O	O	O	O
This texture has a sporty look and feel.	O	O	O	O	O
Touching this texture feels pleasant.	O	O	O	O	O
I would expect to see this texture with a good touch and feel in a reasonably priced car.	O	O	O	O	O
Touching this texture makes me feel warm.	O	O	O	O	O
Touching this texture would make me feel connected when operating the switches in the vehicle.	O	O	O	O	O
With this texture, I would feel comfortable inside the car.	O	O	O	O	O
When I touch this surface I get a sensation of luxury.	O	O	O	O	O
Touching this texture is relaxing.	O	O	O	O	O
I have the impression this texture is modern and contemporary-looking.	O	O	O	O	O
The feel of this texture would help me feel confident with my driving.	O	O	O	O	O

Use this questionnaire for Sample 1

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Vehicle controls with this texture would give good feedback when shifting, pulling, pushing, turning and rotating.	○	○	○	○	○
This texture looks nice quality.	○	○	○	○	○
The look of this texture makes me want to touch it straight away.	○	○	○	○	○
I have the impression that this texture would make my car feel spacious and neat.	○	○	○	○	○
If I gripped a steering wheel which had this texture, it would feel very safe.	○	○	○	○	○
Before I touch this texture, I can see that it would feel grippy.	○	○	○	○	○
This texture does not look overly cheap and plasticky.	○	○	○	○	○
With this texture, I would be able to operate the controls without needing to look.	○	○	○	○	○