Applications of Rasch analysis in consumer research for new food product development

Zheng Li

Submitted in accordance with the requirements for the degree of

Doctor of Philosophy

University of Leeds

School of Food Science and Nutrition

February 2019

The candidate confirms that the work submitted is his own and that appropriate credit has been given where reference has been made to the work of others.

Clarification of contributions

The case studies described in this thesis has been carried out by a research team led by Dr Peter Ho.

The contributions of the other members of the team to the case studies are listed below:

Dr Peter Ho provided guidance for all case studies.

Miss Clara Philippe contributed to data collection for case studies I and II.

Mrs Ifeanyi Okojie contributed to the instrument development and data collection for case studies III and IV.

Acknowledgements

I would like to first thank my family, especially my mother, without whose support I would not have been able to complete this PhD project.

I would also like to thank my main supervisor Dr Peter Ho for offering incredible guidance and suggestions during the research. I greatly enjoyed the time we spent together discussing and debating the academic topics.

My deep appreciation goes out to my colleagues who work with me in the same research team: Ify, Clara, Lily, Hanis and Scarlett. I would particularly like to thank Ify for all of the help over years.

I am indebted to Professor Michael Morgan, from whom I learned how to conduct proper research. He is also the best teacher I have ever met.

I am grateful to the school administration team and lab technicians for supporting all aspects of this research, including Angela, Ian, Miles and many unspecified others.

I owe a special thanks to my examiners Dr Brian Henson and Ms Lauren Rogers for providing me precious suggestions to make this thesis a better one.

Many thanks to my beloved friends from China, UK, and other countries. I will never forget all the great memories we created together. In addition, as a promise, I still have a Guan Dan game to win with my poker partner Dr Zhihua Li against the team of Dr Yi Li and Professor Lingzhao Wang in the future.

I am very lucky to share the office space with some lovely people: Liam, Louise, Ng’andwe, Sadia, Sam, Yue and Zainiu. No words would be enough to express my gratitude to them for brightening up my days.

Finally, the completion of the case studies would never have been possible without the participation of hundreds of research volunteers. I owe everyone a big “thank you”.

Abstract

Introduction

The classical test theory (CTT) and its extensions are the predominant measurement and data processing approaches in current consumer research areas. However, the CTT-based approaches suffer from several theoretical drawbacks such as requiring interval data and lack of quality control procedures. To overcome these drawbacks, an alternative method Rasch analysis, which stands for the analysis using a family of parametric probabilistic response models, can be used. Although Rasch analysis has been broadly used in education and health research areas, it is yet to be used extensively in food-related research and new food product development.

The aims of this research were to demonstrate the benefit which can be obtained by applying Rasch analysis to consumer research for new food product development and to explore the application of Rasch analysis in the development and validation of an instrument which can be used in food-related consumer research tasks

Research activities and outcomes

To achieve the aims of the research, four case studies were conducted in the following order:

Case study I compared the difference between CTT approach and Rasch analysis in the evaluation of a survey of 269 respondents using an existing instrument Health and Taste Attitude Scales. The results indicated that compared to CTT approach, Rasch analysis could identify more meaningful underlying structure of the instrument and interpretable reliability statistics.

Case study II employed the Many-Facet Rasch model in a sensory study for modelling a composite overall liking measure from 8 sensory attributes and a holistic measure from a single overall acceptability item. The psychometric properties of the two models were compared. The ability of the two measures to differentiate between the overall liking of product were also compared. The results suggested that the composite measure has greater ability of differentiating products.

Case study III developed and validated a set of instruments in relation to ready meal consumption under the guidance of Rasch analysis. The instruments can be used for measuring consumers' satisfaction attitudes, decision making patterns and willingness to consume ready meals in different contextual situations. The hierarchical rank order of the items also provided information associated with new food product development opportunities.

Case study IV developed and validated a sensory instrument for benchmarking test on beef lasagne ready meals under the guidance of Rasch analysis. The result verified the expectation of product ranking. The information also reflected panellists' particular needs for product improvement.

Conclusion

Rasch analysis can overcome the limits of CTT approach, improving the quality of measurement in consumer research for new food product development practice. It should be applied to more food-related area.

Table of Contents

Clarification of contributions iii

Acknowledgements iv

Abstract v

Table of Contents...vii

List of Figures xvi

List of Tables..xix

List of Abbreviations xxiii

Chapter 1 Introduction 24

1.1 Definition of measurement and level of measurement 25

1.2 The building blocks of measurement 27

1.2.1 Construct...27

1.2.2 Item response 28

1.2.3 Outcome space 28

1.2.4 Measurement model 28

1.3 The measurement issues in current consumer research for new food product development..29

1.3.1 The misuse of ordinal rating scales...29

1.3.2 The selection between individual measurement and composite measurement 31

1.3.2.1 Individual measurement...31

1.3.2.2 Composite measurement..33

1.4 Applying Rasch analysis to reduce the issues related to the measurement in consumer research for new food product development..35

1.4.1 What is Rasch analysis 35

1.4.2 How can Rasch analysis reduce the issues related to the measurement in consumer research for new food product development?...37

1.4.3 Previous applications of Rasch analysis in consumer research for new food product development..39

1.5 Overview of the research...40

1.5.1 Research aims 40

1.5.2 Thesis structure..40

Chapter 2 Literature review: A comparison between classical test theory (CTT) and Rasch analysis...42

2.1 An overview of classical test theory models and Rasch models.....42

2.1.1 Classical test theory models and the extensions...................42

2.1.1.1 CTT model for individual measurement.........................42

2.1.1.2 CTT model for composite measurement 43

2.1.1.3 The extensions of CTT models.....................................44

2.1.2 The family of Rasch models...46

2.1.2.1 Dichotomous Rasch model...47

2.1.2.2 Polytomous Rasch model...47

2.1.2.3 The Rasch model for more than two measurement facets: the Many-Facet Rasch (MFR) model.............................49

2.2 Estimation of model parameters..49

2.2.1 Estimation of CTT model parameters...................................49

2.2.2 Estimation of Rasch model parameters 50

2.2.2.1 Normal Approximation Algorithm (PROX)......................51

2.2.2.2 Joint Maximum Likelihood Estimation (JMLE) 52

2.3 Quality control procedures...53

2.3.1 Rating scale category effectiveness.....................................53

2.3.1.1 Basic concept of scale categories and category thresholds 53

2.3.1.2 Evaluating of rating scale category effectiveness..........56

2.3.1.3 Optimising rating scale category effectiveness..............61

2.3.2 Local independence 61

2.3.2.1 Definition of local independence...................................61

2.3.2.2 Dimensionality test in CTT...62

2.3.2.3 Tests of unidimensionality in Rasch analysis.................64

2.3.2.4 Test of local item independence 68

2.3.3 Fit..68

2.3.3.1 Fit statistics..68

2.3.3.2 Resolving misfitting item issue.......................................71

2.3.3.3 Considerations on misfitting persons 72

2.3.4 Differential item functioning 72

2.3.4.1 Statistical DIF analysis in CTT.......................................73

2.3.4.2 Statistical DIF analysis in Rasch analysis......................74

2.3.4.3 Resolving DIF 76

2.3.4.4 When should DIF be examined 77

2.3.5 Reliability/Separation 77

2.3.5.1 Reliability in CTT 77

2.3.5.2 Reliability indices in Rasch analysis 80

2.3.6 Standard error of measurement (SE_m) 83
2.3.6.1 SE_m in CTT 83
2.3.6.2 SE_m in Rasch analysis 83
2.3.7 Chi-square tests for fixed effect and random effect in MFR models 85
2.3.7.1 Chi-square test for fixed (all-same) effect 85
2.3.7.2 Chi-square test for random (normal) effect 85
2.3.8 Remarks 86
Chapter 3 Case study I: Comparison between classical test theory approach and Rasch analysis in evaluation of consumer survey – a revisit of the Health and Taste Attitude Scales (HTAS) 87
3.1 Introduction 87
3.2 Instrument and sampling procedures 87
3.2.1 Instrument 87
3.2.2 Participants and sampling procedures 88
3.2.3 Data collection and pre-treatment 90
3.3 Analysis by CTT approach 91
3.3.1 Methods 91
3.3.2 Results 91
3.3.2.1 Structure of the instrument 91
3.3.2.2 Reliability 96
3.3.2.3 Respondents’ total score on each subscale 96
3.4 Rasch analysis 97
3.4.1 Methods 97
3.4.2 Results 98
3.4.2.1 Category effectiveness 98
3.4.2.2 Tests of unidimensionality and local item independence 101
3.4.2.3 Test of individual fit 104
3.4.2.4 Tests for differential item functioning (DIF) 106
3.4.2.5 Wright maps 107
3.4.2.6 Reliability 112
3.5 Differences on the estimation of person parameters between CTT and Rasch analysis 113
3.5.1 Methods 114
3.5.1.1 Comparing the stability of estimation 114
3.5.1.2 Comparing the respondents’ estimates 114
3.5.2 Results 115

3.5.2.1 The stability of estimation of person parameters within different frameworks 115

3.5.2.2 Comparisons between the person estimates obtained by CTT and Rasch analysis 118

3.6 Differences on health and taste attitudes between groups by gender and age 120

3.6.1 Methods 120

3.6.2 Results 121

3.6.2.1 Effect of gender 121

3.6.2.2 Effect of age 122

3.7 Discussion 126

3.7.1 CTT approach vs. Rasch analysis.......................................126

3.7.1.1 Dimensionality of the instrument..................................126

3.7.1.2 Reliability 127

3.7.1.3 Estimation of person parameter...................................128

3.7.1.4 Group effects 129

3.7.2 Effect of collapsing rating scale (Rasch analysis) 129

3.7.3 Resolving DIF..130

3.7.4 Conclusion 130

Chapter 4 Case study II: Measuring the overall liking in sensory acceptability test using Many-Facet Rasch model – a comparison between composite measurement and individual measurement 131

4.1 Introduction..131

4.1.1 Aim..131

4.1.2 Research hypothesis..131

4.2 Participants and sampling procedures 131

4.2.1 Sample selection...132

4.2.2 Sensory attributes and rating scale 132

4.2.3 Experimental design...134

4.2.4 Administration 134

4.3 Rasch analysis procedures 136

4.3.1 Rating scale category effectiveness....................................136

4.3.2 Tests of unidimensionality and local item independence 136

4.3.3 Global model fit 137

4.3.4 Test of individual fit 137

4.3.5 Wright maps..137

4.4 Statistical analysis procedures 137

4.4.1 Reliability..137

4.4.2 Chi-square tests for fixed effect and random effect 137

4.4.3 Modelling of the replicate measures...................................138

4.4.4 Multiple comparisons 138

4.5 Results 139

4.5.1 Rating scale category effectiveness..................................139

4.5.1.1 Optimising rating scale by collapsing categories 140

4.5.2 Tests of unidimensionality and local item independence of Attribute facet 146

4.5.2.1 Tests of unidimensionality - original rating scale 146

4.5.2.2 Tests of unidimensionality - collapsed rating scale......146

4.5.2.3 Test of local item independence 147

4.5.3 Global model fit 148

4.5.4 Test of individual fit 149

4.5.4.1 Product fit 149

4.5.4.2 Attribute fit 149

4.5.4.3 Panellist fit 150

4.5.5 Wright maps 151

4.5.6 Rasch separation and reliability statistics...........................154

4.5.7 Chi-square statistics on fixed effect and random effect.......154

4.5.8 Multiple comparisons on the overall liking of the products ..156

4.6 Discussion 159

4.6.1 Comparison between the composite measurement using attribute ratings and the individual measurement using the overall acceptability ratings 159

4.6.1.1 Model fit 159

4.6.1.2 Separation/Reliability 159

4.6.1.3 Determination of the difference in overall liking of products 160

4.6.2 Effect of collapsing rating scale..160

4.6.3 Conclusion 160

Chapter 5 Case study III: Application of Rasch analysis in instrument development and validation for consumer insights research.....162

5.1 Introduction...162

5.1.1 Aim...162

5.1.2 Rasch analysis and measurement development.................162

5.1.2.1 Rasch analysis obtains a fundamental basis for construct conceptualisation...162

5.1.2.2 Rasch analysis produces test-free person measures and sample-free item measures at individual levels 163

5.1.3 Research object – Ready meal 163

5.1.3.1 The definition of ready meal 163

5.1.3.2 Previous research on ready meal consumption 164

5.1.4 Overview of the study...164

5.2 Defining the building blocks of measurement 167

5.2.1 Defining the initial construct 167

5.2.1.1 Satisfaction attitudes 167

5.2.1.2 Product criteria 167

5.2.1.3 Consumption situations 168

5.2.2 Designing the format of the items with consideration to item responses and outcome space 169

5.2.3 Selecting measurement model...170

5.3 Developing the instruments for measuring the construct.............170

5.3.1 Generating the initial item pool via focus group studies 170

5.3.2 Pilot testing and the composition of the survey questionnaire 171

5.4 Sampling 173

5.5 Data analysis procedures...174

5.5.1 Rasch analysis...174

5.5.2 Segmenting respondents into groups using cluster analysis174

5.5.3 Exploring the relationship between respondent segments and consumption frequency 175

5.6 Results 175

5.6.1 Rating scale category effectiveness...................................175

5.6.2 Tests of unidimensionality and local item independence 177

5.6.2.1 Unidimensionality and local item independence of the satisfaction attitudes instrument 177

5.6.2.2 Unidimensionality and local item independence of the product criteria instrument..178

5.6.2.3 Unidimensionality and local item independence of the consumption situations instrument 179

5.6.3 Test of individual fit 183

5.6.4 Wright maps...189

5.6.4.1 Wright maps of satisfaction attitudes related constructs 189

5.6.4.2 Wright maps of product criteria constructs...................190

5.6.4.3 Wright map of consumption situations.........................190

5.6.5 Reliability..192

5.6.6 Consumer segmentation 193

5.6.6.1 Consumer segmentation based on satisfaction attitudes 193

5.6.6.2 Consumer segmentation based on consumers' degree of consideration on product criteria.................................197

5.6.6.3 Distribution of gender and age in the segments 199

5.6.7 Relationship between consumer segments and the consumption frequency of three type of meals....................200

5.7 Discussion 202

5.7.1 Consumer insights detected from this study 202

5.7.1.1 Consumers' satisfaction attitudes towards ready meals. 202

5.7.1.2 Consumers' decision making patterns and relative importance of product criteria of ready meals.............203

5.7.1.3 The willingness of consuming ready meals at particular contextual situations 204

5.7.2 Consumer segmentation – Rasch vs. CTT 204

5.7.3 Rasch analysis provides a number of additional quality control procedures that can govern the development and validation of the instrument. 205

5.7.4 Summary...205

Chapter 6 Case study IV: Application of the Rasch analysis in instrument development and validation for sensory product benchmarking test on beef lasagne ready meal products 206

6.1 Introduction...206

6.2 Developing the instrument...206

6.2.1 Overview of the instrument development 206

6.2.2 Defining the construct 207

6.2.3 Defining the item responses and outcome space...............207

6.2.4 Defining the measurement model 208

6.2.5 Selecting distinct products for sensory evaluation...............208

6.2.6 Generating the initial attribute pool......................................209

6.2.7 Refining the attribute pool via consumer interviews 209

6.2.8 Composing of the final questionnaire 212

6.3 Procedures of instrument validation using benchmarking test......213

6.3.1 Experimental work...213

6.3.2 Rasch analysis 213

6.3.3 Statistical analysis...214

6.4 Results 215
6.4.1 Rating scale category effectiveness...................................215
6.4.2 Tests of unidimensionality and local item independence 217
6.4.3 Global model fit 219
6.4.4 Test of individual fit 219
6.4.5 Wright maps...221
6.4.6 Reliability..222
6.4.7 Fixed effect and random effect..223
6.4.8 Multiple comparisons on product preference 223
6.4.8.1 Residual analysis..224
6.4.8.2 Multiple comparisons using both parametric and non-parametric procedure 224
6.5 Discussion 226
Chapter 7 Summary and further discussion...228
7.1 Summary of the research 228
7.2 Contribution of knowledge 230
7.3 Additional issues identified from the case studies 231
7.3.1 Optimal number of categories in rating scales 231
7.3.2 Issue associated with local item independence 232
7.3.3 Sample size consideration 232
7.3.3.1 Sample size consideration for constructing stable measures for exploratory research 232
7.3.3.2 Sample size consideration for conducting definitive statistical tests 233
7.3.4 Missing data 233
7.3.4.1 The mechanisms of missing data 233
7.3.4.2 Impact of missing data on Rasch analysis estimation .234
7.3.4.3 The technical considerations when applying Rasch analysis with incomplete data.....................................234
7.3.5 Inconsistency in the person-item response pattern 235
7.4 Disadvantages of applying Rasch analysis in consumer research236
7.5 Limitations of the research 238
7.6 Conclusion..238
Reference 239
Appendix A: The latest version of Stage-Gate® model adapted from (Cooper and Sommer, 2016)...253
Appendix B: The main estimation methods of Rasch model parameters...254

Appendix C: The scree plots obtained from parallel analysis (case study I) Parallel analysis on health-related items Parallel analysis 255

Appendix D: The initial attribute pool (case study IV) 256

Appendix E: The refined attribute pool (case study IV)...................259

List of figures

Figure 1.1 Four building blocks of measurement adapted from Wilson (2004) 27

Figure 1.2 Individual measurement and composite measurement adapted from Martinez-Martin (2010) 32

Figure 1.3 An example of Wright map 36

Figure 1.4 The construct modelling approach within the framework of Rasch analysis adapted from Wilson (2004) 37

Figure 1.5 Transformation of raw score to Rasch measure 38

Figure 2.1 Category probability plot and Rasch-Andrich thresholds 54

Figure 2.2 Expected score ogive and Rasch half-point thresholds 55

Figure 2.3 Cumulative probability curve and Rasch-Thurstone thresholds 56

Figure 2.4 Category probability plot and disordered Rasch-Andrich thresholds 59

Figure 2.5 Standardised residual variance in eigenvalue units 65

Figure 2.6 Correlations between item clusters on first PCAR Contrast 66

Figure 2.7 Standardised residual loadings for Item on first PCAR Contrast 67

Figure 2.8 An example of DIF item 73

Figure 3.1 Individual MSA of health-related items 92

Figure 3.2 Individual MSA of taste-related items 92

Figure 3.3 Tests of unidimensionality of HTAS (original 7-point rating scale) 102

Figure 3.4 Tests of unidimensionality of HTAS (collapsed 4-point rating scale) 103

Figure 3.5 Wright map based on health-related subscales with original rating scale 109

Figure 3.6 Wright map based on health-related subscales with original rating scale 110

Figure 3.7 Wright map based on health-related subscales with collapsed rating scale 111

Figure 3.8 Wright map based on taste-related subscales with collapsed rating scale 112

Figure 3.9 Comparisons between the SE_m estimated by CTT (in red) and the model SE estimated by Rasch analysis on the three health-related subscales. 116

Figure 3.10 Comparisons between the SE_m estimated by CTT (in red) and the real SE estimated by Rasch analysis on the three health-related subscales 117

Figure 3.11 Comparisons between the raw scores and Rasch measures on the three health-related subscales....119

Figure 3.12 Difference between Females (n=194) and Males (n=75) groups for Health and Taste related subscales 124

Figure 3.13 Difference between 16~24 (n=114), 25~34 (n=68) and 35+ (n=87) age groups for health and taste related subscales....125

Figure 4.1 Sensory booth....135

Figure 4.2 Sensory booth....135

Figure 4.3 Category probability plot - Rasch ATTRIBUTES 144

Figure 4.4 Category probability plot - Rasch OA 145

Figure 4.5 Wright maps for Rasch ATTRIBUTES with original rating scale 152

Figure 4.6 Wright maps for Rasch ATTRIBUTES with collapsed rating scale....152

Figure 4.7 Wright maps for Rasch OA with original rating scale....153

Figure 4.8 Wright maps for Rasch OA with collapsed rating scale 153

Figure 5.1 Three initial constructs defined at the beginning of the study 168

Figure 5.2 Tests of unidimensionality of satisfaction attitudes instrument 180

Figure 5.3 Tests of unidimensionality of product criteria instrument.181

Figure 5.4 Tests of unidimensionality of consumption situations instrument 181

Figure 5.5 Wright maps for the four satisfaction-related constructs ..191

Figure 5.6 Wright maps for product criteria constructs and consumption situations construct 192

Figure 5.7 Dendrogram of DIANA clustering on respondents' measures of the four satisfaction attitudes instruments....195

Figure 5.8 Dendrogram of DIANA clustering on respondents' raw mean scores of the four satisfaction attitudes instruments 195

Figure 5.9 Dendrogram of DIANA clustering on respondents' Rasch measures of the product criteria instruments....197

Figure 5.10 Dendrogram of DIANA clustering on respondents' raw mean scores of the product criteria instruments....198

Figure 5.11 Multiple comparisons on consumption frequency of three types of meals by segments based on respondents satisfaction levels....201

Figure 5.12 Multiple comparisons on consumption frequency of three types of meals by segments based on respondents' consideration levels..201

Figure 5.13 Multiple comparisons on consumption frequency of three types of meals by gender or age 202

Figure 6.1 Construct and orientation of the measurement elements defined in this study...207

Figure 6.2 Examples of the attribute cards..210

Figure 6.3 Examples of beef lasagne ready meal product tested in this study 211

Figure 6.4 An example of the computer interface of the benchmarking test 213

Figure 6.5 Category probability plot of the collapsed 4-point scale....215

Figure 6.6 Wright map of the benchmarking test.................................222

List of Tables

Table 1.1 The four levels of measurement (Stevens, 1946) and their basic properties 26

Table 1.2 Some examples of ordinal rating scales used in consumer research for new food product development 29

Table 1.3 The assignment of numbers (raw score) to 7-point Likert scale 30

Table 1.4 Previous applications of Rasch analysis in consumer research 39

Table 2.1 An overview of Rasch models 46

Table 2.2 Criteria for diagnosis of the rating scale category effectiveness – adapted from Linacre (2002a) 57

Table 2.3 Example of probablematic rating scale with unclear categories 58

Table 2.4 Minimum Rasch-Andrich threshold advances (Linacre, 2006) 60

Table 2.5 Interpretation of disattenuated correlation coefficient (Linacre, 2014c)....66

Table 2.6 Guidelines for the interpretation of MNSQ and ZSTD values, adopted from Linacre (2002b)....71

Table 2.7 Major types of reliability indices studied in CTT (adapted from DeVellis, 2011)....78

Table 3.1 Health-related subscales and items in original HTAS instrument 89

Table 3.2 Taste-related subscales and items in original HTAS instrument 89

Table 3.3 Distribution of respondents by gender and age 90

Table 3.4 Results of KMO test and Bartlett's test of sphericity on raw scores of HTAS 92

Table 3.5 Factor loadings of health-related items with the 4-factor solution....94

Table 3.6 Factor loadings of health-related items with the 3-factor solution....94

Table 3.7 Factor loadings of taste-related items under the 5-factor solution....95

Table 3.8 Factor loadings of taste-related items under the 4-factor solution....95

Table 3.9 The Cronbach's alpha of each subscale (and its 95%CI)....96

Table 3.10 Summary of respondents' total score on each subscale.....96

Table 3.11 Statistics for the original 7-point rating scale.......................99

Table 3.12 Statistics for the collapsed 4-point rating scale 100

Table 3.13 Measure, model SE and outfit MNSQ statistics of items in health-related subscales..104

Table 3.14 Measure, model SE and outfit MNSQ statistics of items in taste-related subscales 105

Table 3.15 Counts and proportion of misfitting respondents..............106

Table 3.16 DIF detection by gender group..107

Table 3.17 Rasch reliability statistics of each subscales.....................113

Table 3.18 Significant cases reported from independent t-tests between the respondents' estimates obtained by CTT and Rasch analysis (total number=269)...120

Table 3.19 Comparisons of raw total scores or Rasch measures between gender groups on each subscale...122

Table 3.20 Comparisons of raw total scores or Rasch measures between age groups on each subscale...123

Table 4.1 24 Products evaluated in this study......................................133

Table 4.2 Sensory attributes evaluated in this study............................133

Table 4.3 Example of the crossover design..134

Table 4.4 Statistics for the original 11-point hedonic scale for different MFR-RS models 141

Table 4.5 Effects of collapsing categories to Rasch-Andrich thresholds 142

Table 4.6 Statistics for the collapsed 5-point rating scale categories for different MFR-RS models..143

Table 4.7 Tests of unidimensionality 147

Table 4.8 The item pairs that exhibited potential local item dependence 148

Table 4.9 Counts and proportion of extreme residuals........................148

Table 4.10 Overview of product fit for all models 149

Table 4.11 Measure, standard error (SE) and outfit MNSQ statistics of Attribute facet for all models 150

Table 4.12 Counts and proportion of misfitting panellists for all models 150

Table 4.13 The estimates of Rasch reliability indices for all models ..154

Table 4.14 Chi-square statistics on fixed effect....................................155

Table 4.15 Chi-square statistics on random effect 155

Table 4.16 Results of ANOVA, Kruskal-Wallis test, and the residual analysis for ANOVA assumption based on the estimates of refitted model 157

Table 4.17 Multiple comparisons on the overall liking of the products 158

Table 5.1 Previous research on ready meal consumption 166

Table 5.2 Designated ratings scales used in this study 169

Table 5.3 Classification of focus group participants and the demographic distribution 171

Table 5.4. The procedures and topics of the focus group studies 172

Table 5.5 The composition of the survey questionnaire 173

Table 5.6 Distributions of gender and age 173

Table 5.7 Rasch-Andrich thresholds and the effect of collapsing categories 176

Table 5.8 Item pairs that had a residual correlation equal to or greater than 0.30 182

Table 5.9 The proportion of misfitting respondents 183

Table 5.10 The measure, standard error (SE) and outfit MNSQ statistics of items within the instrument about consumers' satisfaction towards product purchase properties of ready meals 184

Table 5.11 The measure, standard error (SE) and outfit MNSQ statistics of items within instrument about consumers' satisfaction towards the health benefit properties of ready meals....................185

Table 5.12 The measure, standard error (SE) and outfit MNSQ statistics of items within instrument about consumers' satisfaction towards product consumption properties of ready meals 186

Table 5.13 The measure, standard error (SE) and outfit MNSQ statistics of items within the instrument about consumers' satisfaction towards product characteristics of ready meals 187

Table 5.14 The measure, standard error (SE) and outfit MNSQ statistics of items belong to intrinsic criteria instrument....................188

Table 5.15 The measure, standard error (SE) and outfit MNSQ statistics of items within extrinsic criteria instrument....................188

Table 5.16 The measure, standard error (SE) and outfit MNSQ statistics of items within consumption situations instrument 189

Table 5.17 Rasch reliability statistics of the instruments 193

Table 5.18 Segmenmtation based on respondents' Rasch measures of the satisfaction attitudes instruments 196

Table 5.19 Consumer segmenmtation based on resondents' raw mean scores of the satisfaction attitudes instruments 196

Table 5.20 Consumer segmenmtation based on respondents' Rasch measures of the product criteria instruments....................198

Table 5.21 Consumer segmenmtation based on respondents' raw mean scores of the product criteria instruments....................................198

Table 5.22 Gender and Age distribution in consumer segments based on Rasch measures 199

Table 5.23 Gender and Age distribution in consumer segments based on raw mean scores..199

Table 6.1 6 Beef lasagne ready meal product used in this research...208

Table 6.2 Final questionnaire of the benchmarking test 212

Table 6.3 Rasch-Andrich thresholds of the original 9-point scale 216

Table 6.4 Tests of unidimensionality 218

Table 6.5 The item pairs that exhibited potential LID 218

Table 6.6 Counts and proportion of extreme residuals 219

Table 6.7 Misfitting panellists 219

Table 6.8 Measures and outfit MNSQ statistics of products................220

Table 6.9 Measure and outfit MNSQ statistics of attributes.................220

Table 6.10 Rasch reliability statistics of the three facets.....................223

Table 6.11 Chi-square tests for fixed effect and random effect...........223

Table 6.12 Results for ANOVA, Kruskal-Wallis test and residual analysis 225

Table 6.13 Multiple comparisons between products 226

List of Abbreviations

ANOVA	Analysis of variance
CFA	Confirmatory factor analysis
CI	Confidence interval
CMLE	Conditional maximum likelihood estimation
CTT	Classical test theory
d.f.	Degrees of freedom
DIANA Clustering	DIvisive ANAlysis Clustering
DIF	Differential item functioning
EFA	Exploratory Factor analysis
HTAS	Health and Taste Attitude Scales
Infit	Inlier-sensitive fit
JMLE	Joint maximum likelihood estimation
KMO test	Kaiser-Meyer-Olkin test
KR-20	Kuder-Richardson formula 20
LID	Local item dependence
MFR model	Many-Facet Rasch model
MFR-RS model	Many-Facet Rasch Rating scale model
M-H test	Mantel-Haenszel test
MLE	Maximum likelihood estimation
MMLE	Marginal maximum likelihood estimation
MNSQ	Mean-square fit statistics
MSA	Measure of sample adequacy
Outfit	Outlier-sensitive fit
PAF	Principal axis factoring
PAIR	Pairwise conditional likelihood Estimation
PCA	Principal component analysis
PCAR	Principal component analysis on residuals[1]
PC model	Partial credit model
PROX	Normal Approximation Algorithm
RMSE	Root mean square error
RS model	Rating scale model
SE	Standard error
SE_m	Standard error of measurement
ZSTD	Z-standardised fit statistics

[1] Standardised model residuals in Rasch analysis

Chapter 1 Introduction

Fuller (2011) defined new product as:

(1) "*A product not previously manufactured by a company and introduced by that company into its marketplace or into a new marketplace, or*"

(2) "*The presentation or rebranding by a company of an established product in a new form, a new package or under a new label into a market not previously explored by that company*."

The development of different types of new food products requires the implementation of different strategies. For example, brand awareness and extensibility should be investigated by consumers if a company decide to create an "Equity transfer product[1]". However, no matter which type of product is to be developed, the project usually starts from searching market opportunity until product sales (Cooper and Kleinschmidt, 1986). The task of understanding consumers' needs and perceptions towards the product plays an important role in projects involving entirely new product development. For example, in the Stage-gate® model (Cooper, 1988), which is arguable the most popular new product development framework[2], consumers' feedback and/or consumer test data need to be collected in every stage of the project.

To collect consumer data, a variety of data collection methods and tools have been developed in the past decades. For example, online survey tools such as SurveyMonkey® have been widely used for data collection. While in the consumer sensory test area, the traditional lab-based test program has been updated to cloud-based system (*e.g.* Compusense® cloud and Redjade®). However, the quantity of data does not directly link to the quality of measurement. The more important thing is developing the right measurement system and interpreting data in the right way. Currently, the consumer research practitioners are mainly relying on a series of linear models, which were developed on the basis of a theory of measurement called classical test theory (CTT), to interpret the observed data collected from consumer research (Ganglmair and Lawson, 2003). Although the CTT models can obtain relatively simple solution for data analysis, they not only make unrealistic assumptions but also split the connections between the characteristics of persons and test items, which would

[1] Equity transfer products are the products new to a specific product category with brand name and brand image that already be recognised by consumers for another category. An example is chocolate bar made under the ice cream brand "Magnum".

[2] The latest version of Stage-gate ® model can be seen in Appendix A

compromise the quality of measurement. The main issues in current practice using CTT approach will be briefly discussed in section 1.3. Before that, an introduction about measurement is given in section 1.1 and 1.2. In this research, unless specified, the term “measurement” refers to the measurement of persons’ non-physical properties such as attitudes and sensory perceptions.

1.1 Definition of measurement and level of measurement

A classic definition of (fundamental) measurement was given by Campbell (1920) as the assignment of numerals to represent the properties of objects according to scientific laws, where the objects must be in an ordered relationship and satisfy a process of concatenation (*i.e.*, addition). Therefore the relationship between the properties can then be measured by evaluating the relationship between the arbitrary numbers assigned. This definition fits well in the realm of natural science. For instance, the length of roads can be measured and compared using the standard unit meter as reference. However, Campbell’s definition fails in the field of social sciences because the variables measured in this area do not possess the additivity in Campbell’s sense. Therefore, Campbell (1940) asserted that the psychological properties cannot be measured scientifically.

To contradict Campbell’s conclusion, Stevens (1946) developed an alternative framework called “level of measurement theory”, under which the measurement is classified into four different levels according to how could the numbers assigned to the variables can be transformed to information. Table 1.1 tabulates the properties of these levels. Under this framework, Stevens redefined the concept of measurement as “*the assignment of numerals to objects and events according to rules*”. Stevens’ level of measurement theory provided a new idea of measurement. Nowadays, it is still broadly accepted as the fundamental basis of measurement, although it is still being critiqued[3].

[3] More detailed information about the debates can be seen in the literature, such as Hand (1996) and Michell (2002). In addition, a few alternative typologies which attempted to redefine the levels of measurement had been proposed (Chrisman, 1998; Mosteller and Tukey, 1977; and Nelder, 1990). But these typologies were not broadly adopted.

Table 1.1 The four levels of measurement (Stevens, 1946) and their basic properties

Level of measurement	**Scale property**	**Mathematic operation**	**Measure of Central tendency**	**Measure of Dispersion**
Nominal	The numbers assigned to the variables of this type are merely symbols, which do not hold any quantitative meaning	= and ≠	Mode	Percent distribution
Ordinal	Nominal scale + monotonic order/rank The distances between scale categories are unknown	=, ≠, <, and >	Median (preferred) Mode	Range (preferred), Interquartile range (preferred), Percentiles, and Percent distribution
Interval	Ordinal scale + specific distances between categories The value “zero” does not mean “nothing”	=, ≠, <, >, +, and -	Mean (preferred) Median Mode	Standard deviation (preferred), Variance (preferred), Range, Interquartile range, Percentiles, and Percent distribution
Ratio	Interval scale + a fixed starting point of “zero” represents “nothing”	=, ≠, <, >, +, -, x, and ÷	Mean (preferred) Median Mode	Standard deviation (preferred), Variance (preferred), Range, Interquartile range, Percentiles, and Percent distribution

1.2 The building blocks of measurement

The measurement can be decomposed into four building blocks (Wilson, 2004), which are the construct, item responses, outcome space and measurement model. Figure 1.1 explains the basic connections between the four building blocks.

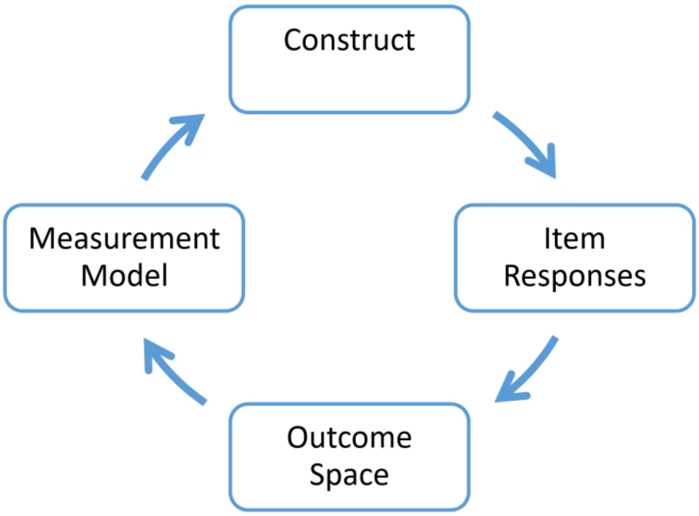


Figure 1.1 Four building blocks of measurement adapted from Wilson (2004)

1.2.1 Construct

A construct can be defined as a hypothesised characteristic of people (Cronbach and Meehl, 1955). It is the theoretical object of researchers' interest in respondents. Therefore, the first step of developing a measurement system is defining the construct to be measured.

A construct often cannot be observed directly, but can be reflected and interpreted via persons' responses to delicately designed tasks or question. These tasks or questions are called items. After the ideation of the construct, the following task is designing the items. In theory, there are a near infinite number of items that can reflect the construct of interest. The selection of items usually starts from developing an initial item pool that consists of most representative items (DeVellis, 2011). Next, the items are carefully inspected so that the size of the initial item pool can be reduced by removing some of the questionable items, such as the items that are highly similar to the others or items which have ambiguous meanings. Finally, a pre-test can be conducted to decide the final set of items to use.

It should be noted that in many occasions the known information at the beginning of measurement development may not be sufficient for defining a clear construct. In that case, the construct should be refined during the other stages of measurement.

1.2.2 Item response

The second building block is the item responses, which refer to persons' responses to specific items. In measurement, the information about the construct is gathered by collecting the item responses. Therefore the format of the item responses is important to the measurement.

A number of formats of item responses have been developed. The most commonly used formats in consumer research are open-end response and fixed-option response. The former is often used in qualitative research, whereas the latter one is usually employed in quantitative studies. Sometimes, a hybrid response format can be seen. For example, in a flavour profile, the panellists would be asked to elicit a list of attributes in open-end format first, followed by rating the intensity of the attributes on an intensity scale using fixed-option (Lawless and Heymann, 2010). The selection of the format of item responses is affected by various factors, such as the age and education levels of targeted participants. One can present the items in the most commonly used format such as Likert item (Likert, 1932), or develop a particular format based on the item statement and the background of the research participants.

1.2.3 Outcome space

The outcome space, which was first introduced by Marton (1981), is the third building block of measurement. The outcome space can be considered as a scoring guide in most of the cases. It defines the rules of categorising the item responses for inference. Numbers are assigned to the item responses on particular levels of measurement according to these rules. By doing that, the item responses are transformed to item scores for further analysis.

1.2.4 Measurement model

The last building block is the measurement model. It is used for computing the inferential statistics from the item scores. These inferential statistics are used to relate the item responses back to the construct. Therefore the whole cycle of the measurement can be completed.

1.3 The measurement issues in current consumer research for new food product development

1.3.1 The misuse of ordinal rating scales

In consumer research, ordinal rating scales are often used as a medium for transferring persons' responses to data. A few examples of ordinal rating scales that are broadly used in new food product development related consumer research can be seen in table 1.2.

Table 1.2 Some examples of ordinal rating scales used in consumer research for new food product development

Scale	**Scale Categories**	**Application**	**Source**
Likert Scale	5-point or 7-point from "Strongly disagree" to "Strongly agree"	Consumer attitude	Likert, 1932
Purchase Intent Scale	5-point or 11-point from "Definitely won't buy" to "Definitely will buy"	Consumer purchase intent	Juster, 1966
Hedonic Scale	9-point[1] from "Dislike extremely" to "Like extremely"	Consumer sensory acceptability test	Peryam and Pilgrim, 1957

[1] 9-point is the standard format of hedonic scale in food industry. The other formats exist (*e.g.* 11-point hedonic scale with two additional scale categories "Dislike greatest imaginable" and "Like greatest imaginable" anchored at the two ends of the scale).

In practice, when analysing consumers' responses to questions within the frame of these ordinal rating scale reference, one normally treats the data as interval data that fits in Stevens' typologies (see table 1.1). The successive scale categories are assumed to be equal spaced, therefore the parametric statistical inferences can be computed. For example, in the analysis of data collected using 7-point Likert scale (table 1.3), arbitrary scores of 1-7 are assigned to the scale categories from "1=Strongly Disagree" to "7=Strongly Agree". One assumption of this rule of assignment is that the spacing between adjacent categories is "1" constantly (Likert, 1932).

Table 1.3 The assignment of numbers (raw score) to 7-point Likert scale

Scale category	Raw score	Assumed spacing to next category
Strongly disagree	1	1
Disagree	2	1
Slightly disagree	3	1
Neither agree nor disagree	4	1
Slightly agree	5	1
Agree	6	1
Strongly agree	7	-

However, in measurement involving human participants, the assumption of equal spacing cannot hold. The spacing between each category of such types of rating scales is in fact unknown and uncontrolled. The numbers assigned to the categories are arbitrary numbers that only stand for the ordered ranks, rather than the actual locations on the scale. As pointed out by Wright and Linacre (1989), "*all observations begin as ordinal, if not nominal, data*".

This raises a question about whether parametric tests (*e.g.* ANOVA), which require data to be interval, can be employed to analyse data at the ordinal level of measurement. This issue has been debated in the literature for several decades (Carifio and Perla, 2008; Norman, 2010). The continued use of parametric tests for this type of data is based on reasons that parametric tests are robust on ordinal data to some extent (Carifio and Perla, 2008; Norman, 2010; Sullivan and Artino Jr, 2013), while others have argued that unless sufficient justification is provided, parametric tests should not be used (Jamieson, 2004; Kuzon *et al.*, 1996). There are several reasons that the parametric tests should not be employed to ordinal data collected. Firstly, one cannot perform arithmetic operations on ordinal data, thus comparing the means and standard deviations of ordinal data is meaningless (Marcus-Roberts and Roberts, 1987). Secondly, the distribution of ordinal data may be skewed. Therefore the data collected from the scales may violate the assumption of normal distribution (Hsu and Feldt, 1969; Villanueva *et al.*, 2000). Thirdly, the extreme scale categories are less used than the central categories in practice. Consequently, the location of extreme scale categories would be further apart from the centre of the scale (Bishop and Herron, 2015). Finally, the respondents vary in the ways of interpreting and using scale categories. For instance, some respondents may intend to use middle categories if they are not sure about how to respond to the items, while others may avoid using it. Hence the results computed from uncalibrated raw scores would be ambiguous and population-dependent.

A number of solutions have been proposed for dealing with the ordinal scale data. Allen and Seaman (2007), and Jamieson (2004) suggested that only non-parametric procedures should be employed when using ordinal rating scales such as the Likert scale. However, compared to the parametric counterparts, the non-parametric tests are normally less powerful (Whitley and Ball, 2002). Other practitioners have developed various alternative scales such as the Visual Analogue Scale (VAS) for attitude and sensory study, Labelled Magnitude Scales (LMS) and Labelled Affective Magnitude scale (LAMS) for sensory evaluation (Green *et al.*, 1993; Schutz and Cardello, 2001). They are all category-ratio type of scales, which require the respondents to rate on a line with anchored labels so that interval data may be produced. However, some researchers found that these scales may be less valid when using children or untrained participants (Hasson and Arnetz, 2005; Lim *et al.*, 2009).

In short, there is a need to solve this contradiction rooted in the nature of ordinal rating scales such as the Likert scale (for consumer survey) and the 9-point hedonic scale (for sensory test). Wright (1999) suggested that one should use a measurement model to convert the raw data into linear measures before conducting the linear statistical analysis[1].

1.3.2 The selection between individual measurement and composite measurement

Measurement can be classified into two forms in another way, according to the number of items in the instrument (*i.e.* questionnaire): **Individual Measurement** and **Composite Measurement** (figure 1.2).

1.3.2.1 Individual measurement

Individual measurement is the simplest form of measurement, in which a single item might be used to obtain a holistic score or measure of the variable of interest. This type of measurement is often seen in the evaluation of consumers' overall perception towards the construct (*e.g.*, consumers' purchasing intent, and overall liking of a product, *etc.*). Time and cost saving are the main benefits of using individual measurement (Martinez-Martin, 2010).

[1] The scores are often misused as measures in current practice. The scores and measures are two different concepts used for the same purpose of describing the measurement results. See section 1.4.1 for the explanation.

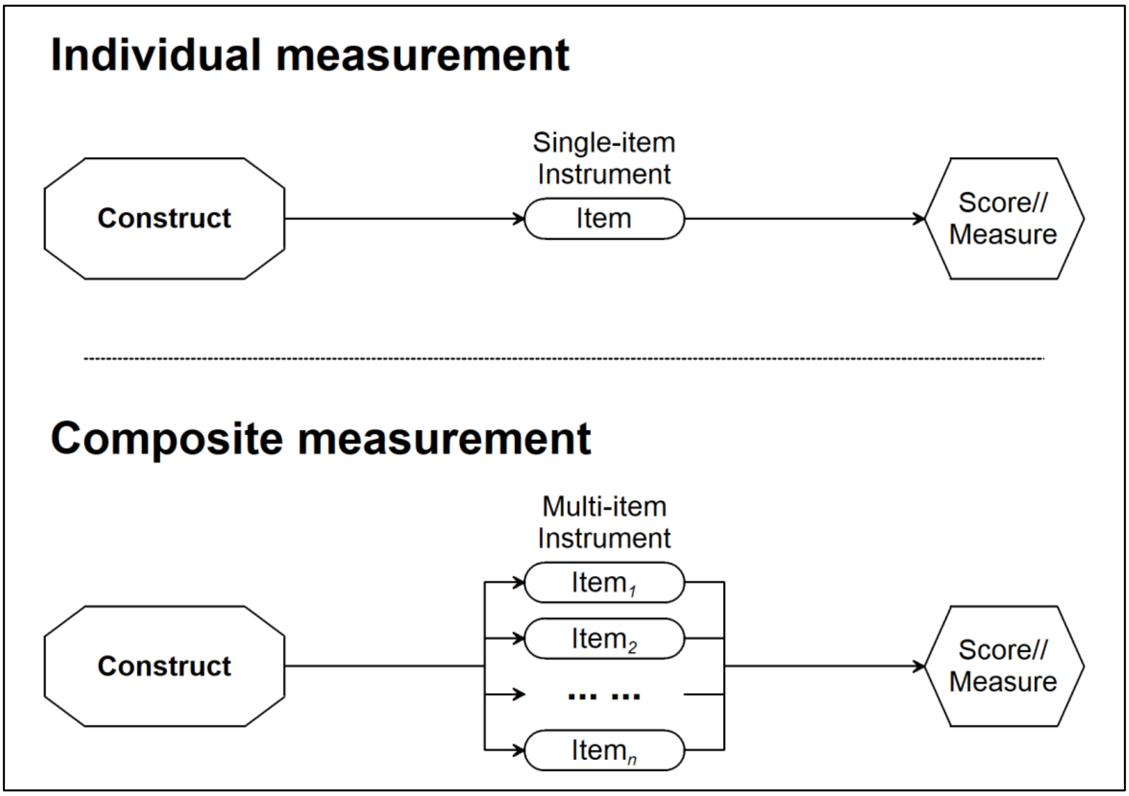


Figure 1.2 Individual measurement and composite measurement adapted from Martinez-Martin (2010)

However, there are several issues that need to be considered before using the individual measurement.

Firstly, the individual measurement can only be used for the evaluation of a concrete attribute, which is defined by Rossiter (2002) as an attribute that "*has virtually unanimous agreement by raters as to what it is, and they clearly understand that there is only one, or holistically one, characteristic being referred to when the attribute is posed, as in a questionnaire or interview, in the context of the to be rated object*". To fulfil this requirement, in the development of the single item, a group of expert raters are needed (Bergkvist and Rossiter, 2007). The effort of using expert raters for deciding the single item for individual measurement, however, is sometimes questionable (Sarstedt *et al.*, 2016). One reason is the number of expert raters needed and the how they qualify for the task is not clear in the literature. Another reason is, as McIver and Carmines (1981) indicated, "*it is very unlikely that a single item can fully represent a complex theoretical concept or any specific attribute for that matter*". Loo (2002) also argued that the individual measurement should only be used for extreme homogeneous construct.

Secondly, the score or measure obtained from the individual measurement only concerns an individual's holistic perception towards the construct. However, the other information cannot be extracted from it for explaining variability, which is often the interesting part in a consumer research. For example, in consumer satisfaction study, it is often important to know not only the consumers' overall satisfaction towards the product or service, but also the degree of their satisfaction towards the particular aspects of the product or service; by doing this, the unsatisfied aspect could be improved on.

Thirdly, the individual measurement is more vulnerable to the random measurement error and the bias than the composite measurement using multiple items (Hoeppner *et al.*, 2011). This is because the measurement error can be averaged out when multiple items are used to obtain a total score or measure (Nunnally and Bernstein, 1994). In the meantime, using the multiple items developed for covering the wider range of the construct can reduce the risk of introducing bias generated by respondents' misinterpretation of the single abstract item.

Last but not least, the most commonly used reliability coefficient[2] "Cronbach's alpha", which is an estimate of internal consistency of items, cannot be computed using a single item.

In conclusion, although the individual measurement is adequate for some simple construct, it is not preferred for measuring complex constructs in consumer research for new food product development. As suggested by Churchill Jr in his broadly cited paper (1979), "*marketers are much better served with multi-item than single-item measures of their constructs, and they should take the time to develop*".

1.3.2.2 Composite measurement

Composite measurement is the measurement where a set of items are used for evaluating different components of the same underlying construct. Although a variety of multidimensional analysis approaches have been developed for measuring multidimensional constructs in recent decades, only unidimensional modelling is discussed in this thesis. In other words, for a multidimensional construct, instead of performing multidimensional analysis, the instrument would

[2] The reliability will be discussed in details in section 2.3.5.

be decomposed to unidimensional sub-instruments first according to the results of dimensionality test, then the sub-instruments are analysed individually.

In food-related consumer research, composite measurement is often employed to obtain a single score or measure of a complex concept that contains several aspects. For example, the food choice questionnaire (Steptoe *et al.*, 1995) was developed to evaluate the determinants of consumers' food choice. It consisted of thirty-six items that were differentiated to nine subscales by exploratory factor analysis. A summated single score of each subscale was calculated to represent respondents' overall attitude towards the corresponded underlying factor. Another example is UC Davis' 20-point wine-scoring method (Amerine *et al.*, 1959; Ough and Winton, 1976), which obtains a composite quality rating for a given wine computed from nine specific attributes scores and a general quality score.

Assessing several items which belong to the same construct can increase the information used for computing the final score or measure of the construct. Therefore the reliability and validity of the measurement can be improved compared to the individual measurement which uses only one item.

Under the framework of standard approach (*i.e.* the CTT and its extensions, see chapter 2 for details), the composite score of the measurement is usually calculated by summating or averaging the scores of the individual items. Other methods exist, such as standardising the scores to percentage or a specific mean and standard deviation (DeVellis, 2006; Martinez-Martin, 2010), and using a weighted mean (Lord and Novick, 1968), although these methods are less frequently used.

However, no matter which method is applied, in the context of CTT, the composite score is "built up" above the assumption of data being interval, which cannot hold for the ordinal rating scale. The impact of the issue related to the ordinal nature of the raw observation scores is even greater in composite measurement than individual measurement. It is assumed that the same composite score has the same meaning on the construct level, which is not true because the same score can be summated by different combinations of ordinal responses. For instance, one cannot say the sum of "Strong disagree" and "Strongly agree" has equal meaning with the sum of "Slightly disagree" and "Slightly agree" when using the 7-point Likert scale in a test consisting of two items, even if that their summated scores are identical (under the rules of assignment of numbers illustrated in table 1.3).

In addition to that, the individual items are merged in the composite measurement, where only a total score is computed within the framework of the standard approach. The characteristics of the items do not contribute to the modelling of the result. Consequently, the result is only inferential at the global level of the construct. The information about the particular aspects of the construct represented by individual items is hidden.

Moreover, the merge of items increases the difficulty of evaluating item quality. The “bad” item can be covered. Although the thorough item analysis[3] is a routine part of analysis in the studies of psychometrics, the statistics calculated in standard approach are biased by the distribution of construct levels in the given population (Embretson, 1999). They are population-dependent.

Therefore, the composite scale must be evaluated under a new paradigm other than the standard approach, which should be able to provide bias-free meaningful statistics at both global level and individual levels.

1.4 Applying Rasch analysis to reduce the issues related to the measurement in consumer research for new food product development

To overcome the issues addressed in section 1.3, Rasch analysis can be applied.

1.4.1 What is Rasch analysis

Rasch analysis stands for the analysis based on a family of parametric probabilistic response models (named after G. Rasch). Rasch analysis intends to model the ordinal raw scores into interval measures, using the probabilities of specific responses as a function of person and item parameters[4].

One should distinguish the difference between the concepts of “scores” and “measures”. The scores are only discrete cumulative counts of the arbitrary numbers assigned to specific responses by certain rules during the measurement. They are uncalibrated due to the ordinal nature of the data, thus merely reflecting the initial observations of a measurement. In contrast, the measures are the continuous real numbers modelled from the scores, which

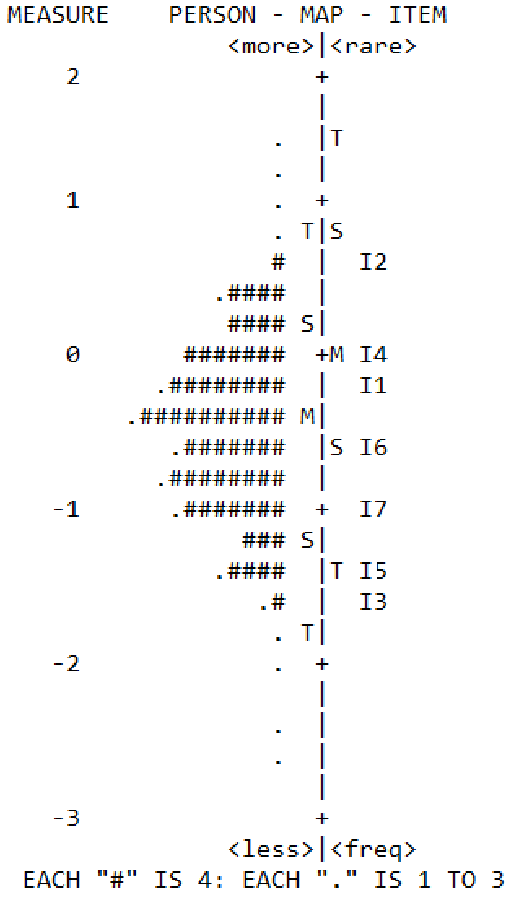
[3] Such as the evaluation of the mean, standard error, and item-total correlation, *etc.*
[4] And other parameter in a multi-facet model (section 2.1.2.3)

explain the calibrated locations of the elements (*i.e.* items, persons, *etc.*) on the measurement continuum of the construct.

In order to provide comparable interval measures, Rasch analysis expects the responses follow a hierarchical probabilistic pattern:

(1) Any person should have more chance to endorse an item than all other items which are more difficult to endorse.
(2) Any person who ranks higher on the scale than another person has greater probability to endorse all items than the other one.

The location of the measurement elements and the hierarchy pattern on the construct continuum can be presented on a graph, which is usually called a "Wright map" in honour of B. Wright for his contributions to the measurement theory. Figure 1.3 exhibits an example of a Wright map, where the person and item are anchored on the map according to their measures.



```
MEASURE    PERSON - MAP - ITEM
               <more>|<rare>
    2                +
                     |
                  .  |T
                  .  |
    1             .  +
                  . T|S
                  #  |  I2
              .####  |
               #### S|
    0       #######  +M I4
          .########  |  I1
        .########## M|
           .#######  |S I6
          .########  |
   -1      .#######  +  I7
                ### S|
              .####  |T I5
                 .#  |  I3
                  . T|
   -2             .  +
                     |
                  .  |
                  .  |
                     |
   -3                +
               <less>|<freq>
 EACH "#" IS 4: EACH "." IS 1 TO 3
```

Figure 1.3 An example of Wright map

(adapted from the Wright map obtained from the case study I of this research)

The Wright map can also be used for conceptualising the initial construct, which was defined as the first building block of measurement (Wilson, 2004). Thus, the development tasks of measurement, within the framework of Rasch analysis, can be planned using Wilson's construct modelling approach (section 1.2), as shown in figure 1.4. This approach had been employed in two of the case studies in this research (case study III and IV).

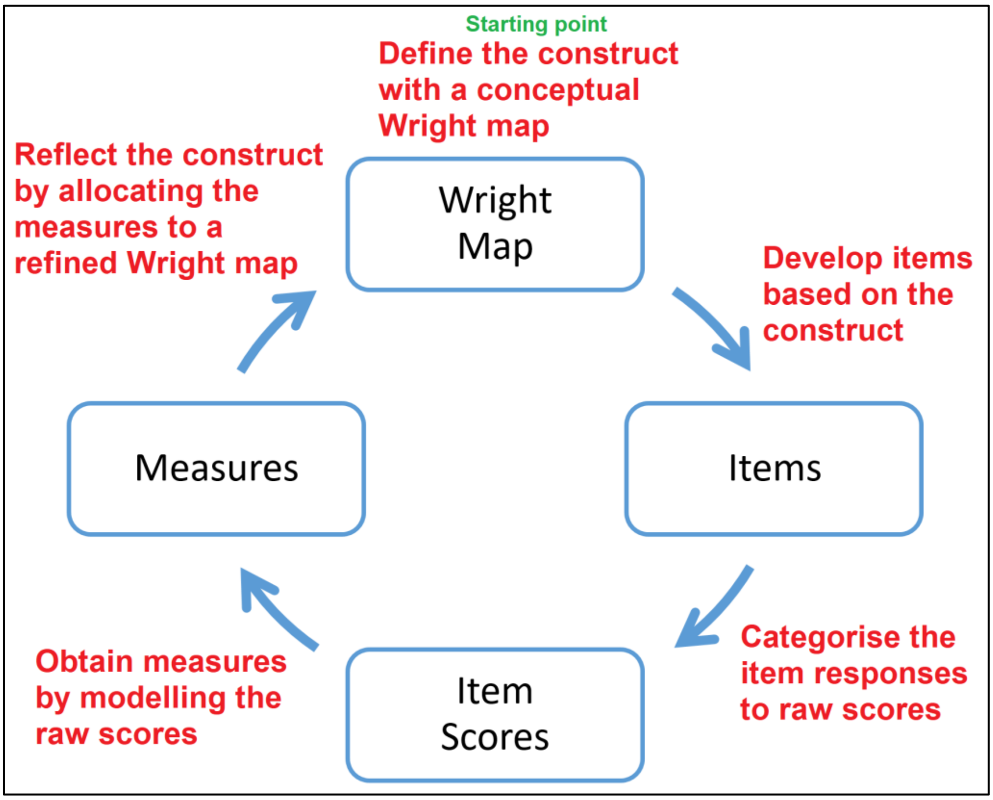

Starting point
Define the construct with a conceptual Wright map
Wright Map
Develop items based on the construct
Items
Categorise the item responses to raw scores
Item Scores
Obtain measures by modelling the raw scores
Measures
Reflect the construct by allocating the measures to a refined Wright map

Figure 1.4 The construct modelling approach within the framework of Rasch analysis adapted from Wilson (2004)

1.4.2 How can Rasch analysis reduce the issues related to the measurement in consumer research for new food product development?

Firstly, the conversion of ordinal raw scores into interval measures in Rasch analysis can solve the issues of conducting interval-level statistical analysis on ordinal data.

In Rasch analysis, the measures are reported in logit, which is defined as the natural log of the odd ratio (Cox, 1970). Logit is the unit of the logistic ogive. The transformation from raw scores to logits does not assume any particular sample or item distribution in Rasch analysis. Figure 1.5 displays an example of model transformation from the ordinal raw scores of a 7-point rating scale to the linear Rasch measures in logit.

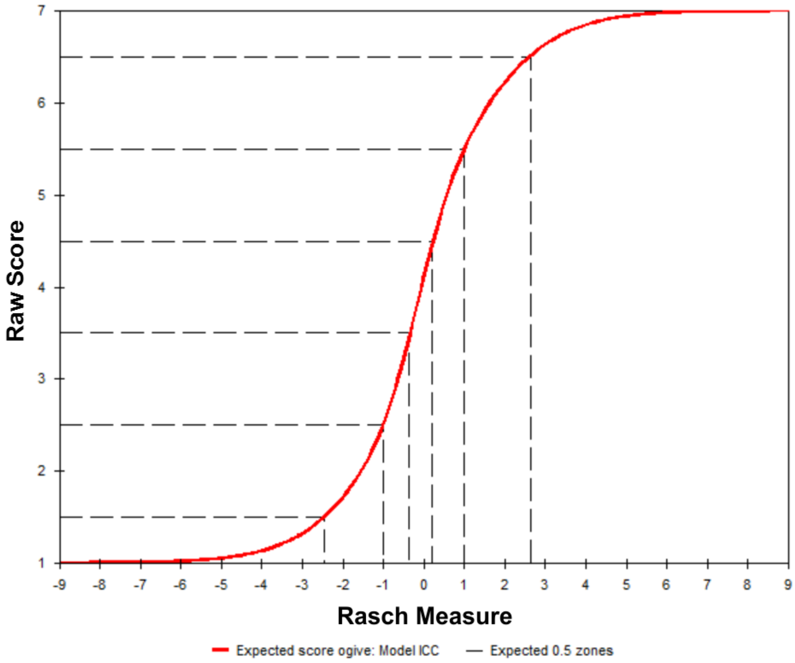


Figure 1.5 Transformation of raw score to Rasch measure

The “expected 0.5 zone” line shows the Rasch-half-point thresholds, corresponding to the expected value of 0.5 points in the unit of raw scores, see section 2.3.1.1

Secondly, unlike the standard approach (*i.e.* CTT) that mainly focuses on the test level of measurement by treating all items as a group (DeVellis, 2006), Rasch analysis obtains estimates and standard error of measurement[5] at the individual levels, taking the characteristics of individual items into account. One can inspect both overall and specific information after modelling. Firstly, for each person, a single measure of the underlying construct can be modelled by multiple items using Rasch analysis. This should be used to replace the summated score of the composite measurement computed by standard approach using CTT models. Secondly, individual items are calibrated on the same scale together with persons. The individual items can be evaluated in a hierarchical order according to their estimates and standard error of measurement (SE_m) during the measurement.

[5] The computation of standard error of measurement can be seen in section 2.3.6.

1.4.3 Previous applications of Rasch analysis in consumer research for new food product development

Despite the fact that Rasch analysis is commonly employed in education and health studies as one of the standard methods for data processing, it has not raised broad attention to the practitioners in the field of consumer research for new food product development areas, especially those working on food-related researches. Table 1.4 depicts some of the previous research that utilised the Rasch analysis in measuring consumers-related aspects. However, none of them were directly linked to new food product development. Therefore, there is a need of exploring the usage of Rasch analysis further in food-related area, and establishing the standard procedures for the applications thereafter.

Table 1.4 Previous applications of Rasch analysis in consumer research

Consumer-related area	**Researchers**
Evaluating perceived quality of product or service	De Battisti *et al.* (2005) García *et al.* (1996)
Measuring consumer's affective response to product attributes	Camargo and Henson (2015)
Investigating consumer satisfaction	De Battisti *et al.* (2010)
Exploring the influence of contextual conditions on consumers' purchasing motivation	Baranowski *et al.* (2008) Soutar and Cornish-Ward (1997) Tanner *et al.* (2004)
Developing international marketing strategies via cross-national study	Pantouvakis and Renzi (2016) Salzberger *et al.* (2009)
Monitoring sensory panel selection and performance	Álvarez and Blanco (2000) Thompson (2003)

1.5 Overview of the research

1.5.1 Research aims

The main aims of this research are:

(1) Demonstrating the benefit of applying Rasch analysis to consumer research for new food product development study;

(2) Exploring the application of Rasch analysis in development and validation of instrument for food-related consumer research tasks.

1.5.2 Thesis structure

The whole thesis is made up of seven chapters.

Chapter 2 provides a literature review about the classical test theory (CTT) and Rasch analysis. The models and model assumptions, methods for estimation of parameters, and quality control procedures of both approaches are introduced.

Chapters 3 to 6 report the details of the four case studies, one study per chapter.

Chapter 3 reports a study that revisits an existing consumer attitude instrument – Health and Taste Attitude scales (Roininen *et al.*, 1999). Both CTT approach and Rasch analysis are applied to the survey data of 269 respondents. The results computed from the two methods are compared.

Chapter 4 is related to a sensory acceptability study on twenty-four food and beverage products. A Many-Facet Rasch Rating scale (MFR-RS) model[6] is applied to obtain the composite measures of overall liking modelled from 8 sensory attributes and the holistic measures of overall liking modelled using a single overall acceptability item. The test hypothesis is the composite measure modelled using attribute ratings can provide greater power in differentiating the difference between the overall liking of the products than the holistic measure.

Chapter 5 refers to the development of a series of consumer survey instruments within the framework of Rasch analysis. The instruments are associated with three aspects of consumer insights towards ready meals. The data is collected from 333 participants, who are further segmented by performing a cluster analysis on their measures related to their satisfaction attitudes towards ready meals, and their decision making pattern, respectively. In addition, the cluster analysis is also

[6] About the model, see section 2.1.2.3.

performed to the related raw scores for comparison. The predictive power of the instruments are evaluated by examining the relationship between the segmentations and the recorded consumption frequencies of three types of meals including ready meal, restaurant meals and take-away meals.

Chapter 6 reports an application of Rasch analysis in the development of a consumer sensory benchmarking test for beef lasagne ready meal products. It is a two-stage study. A list of sensory attributes are generated from 45 research volunteers via 1-to-1 interviews in the first stage. After that, a benchmarking test involving 96 panellists is conducted in the second stage. The composite measures of the product overall liking are then modelled using the Many-Facet Rasch Rating scale (MFR-RS) model and compared.

Chapter 7 refers to a summary of the research and a discussion of some special issues elicited from the application of Rasch analysis in the case studies, such as the sample size and the appropriate number of categories in the rating scale.

Chapter 2 Literature review: A comparison between classical test theory (CTT) and Rasch analysis

2.1 An overview of classical test theory models and Rasch models

2.1.1 Classical test theory models and the extensions

The foundation of classical test theory (CTT) was laid by Spearman's work (1904). Nowadays, CTT is still the predominant measurement theory in research.

The main concern of CTT is the estimation of measurement error. Under the framework of CTT, the observed score variable is decomposed as the sum of a true score variable and an error variable. According to Lord and Novick (1968), the "true score" can be defined as "*the expected observed score with respect to the propensity distribution of a given person on a given measurement*". And the "error" is defined as the disturbance to the measurement which is caused by uncontrolled factors in the measurement procedure.

2.1.1.1 CTT model for individual measurement

The equation of CTT model for individual measurement can be expressed as:

$$X = T + E \tag{2.1}$$

where

X is the observed variable taking value x,

T is the true score variable taking value τ,

E is the error variable taking value ε.

The application of CTT model requires several assumptions, which are written in equations 2.2~2.5:

(1) The expected true score is equal to expected observed score:

$$\mathcal{E}T = \mathcal{E}\mathrm{X} \tag{2.2}$$

where

$\mathcal{E}\mathrm{T}$ is the expected true score variable,

$\mathcal{E}\mathrm{X}$ is the expected observed score variable.

(2) The expected error score is zero:

$$\mathcal{E}E = \mathcal{E}X - \mathcal{E}T = 0 \qquad (2.3)$$

where

$\mathcal{E}\mathrm{E}$ is the expected error score variable that has a value of 0,

$\mathcal{E}\mathrm{X}$ is the expected observed score variable,

$\mathcal{E}\mathrm{T}$ is the expected true score variable.

(3) True score and the error score are uncorrelated:

$$\rho_{TE} = 0 \qquad (2.4)$$

where

ρ_{TE} is the correlation coefficient between true score variable T and error score variable E.

(4) The variance of observed score is therefore equivalent to the sum of the variance of true score and error score:

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2 + 2\sigma_{TE} = \sigma_T^2 + \sigma_E^2 + 2\rho_{TE}\sigma_T\sigma_E = \sigma_T^2 + \sigma_E^2 \qquad (2.5)$$

where

σ_X^2 is the variance of the observed score variable X,

σ_T^2 is the variance of the true score variable T,

σ_E^2 is the variance of the error score variable E,

σ_{TE} is the covariance of true score variable T and error score variable E,

ρ_{TE} is the correlation coefficient between true score variable T and error score variable E that has a value of 0.

2.1.1.2 CTT model for composite measurement

In a composite measurement, the person's composite score (in other words, the total score) is normally computed by summating or averaging the observed scores of individual items of the instrument. Within CTT framework, the observed score should also be decomposed as the sum of true score" and error score. Therefore, for a measurement composed by n items, using the summating method:

$$X = \sum_{i=1}^{n} Y_i; \qquad T = \sum_{i=1}^{n} T_i; \qquad E = \sum_{i=1}^{n} E_i \tag{2.6}$$

where

X is the composite score variable of the measurement taking the value x,

Y_i is the observed score variable on item i taking value y_i,

T is the composite true score variable,

T_i is the true score variable on item i taking value τ_i,

E is the composite error variable,

E_i is the error variable on item i taking value ε_i.

Likewise, a few assumptions derived from the CTT model for individual measurement are required in the application of CTT model for composite measurement, such as:

(1) The sum of expected true score on n items is equal to the sum of individual expected observed scores on the same item set.

$$\sum_{i=1}^{n} \mathcal{E}T_i = \mathcal{E}T = \mathcal{E}\mathrm{X} = \sum_{i=1}^{n} \mathcal{E}Y_i \tag{2.7}$$

where

$\sum_{i=1}^{n} \mathcal{E}T_i$ is the sum of expected true score on each item,

$\mathcal{E}$T is the expected true score variable,

$\mathcal{E}$X is the expected composite score variable,

$\sum_{i=1}^{n} \mathcal{E}Y_i$ is the sum of expected observed score on each item.

(2) The expected value of the error score is zero in each sub-population of observational unit:

$$\sum_{i=1}^{n} \mathcal{E}E_i = \mathcal{E}E = \mathcal{E}X - \mathcal{E}T = 0 \tag{2.8}$$

where

$\sum_{i=1}^{n} \mathcal{E}E_i$ is the sum of expected error score on each item,

$\mathcal{E}$E is the expected error score variable that has a value of 0,

$\mathcal{E}$X is the expected observed score variable,

$\mathcal{E}$T is the expected true score variable.

2.1.1.3 The extensions of CTT models

A few extensions of CTT models have been developed in the past decades. The discussion of them is beyond the scope of this research, thus only a very brief introduction about three most representative models are discussed here.

(1) The generalisability theory (G theory)

The G theory was devised by Cronbach and his colleagues (1963; 1972). It further decomposes the error variable to several components on the basis of CTT. The main concern of G theory is the analysis of sources of errors, where the Analysis of Variance (ANOVA) should be employed. In G theory, the reliability coefficient[1] is replaced by an analogue term under the name of "generalisability". The G theory model equation can be written as:

$$X = T + E_1 + E_2 + \cdots + E_k \tag{2.9}$$

where

X and T share the same definition with those in CTT (*i.e.* X is the observed score variable, T is the true score variable),

E_1, E_2, … E_k are k components of error variable associated with individual source such as items, persons, *etc.*

(2) Factor analysis

The factor analysis (Spearman, 1904; Thurstone, 1931b; Thurstone, 1934) decomposes the true score variable into a number of components associated with specific underlying factors . For a factor analysis model made up of n items and m factors, the equation (adapted from Engelhard, 2013) can be written as:

$$\begin{aligned} X_1 &= \lambda_{11}F_1 + \lambda_{12}F_2 + \cdots + \lambda_{1m}F_m + E_1 \\ X_2 &= \lambda_{21}F_1 + \lambda_{22}F_2 + \cdots + \lambda_{2m}F_m + E_2 \\ &\cdots\cdots \\ X_n &= \lambda_{n1}F_1 + \lambda_{n2}F_2 + \cdots + \lambda_{nm}F_m + E_n \end{aligned} \tag{2.10}$$

where

X_1 ~ X_n are the observed score variables of the n items,

λ_{11} ~ λ_{nm} are the factor loadings,

F_1 ~ F_m are the factor scores variables of the m factors,

E_1 ~ E_n are the error variables of the n items.

(3) Structural Equation Modelling (SEM)

Structural Equation Modelling (SEM) is a combination of structural models and factor analysis models. The first practicable SEM model (Jöreskog, 1973) and first computer software LISERL (Jöreskog and van Thiilo, 1972) were developed by Jöreskog and his associate. The model is not provided here.

[1] Reliability will be discussed in section 2.3.5.

2.1.2 The family of Rasch models

Since Rasch spelled out his individual-centred statistical technique for measuring dichotomous items in his book firstly published in 1960, the basic dichotomous Rasch model has been extended in numerous ways by other researchers such as Andrich (1978a) , Masters (1982) and Linacre (1989). Table 2.1 provides an overview of Rasch models and how to distinguish them.

Table 2.1 An overview of Rasch models

Criteria	**Models**
Dimensionality of construct[1]	Unidimensional model – measurement model is based on the assumption of unidimensionality; Multidimensional model (*e.g.* Kelderman and Rijkes, 1994; Meiser, 1996) – measurement model that are composited by multiple dimensions.
Number of rating scale categories	Dichotomous model (Rasch, 1960/1980) – if the rating scale is binary (*e.g.* Yes/No); Polytomous model (Andrich, 1978a; Masters, 1982) – if there are 3 or more scale categories (*e.g.* 7-point Likert scale); Hybrid model (Masters, 1982) – the instrument consists of both dichotomous item and polytomous item.
Number of measurement facets	General model – if there are 2 measurement facets (*i.e.* person and item); Many-Facet model (Linacre, 1989) – if there are 3 or more measurement facets (*e.g.* sensory evaluation made up of Panellist, Sample and Attribute facets).
The usage of rating scale	Rating scale model (Andrich, 1978a) – if the items share the same scale structure; Partial credit model (Masters, 1982) – if the items do not share the same scale structure (*e.g.* an instrument is made up of both dichotomous item and polytomous item, or the use of same scale categories are different among items).
Number of attempts to an item	General models – if each item is rated only once; Binomial trials model (Andrich, 1978b) – if each item is rated multiple times (the number of attempts is restricted to an upper limit; Poisson counts model (Rasch, 1977) – if each item is rated multiple attempts (the number of attempts is not restricted).
Other extensions	Log-linear model (Kelderman, 1984); The mixture model bind Rasch model with latent class analysis (Rost, 1990); *etc.*

1: This research only concerned the unidimensional Rasch models

2.1.2.1 Dichotomous Rasch model

The dichotomous Rasch model (Rasch, 1960/1980) is the root of all others in the family of Rasch models. The dichotomous Rasch model can be applied to binary data. The equation of dichotomous Rasch can be written as:

$$\ln\left(\frac{P_{ni}}{1-P_{ni}}\right) = B_n - D_i \tag{2.11}$$

Or in an alternative format as:

$$P_{ni} = \frac{\exp(B_n - D_i)}{1+\exp(B_n - D_i)} \tag{2.12}$$

where P_{ni} is the probability of person n who locates at construct level B_n succeeding on item i that locates at construct level D_i. The estimates of parameters are π_{ni} for the probability, β_n for persons and δ_i for items.

2.1.2.2 Polytomous Rasch model

The polytomous scale item can be considered as a set of Dichotomous scale items. The polytomous Rasch model can be considered as an extension of dichotomous Rasch model with an additional parameter refers to the function of rating scale categories. The Rating scale (RS) model and the partial credit (PC) model are the two widely-used extended formats of Rasch models for polytomous items.

(1) Rating scale (RS) model

Rating scale model (Andrich, 1978a) adds a threshold parameter F_x to represent the relative endorsability of the transition from one category of the rating scale to the successive one. This threshold is usually called Rasch-Andrich threshold, or step calibration. The equation of RS model can be written as

$$\ln\left(\frac{P_{nix}}{P_{ni(x-1)}}\right) = B_n - D_i - F_x \tag{2.13}$$

where

P_{nix} is the probability of person n responds category x of rating scale to item i with an estimates of π_{nix},

$P_{ni(x-1)}$ is the probability of the same person choose adjacent category x-1 of a rating scale on the same item with an estimates of $\pi_{ni(x-1)}$,

B_n is the agreeability of person n with an estimate of β_n,

D_i is the endorsability of item i being endorsed with an estimate of δ_i, and

F_x is the relative endorsability of responding category x to category x-1 with an estimate of τ_x.

For example, for a questionnaire using the 7-point Likert scale from “Strongly disagree” to “Strongly agree”, the equations become:

$$\ln\left(\frac{P_{ni(Disagree)}}{P_{ni(Strongly\ Disagree)}}\right) = B_n - D_i - F_{Disagree};$$

$$\ln\left(\frac{P_{ni(Slightly\ Disagree)}}{P_{ni(Disagree)}}\right) = B_n - D_i - F_{Slightly\ Disagree};$$

$$\dots\ \dots$$

$$\ln\left(\frac{P_{ni(Strongly\ Agree)}}{P_{ni(Agree)}}\right) = B_n - D_i - F_{Strongly\ Agree}. \qquad (2.14)$$

where

$P_{ni(Strongly\ Disagee)}$, $P_{ni(Disagee)}$, $P_{ni(Slightly\ Disagee)}$, …, $P_{ni(Agree)}$ and $P_{ni(Strongly\ Agree)}$ are the probabilities of person n select answer option “Strongly Disagree”, “Disagree”, “Slightly Disagree”, …, “Agree” and “Strongly Agree” for item I, respectively,

B_n is the tendency of person n towards agreement, D_i is the endorsability of item i being agreed,

$F_{Disagree}$, $F_{Slightly\ Disagree}$, and $F_{Strongly\ Agree}$, are the relative endorsability of a respondent giving the answer of “Disagree” to “Strongly Disagree”, “Slightly Disagree” to “Disagree” and “Strongly Agree” to “Agree” by orders.

(2) Partial credit (PC) model

If the items share the same scale categories, then the F_x for these items are constant. In some cases, however, when a mixture of scale categories have been used in a scale, the F_x are not constant for all items. Masters (1982) developed the partial credit model to provide calibration against this situation. The PC model replace *F_x* used in rating scale model by a new parameter *F_{ix}*, which estimates the function of scale categories independently for each item or item groups. The basic equation of PC model is:

$$\ln\left(\frac{P_{nix}}{P_{nix-1}}\right) = B_n - D_i - F_{ix} \qquad (2.15)$$

where

P_{nix} is the probability of person n selects category x of rating scale on item i,

$P_{nix\text{-}1}$ is the probability of the same person responds adjacent category x-1 of a rating scale to the same item,

B_n is the agreeability of person n with an estimate of β_n,

D_i is the endorsability of item i being endorsed with an estimate of δ_i, and

F_{ix} is the relative endorsability of responding scale category x to category x-1 on item i with an estimate of τ_{ix}.

2.1.2.3 The Rasch model for more than two measurement facets: the Many-Facet Rasch (MFR) model

The general dichotomous and polytomous Rasch models only take two facets of the measurement (*i.e.* the items and persons) into account. However, there may be more aspects or measurement situations that interact with items and person. For example, in the sensory evaluation research, consumer's overall acceptability ratings on particular product is a compromise among personal perceptions on many attributes (*e.g.*, flavour, aroma, appearance, *etc.*). If we want to measure these perceptions individually, one extra facet that accounts for the attributes should be added to the model, since the standard two-facet Rasch model is not applicable here.

Linacre (1989) developed the Many-Facet Rasch (MFR) Model for analysing this more complicated situation. In rating scale format, the equation of MFR model is:

$$\ln\left(\frac{P_{nijx}}{P_{nij(x-1)}}\right) = B_n - D_i - F_x - C_j \tag{2.16}$$

where in a sensory liking test with several attributes,

P_{nijk} is the probability of product n receiving a rating of x on attribute i from panellist j,

$P_{nij(x-1)}$ is the probability of the same panellist j rate attribute i of product n with rating category x-1,

B_n is the overall liking of product n with an estimate of β_n,

D_i is the endorsability of attribute i being endorsed with an estimate of δ_i,

F_x is the relative endorsability of responding rating category x to x-1, and

C_j is the panellist j's overall liking level with an estimate of λ_j,

Note: in partial credit format, the variable F_x would be replaced by F_{ix} in the equation of MFR Model.

2.2 Estimation of model parameters

2.2.1 Estimation of CTT model parameters

In CTT, the person's true location on a construct is measured by the person's true score, which cannot be observed. The observed score is usually used as the estimate of person's true location because in a "perfect" test, the expected true score is equal to the expected observed score (see section 2.1.1.1 for more details). However, this "perfect" test can never be administrated. Therefore as a supplement to the observed score, a confidence interval, within which the person’s true score is expected to fall, is computed from the observed score and standard error of measurement (SE_m)[2]. In addition to that, the reliability statistics[3] are used to estimate to what extent the observed score is close to the true score.

In addition, in CTT, the item discrimination property is usually evaluated by determining the correlation between the observed item score and the observed total score of the measurement with this item included or excluded, which are named as item-total correlation and item-rest correlation, respectively.

2.2.2 Estimation of Rasch model parameters

In Rasch analysis, the model parameters (*e.g.* item and person, *etc.*) are estimated separately, which eliminates the dependence between each other. A number of estimation methods had been devised in the past decades. The most popular estimation methods for Rasch model parameters can be seen in Appendix B.

Linacre (1999) argued that none of the estimation methods of Rasch model parameters is the “*one best method*”. They are approximately either the same method under different conditions or different methods under same condition. Although different degrees of imprecision and inaccuracy constantly exist behind the estimates of the parameters, these methods are likely to obtain statistical equivalent estimates (Linacre, 1999).

A comprehensive discussion about the advantages and disadvantages of each estimation procedure can be found in the user manual of the software WINSTEPS (Linacre, 2014d). It will be of interest to compare the difference between the estimation methods via case studies. This is, however, beyond the scope of this research. Therefore, only a short introduction about the estimation methods utilised in software WINSTEPS (Linacre, 2014d) and Facets (Linacre, 2014a) is provided here.

[2] More details can be seen in section 2.3.6
[3] More details can be seen in section 2.3.5

2.2.2.1 Normal Approximation Algorithm (PROX)

The PROX method was firstly proposed by Cohen (1979). The original algorithm for dichotomous data can be used for obtaining the adjusted sample-free and test-independent estimates calculated on the initial set of estimates. Linacre (1994a) improved the algorithm to solve the missing value issue. He further extended the method to the analysis of polytomous data (Linacre, 1995b). Currently, this method is used in software WINSTEPS and Facets to give the starting values of estimates.

For a polytomous instrument, the estimation begins with an initial set of estimates for every person measure, item measure and step calibration (*i.e.* the Rasch-Andrich threshold F_x or F_{ix}, see section 2.1.2.2), unless pre-determined "anchor" values are provided by the researcher. To be more specific:

(1) The item measure and person measure are estimated at the origin of the measurement scale;

(2) All items are treated to have the same measure;

(3) Each person is estimated to have the same measure;

(4) The Rasch-Andrich thresholds are unified as 0.

Next, an iterative algorithm is used. The iteration stops when a rough convergence to the observed data pattern is found or the pre-defined number of iterations is reached.

During the iteration, the item estimates are obtained by equation 2.17:

$$D_i = \mu_i - \sqrt{1 + \frac{\sigma_i^2}{2.89}} log_e \left(\frac{R_i}{N_i - R_i} \right) \tag{2.17}$$

where

D_i is the revised estimate of item I,

μ_i is the mean measure of the persons on item I,

σ_i is the standard deviation of the person measures,

R_i is the raw score observed on item I, and

N_i is the most possible item score estimated by the same persons.

The person estimates are produced similarly by applying the equation 2.18:

$$B_n = \mu_n + \sqrt{1 + \frac{\sigma_n^2}{2.89}} log_e \left(\frac{R_n}{N_n - R_n}\right) \qquad (2.18)$$

Where

$\mathbf{B_n}$ is the revised location of person n,

$\boldsymbol{\mu_n}$ is the mean measure of the items for person n,

$\boldsymbol{\sigma_n}$ is the the standard deviation of the item measures,

$\mathbf{R_n}$ is the raw score given by person n, and

$\mathbf{N_n}$ is the most possible item score on the items.

In addition, the estimate of Rasch-Andrich threshold between category x and x-1 is computed by normalising $\log(Observed\ count_x /\ Observed\ count_{x-1})$ to a sum of zero.

2.2.2.2 Joint Maximum Likelihood Estimation (JMLE)

After using PROX procedure to produce the starting estimates of parameters, the software WINSTEPS and Facets continue the estimation by applying iterative JMLE method to obtain more exact estimates, standard errors and fit statistics. The measures are reported in Logits (log-odds units) unless they are rescaled to a certain range (*e.g.* 0-100). Fit statistics[4] are reported as mean-square residuals, which have approximate chi-square distributions. The mean square residuals can also be transformed to z scores.

The JMLE method was proposed by Wright and Panchapakesan (1969). It is also called unconditional maximum likelihood estimation (UCON). The algorithm employs a modified Newton-Raphson iteration method to estimate the parameters. All parameters are estimated simultaneously to maximise the joint likelihood. In each iteration, the expected values of parameters are produced. They, together with the marginal sums of them, are compared with the correlative observed values.

[4] Fit statistics will be introduced in section 2.3.3 in details.

For persons and items, the estimates are improved from the current estimated measures by equation 2.19:

$$y'_{n\ or\ i} = y_{n\ or\ i} + \frac{Observed\ score - Expected\ score\ based\ on\ current\ estimates}{Modelled\ variance} \tag{2.19}$$

where

$y'_{n\ or\ i}$ is the improved estimate for person n or item I,

and $y_{n\ or\ i}$ is the current estimate for person n or item i.

For the Rasch-Andrich thresholds of the polytomous items, the improved estimates are obtained by equation 2.20:

$$y'_x = y_x - \log(\frac{Observed\ count_x}{Observed\ count_{x-1}}) + \log(\frac{Estimated\ count_x}{Estimated\ count_{x-1}}) \tag{2.20}$$

where

$y'_k$is the improved estimate for Rasch-Andrich thresholds x,

y_k is the current estimate for Rasch-Andrich thresholds x.

2.3 Quality control procedures

2.3.1 Rating scale category effectiveness

2.3.1.1 Basic concept of scale categories and category thresholds

Within the framework of CTT, the labelled categories of a rating scale such as Likert scale (Likert, 1932) and 9-point hedonic scale (Peryam and Pilgrim, 1957) are usually considered as equally spaced single points on the scale. Thus the arbitrary numbers can be assigned to the categories to record people's responses in the format of raw scores. Thereafter the person's location on the scale can be estimated based on the raw scores.

By contrast, in Rasch analysis, each scale category represents an interval along the measurement continuum. Not only the persons but also the items and the other elements in additional facets are modelled on the same continuum. The transitions between the categories could be described using the threshold statistics. The names of three threshold statistics are often seen in the literature, which conceptualise the rating scale in different ways with the same measurement information (*i.e.* the response data) (Linacre, 2014c).

(1) Rasch-Andrich threshold

The Rasch-Andrich threshold is the transitioning point of two adjacent scale categories, where the probability of either category being used by the respondent is equivalent. It is the parameter τ in the polytomous Rasch model (see section 2.1.2.2). Figure 2.1 illustrates an example of the Rasch-Andrich thresholds on the category probability plot, which is a graph shows the expected probability of each scale category being chosen by the person.

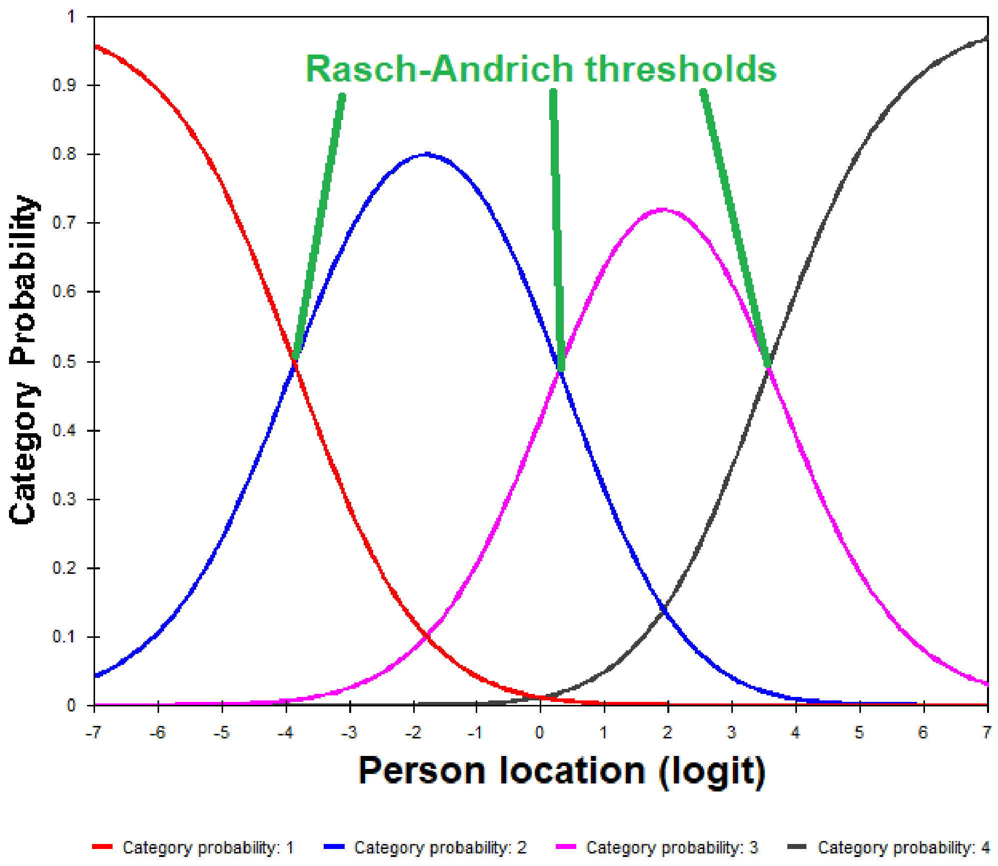


Figure 2.1 Category probability plot and Rasch-Andrich thresholds

The estimates of Rasch-Andrich thresholds are the x-values of the transitioning point

(2) Rasch-half-point threshold

The Rasch-half-point threshold is also called “average score threshold”. It is the boundary between the two categories, where the expected measure corresponds to the expected 0.5 score point. It is usually labelled in the score ogive curve, which reflects the transformation of ordinal non-linear raw scores to interval linear Rasch measures. Figure 2.2 shows an example of Rasch-half-point thresholds using the expected score ogive[5].

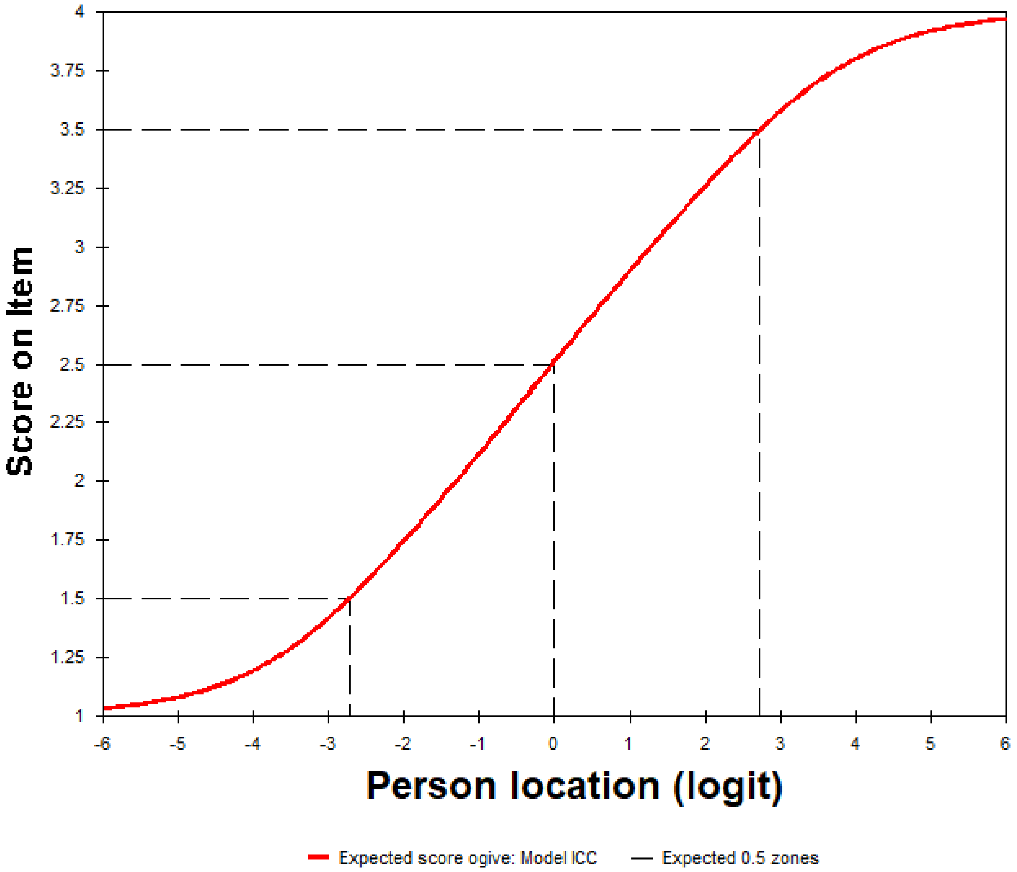


Figure 2.2 Expected score ogive and Rasch half-point thresholds

The end points of the dash lines on x-axis indicate the Rasch half-point thresholds

(3) Rasch-Thurstone threshold

The Rasch-Thurstone threshold is the median cumulative probability point, where there is 50% chance for the lower categories to be rated and 50% chance for all other categories that represent higher levels of the measurement to be rated.

[5] It is also called model item characteristic curve (ICC)

Figure 2.3 displays an example of the Rasch-Thurstone thresholds using a cumulative probability plot.

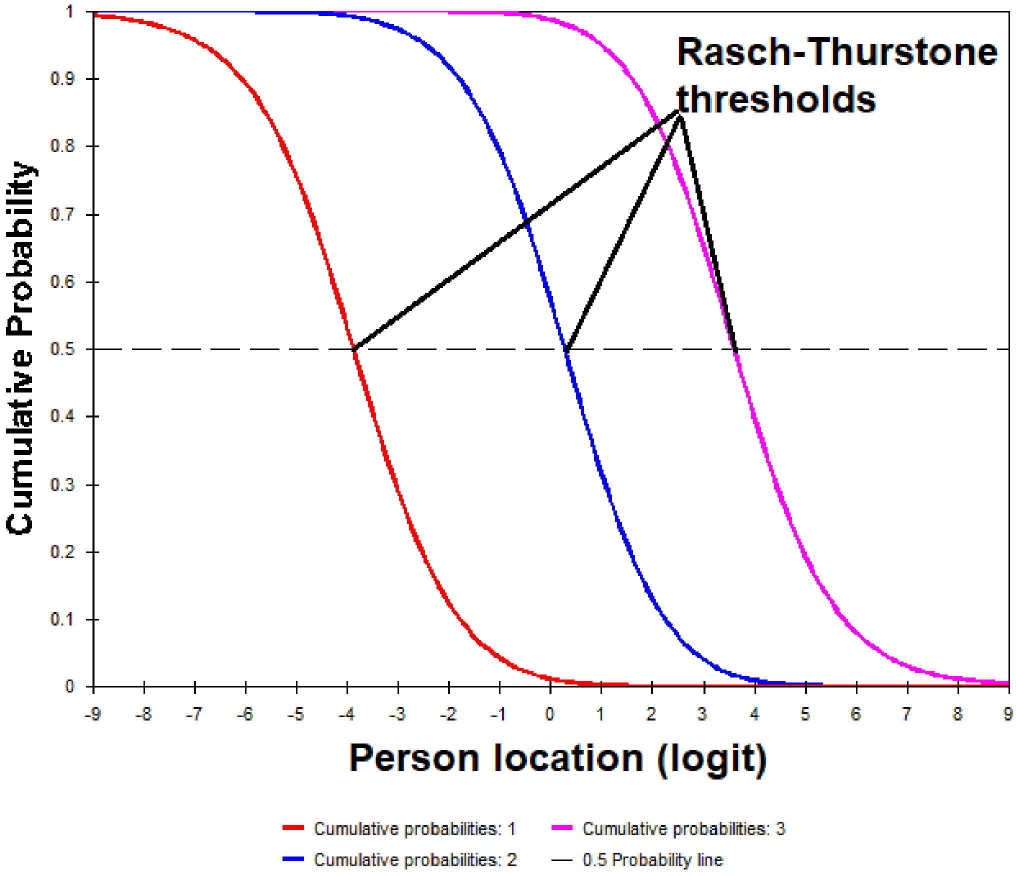


Figure 2.3 Cumulative probability curve and Rasch-Thurstone thresholds

The estimates of Rasch-Thurstone thresholds are the x-values of the labelled point

2.3.1.2 Evaluating of rating scale category effectiveness

When using CTT approach for data analysis, there is no formal procedure to evaluate whether the rating scale is functioning properly (Petrillo *et al.*, 2015). However, as part of the measurement device, how well the rating scale are functioning would have an influence on the quality of measurement. An important source of item misfit is affiliated with respondents' use of the rating scale in an inconsistent manner. (Pallant and Tennant, 2007). In Rasch analysis, additional quality control procedures have been suggested for evaluating rating scale categories effectiveness. Linacre (2002a) proposed a set of criteria for the examination, which can be classified as essential criterion or helpful criterion. Table 2.2 depicts the details of these criteria, which were employed in all four case studies of this research for the diagnosis of rating scale category effectiveness.

Table 2.2 Criteria for diagnosis of the rating scale category effectiveness – adapted from Linacre (2002a)

Criteria		**Measure Stability**	**Measure Accuracy (Fit)**	**Description of this sample**	**Inference for next sample**
prerequisite	Scale oriented with construct	Essential	Essential	Essential	Essential
1	At least 10 observations of each category.	Essential	Helpful		Helpful
2	Regular observation distribution.	Helpful			Helpful
3	Average measures advance monotonically with category.	Helpful	Essential	Essential	Essential
4	OUTFIT mean-squares less than 2.0.	Helpful	Essential	Helpful	Helpful
5	Rasch-Andrich thresholds advance.				Helpful
6	Ratings imply measures, and measures imply ratings.		Helpful		Helpful
7	Rasch-Andrich thresholds advance by at least 1.4 logits[1].				Helpful
8	Rasch-Andrich thresholds advance by less than 5.0 logits	Helpful			

[1] For a 3-point rating scale

(1) The mean measures of the rating scale categories should monotonically advance

The rating scale categories should follow a monotonic advancing order, because the higher categories should represent higher performance level on the scale. In practice, this can be verified by comparing the average measure of persons who respond to the items using the category.(Linacre, 2014c). This is because in theory, the persons who choose higher scale categories should have higher mean measures.

Disordering in the average measures of categories may be observed if the categories were not clearly defined. Table 2.3 displays an example provided by Linacre (2014c), where the categories with similar meanings were used. In this example, the meanings of “Occasionally” and “Sometimes” are close, while the other two categories “Often” and “Frequently” are similar. The respondents may interpret them in different way, resulting in the disorder issue.

In addition, this issue may also happen if a category was only used by a few persons, including someone with unexpected measure. If it could be used more frequently, then the impact of unexpected measure to the means would be much smaller. Therefore, before examining the ordering of the mean measures of categories, one should check if there are enough observations for every category (*i.e.* at least 10 observations).

Table 2.3 Example of probablematic rating scale with unclear categories

Scale category	**Raw score**
Never	1
Rarely	2
Occasionally[1]	3
Sometimes[1]	4
Often[2]	5
Frequently[2]	6
Always	7

The two pairs of scale categories which may make respondents confused are labelled in 1 and 2, respectively.

(2) Rasch-Andrich thresholds should monotonically advance

As noted earlier in section 2.3.1.1, the Rasch-Andrich thresholds are the transitioning points between adjacent categories, where the adjacent categories share the same probability of being observed. Therefore, the Rasch-Andrich thresholds should advance monotonically, consistent with the levels of the construct being measured (Andrich, 2013; Pallant and Tennant, 2007).

Disordered Rasch-Andrich thresholds may indicate that some of the categories only occupied a narrow section on the scale. This means the respondents might not be able to effectively use the rating scale system, or there were too many categories. This can be revealed in the category probability plot. Figure 2.4 shows an example of disordered Rasch-Andrich thresholds in a 7-point rating scale, where the threshold between category 4 and 5 is smaller than that between category 3 and 4 on the plot, and category 4 is never modal.

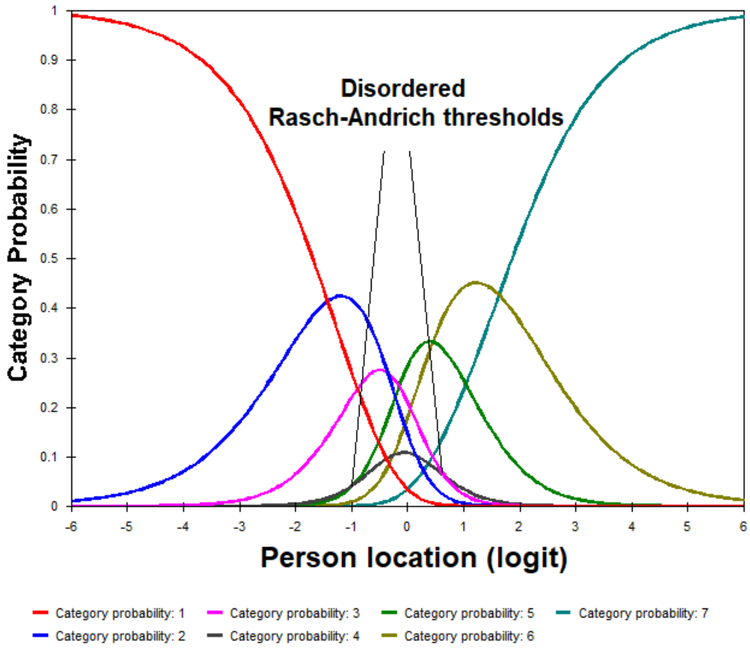


Figure 2.4 Category probability plot and disordered Rasch-Andrich thresholds

In addition, since the Rasch-Andrich threshold is a parameter of polytomous Rasch models, its standard error could be estimated at the individual level (see section 2.3.6). Salzberger (2015) suggested that one can test whether the Rasch-Andrich-threshold advance monotonically by performing t-tests to compare the adjacent Rasch-Andrich thresholds using their estimates and standard errors.

(3) Rasch-Andrich thresholds should not only monotonically advance, but also advance in a certain range

As part of the criteria (table 2.2), Linacre (2002a) suggested that the advancing distance between the adjacent Rasch-Andrich thresholds should be in a reasonable range.

In theory, a polytomous item could be decomposed into several independent dichotomous items. Therefore, Linacre (2006) proposed using the method described below to calculate the lower limit of advancing distance for Rasch-Andrich thresholds, if the decomposition of the polytomous item to the dichotomous items can be made (using a 5-point rating scale as example):

Step 1: Calculating the required Rasch-Andrich thresholds for the decomposition using the equation 2.21:

$$\ln(\frac{x}{m-x+1}) \tag{2.21}$$

where

x is the category before the 4 transitioning point (in this case:1~4),

and m is the number of Rasch-Andrich thresholds (in this case: 4).

The required Rasch-Andrich thresholds are -1.386, -0.405, 0.405, 1.386.

Step 2: Comparing the advancing distances between the adjacent Rasch-Andrich thresholds calculated using equation 2.21. The lower limit of the advancing distance between the adjacent ones is 0.81 in this example.

Table 2.4 tabulates the minimum advancing distance calculated in this way for Rasch-Andrich thresholds in 3-~11-point rating scales according to Linacre (2006).

Table 2.4 Minimum Rasch-Andrich threshold advances (Linacre, 2006)

Number of Categories	**Minimum advancing distance between Rasch Andrich Thresholds**								
	1 to 2	**2 to 3**	**3 to 4**	**4 to 5**	**5 to 6**	**6 to 7**	**7 to 8**	**8 to 9**	**9 to 10**
3	**1.39**								
4	**1.10**	**1.10**							
5	0.98	**0.81**	0.98						
6	0.92	**0.69**	**0.69**	0.92					
7	0.88	0.63	**0.58**	0.63	0.88				
8	0.85	0.59	**0.51**	**0.51**	0.59	0.85			
9	0.83	0.56	0.47	**0.45**	0.47	0.56	0.83		
10	0.81	0.54	0.44	**0.41**	**0.41**	0.44	0.54	0.81	
11	0.80	0.52	0.42	0.38	**0.36**	0.38	0.42	0.52	0.80

The minimum advancing distances for 3- to 11-point rating scales are labelled in **bold**.

In addition, the advancing distance cannot be too large, otherwise the rating scale will lose its function of discriminating people.

2.3.1.3 Optimising rating scale category effectiveness

If an issue related to rating scale categories is found, then the rating scale should be optimised, depending on the cause of the issues. In most cases, collapsing categories may eliminate the issues. This can be done by recoding the adjacent categories using the same raw scores.

There are a few considerations regarding collapsing categories:

(1) Collapsing the less frequently used categories;

(2) Collapsing the categories associated with disordered mean measure of categories or Rasch-Andrich thresholds;

(3) Renaming the collapsed category in a manner consistent with the rating scale.

2.3.2 Local independence

2.3.2.1 Definition of local independence

Early discussions of local independence can be traced back to Lazarsfeld (1959) in his work about latent class test. A widely accepted definition of local independence given by Lord and Novick (1968) is, within any sample characterised by the same construct, "*the (conditional) distributions of the item scores are all independent of each other*". It is a basic assumption of Rasch analysis and item response theory[1]. It is also considered as a prerequisite in CTT (Lord and Novick, 1968).

Two types of local independence are often discussed in the literature, including trait independence (in other words, unidimensionality) and response independence (namely local item independence).

(1) Unidimensionality

Unidimensionality can be defined as a single construct being able to account for the performance of a set of items (Brentani and Golia, 2007). Thurstone (1931a) stated that unidimensionality is a "*universal characteristic of all measurement*". The violation of unidimensionality may produce biased results in the estimation of model parameters.

[1] Another type of psychometric test theory

(2) Local item independence

Local item independence means the items are related to each other only through the construct. It requires that the responses to an item should be independent from the responses to the other items in a measurement. The violation of the assumption of local item independence is often called local item dependence (LID). The main impact of LID to the measurement is the reliability of the measurement would inflate artificially (Marais and Andrich, 2008). As a consequence, the estimates of model parameters and item discrimination would be biased, therefore the results obtained from succeeding statistical analysis would be misleading (Wang *et al.*, 2005).

2.3.2.2 Dimensionality test in CTT

In CTT, the dimensionality of the construct is primarily evaluated by performing the factor analysis[2] approach (DeVellis, 2006).

Factor analysis is based on the correlation matrix between variables. The two basic types of factor analysis are exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) (Pett *et al.*, 2003). The EFA can be used for exploring the unknown underlying dimension among items, while the CFA can be applied for verifying the known underlying structure of an instrument. In this research, only the EFA approach had been used (in the case study I). Therefore, only the procedures of conducting EFA are described here, which can be divided into four steps:

Step 1. Determining the appropriateness of conducting EFA

A fundamental assumption of EFA is that there exists a set of interpretable underlying factors, which can explain the interrelationships among a larger set of observed items (Kim and Mueller, 1978). This assumption requires that there are significant correlation among the items. To test that, a few approaches had been developed. The two commonly used methods are Kaiser-Meyer-Olkin test for sampling adequacy (Kaiser, 1970; Kaiser and Rice, 1974) and Bartlett's test of sphericity (Bartlett, 1950)

(1) Kaiser-Meyer-Olkin test for sampling adequacy (KMO test)

The KMO test produces a measure (range 0 to 1) that summarises the proportion of variances which can be explained by underlying factors. This measure is called

[2] It should be noted that, in a broad sense, the factor analysis is an extension of CTT, which decomposes the true score into several components associated with underlying factors (see section 2.1.1).

the measure of sample adequacy (MSA). The MSA can be computed using the equation 2.22:

$$KMO = \frac{\sum(correlations)^2}{\sum(correlations)^2 + \sum(partial\ correlations)^2} \quad (2.22)$$

Two types of MSA indices can be obtained. The overall MSA is calculated using the correlation and the partial correlation between all items, while the individual MSA is computed using those correlations involving the particular item.

MSA is an indicator of how small the partial correlation is within the data. It would suggest that there is at least one underlying factor, if the partial correlation is close enough to 0. The smaller the partial correlation, the closer the MSA to 1, thus the data set is more suitable for factor analysis. (Kaiser, 1974) suggests that data with an overall MSA greater than 0.5 can be considered as barely acceptable for factor analysis. Hutcheson and Sofroniou (1999) further recommend that the KMO index is "*mediocre when between 0.5 and 0.7, good when between 0.7 and 0.8, great when between 0.8 and 0.9, and superb when above 0.9*". In this research, a value greater than 0.6 was used as the criterion that suggests the sampling is adequate for both overall MSA and individual MSAs, following the recommendation of Pett *et al.* (2003).

(2) Bartlett's test of sphericity

Bartlett's test of sphericity is a chi-square test, which tests the null hypothesis that there is no relationship among the items. If the test statistic is not significant, then the EFA should not be performed.

Step 2. Extracting the initial factors

If both KMO test and Bartlett's test of sphericity imply that the data is appropriate for EFA, then the next step is extracting the initial factors. A few extraction techniques have been proposed, such as principal axis factoring (PAF) developed by Thurstone (1947) and maximum likelihood method. They all start with the assumption that the initial extractable factors are uncorrelated (Pett *et al.*, 2003).

A discussion for the procedures of the extraction methods and the difference between them is not provided here because it is beyond the scope of this thesis. This part of information has been well documented in the literature such as Kline (1994).

Step 3. Rotating the factors

After the initial factors have been extracted, an arithmetic procedure (namely factor rotation) to simplify the structure of factor matrix can be performed. The varimax rotation (Kaiser, 1958) is arguable the most common choice (Costello and Osborne, 2005).

Step 4. Evaluating the factor loadings and summarising the meanings of the factors

The relationship between each item to each underlying factor is expressed by the factor loading, which can be can be used to decide the composition of the factors. A factor loading equal to or greater than 0.30 implies that this factor has an effect on the item. The item that has a loading less than 0.30 on all extracted factors should be removed at this step, unless it is believed to have a unique contribution to the instrument (Hair, 1995). The other items need to be reviewed so that they can be used for interpreting the extracted factors. A name can be given to each factor according to the common meanings of the items that made up of the factor.

2.3.2.3 Tests of unidimensionality in Rasch analysis

A number of methods for examining the assumption of unidimensionality have been developed in Rasch analysis. Early proposals, such as assessing the fit of person and items (Andrich, 1988), and using a reliability-based unidimensionality coefficients (Wright, 1994b), had been discarded due to their deficiencies. Nowadays, the method of applying principal components analysis on standardised model residuals (PCAR) followed by independent t-tests has been broadly accepted as most appropriate method for detecting dimensionality of the instrument in Rasch analysis.

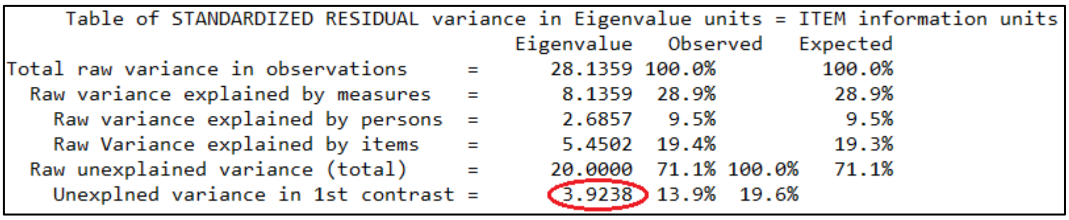
(1) PCAR

The use of principal component analysis (PCA) and factor analysis in the framework of Rasch analysis were firstly proposed by Smith (1996) and Wright (1996a). Later, Linacre (1998) suggested that the residual-type data should be used for this purpose. Smith (2002) introduced an independent t-test approach as a supplement to the PCAR approach, which was further improved by Tennant and Conaghan (2007).

It should be noted that PCAR approach is not the same as normal PCA because the standardised residuals are used instead of raw data here. Therefore, the result cannot be interpreted in the same way as the normal PCA. The hypothesis of the test by default is the measurement is unidimensional, as there is an underlying component (*i.e.* the Rasch dimension) that can explain most of the variance. If the assumption of unidimensionality could hold, then the residuals should only reflect random noise.

In WINSTEPS, the first extracted PCAR contrast is the residual contrast, where the Rasch dimension has already been removed. An example about how to interpret the PCAR results is given below, using the output from WINSTEPS for illustration purpose (figure 2.5~2.7):

After conducting PCA on Rasch residuals, the eigenvalues of explained and unexplained variances should be looked at first (figure 2.5). The raw variance explained by measures refers to the Rasch dimension, which is removed by default in WINSTEPS. The unexplained variances are originated from the additional dimensions and the random errors. The first key statistic in the results is the eigenvalue of unexplained variance of the first PCAR contrast. If it is small enough, then the residuals can be considered at random noise level. A value greater than 3.00 (*i.e.* the strength of 3 items) implies that the residuals of some items share the same pattern. This is an indicator of the existence of an additional dimension. In this example, the eigenvalue of 3.9238 suggests that there might be a second dimension.



```
     Table of STANDARDIZED RESIDUAL variance in Eigenvalue units = ITEM information units
                                           Eigenvalue   Observed   Expected
Total raw variance in observations     =      28.1359 100.0%         100.0%
  Raw variance explained by measures   =       8.1359  28.9%          28.9%
    Raw variance explained by persons  =       2.6857   9.5%           9.5%
    Raw Variance explained by items    =       5.4502  19.4%          19.3%
  Raw unexplained variance (total)     =      20.0000  71.1% 100.0%   71.1%
    Unexplned variance in 1st contrast =       3.9238  13.9%  19.6%
```

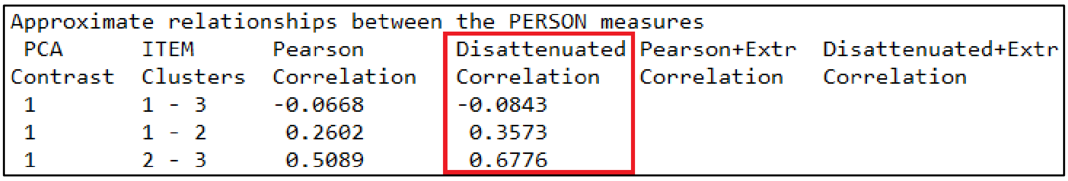
Figure 2.5 Standardised residual variance in eigenvalue units

(the eigenvalue of the first PCAR contrast is circled in red)

If the eigenvalue of the first PCAR contrast is greater than 3.00, then the disattenuated correlation coefficient between item clusters (figure 2.6) should be inspected next. WINSTEPS discerns the items into two or three clusters, including item cluster 1 made up of the items with highest loadings on the first PCAR contrast and those with lowest loadings in cluster 3. The remaining items would be assigned to cluster 2 (sometimes there is no item assigned to cluster 2). The disattenuated correlation coefficients between item clusters are calculated after removing the error variance. It is reported in the range of -1~1. Table 2.5 depicts the interpretation of the their values recommended by Linacre (2014c).

In the example (figure 2.6), we can conclude that cluster 2 and 3 (value=0.6776) are highly correlated, whereas the cluster 1 does not measure the same thing with neither cluster 2 or 3.The whole item set could be split into two subsets (figure 2.7) for further analysis. They are:

(1) Subset 1: Items in Cluster 1; and
(2) Subset 2: Items in Cluster 2 and 3.



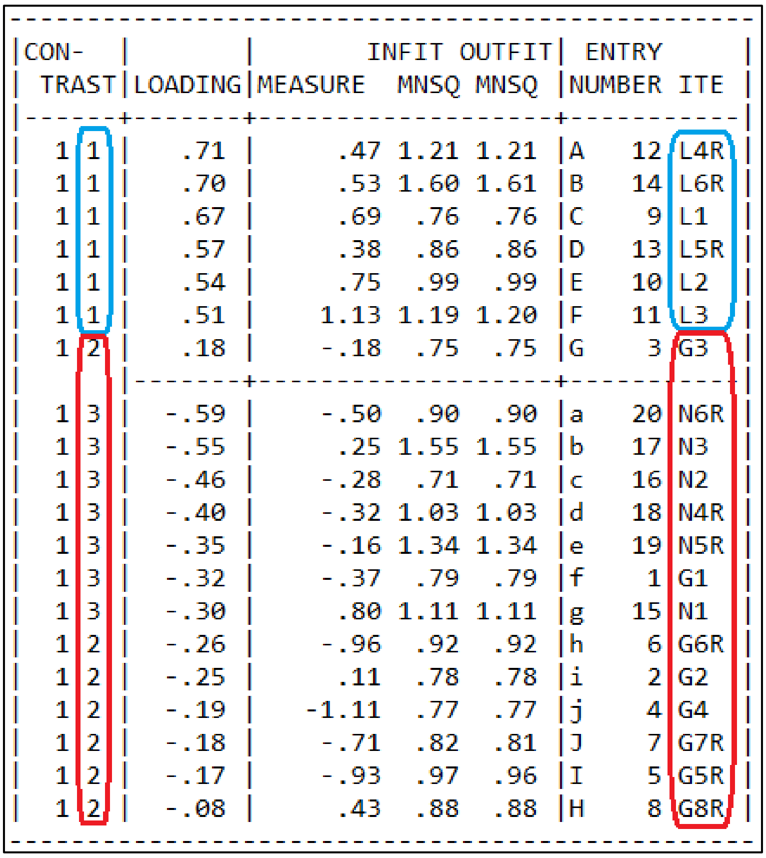
```
Approximate relationships between the PERSON measures
 PCA       ITEM      Pearson        Disattenuated Pearson+Extr  Disattenuated+Extr
Contrast   Clusters  Correlation    Correlation   Correlation   Correlation
 1         1 - 3     -0.0668        -0.0843
 1         1 - 2      0.2602         0.3573
 1         2 - 3      0.5089         0.6776
```

Figure 2.6 Correlations between item clusters on first PCAR Contrast

The disattenuated correlations between item clusters are marked in red

Table 2.5 Interpretation of disattenuated correlation coefficient (Linacre, 2014c)

Disattenuated correlation coefficient	**Interpretation**
<0.57	Cut-off point for the conclusion that the clusters measure different thing
0.71	The clusters are more dependent than independent
0.82	Indicative cut-off point for the conclusion that the clusters measure the same thing
0.87	Definitive cut-off point for the conclusion that the clusters measure the same thing



```
------------------------------------------------------
|CON-  |       |         INFIT OUTFIT| ENTRY      |
| TRAST|LOADING|MEASURE   MNSQ MNSQ  |NUMBER ITE  |
|------+-------+---------------------+------------|
|   1 1|   .71 |      .47 1.21 1.21  |A    12 L4R |
|   1 1|   .70 |      .53 1.60 1.61  |B    14 L6R |
|   1 1|   .67 |      .69  .76  .76  |C     9 L1  |
|   1 1|   .57 |      .38  .86  .86  |D    13 L5R |
|   1 1|   .54 |      .75  .99  .99  |E    10 L2  |
|   1 1|   .51 |     1.13 1.19 1.20  |F    11 L3  |
|   1 2|   .18 |     -.18  .75  .75  |G     3 G3  |
|      |-------+---------------------+------------|
|   1 3|  -.59 |     -.50  .90  .90  |a    20 N6R |
|   1 3|  -.55 |      .25 1.55 1.55  |b    17 N3  |
|   1 3|  -.46 |     -.28  .71  .71  |c    16 N2  |
|   1 3|  -.40 |     -.32 1.03 1.03  |d    18 N4R |
|   1 3|  -.35 |     -.16 1.34 1.34  |e    19 N5R |
|   1 3|  -.32 |     -.37  .79  .79  |f     1 G1  |
|   1 3|  -.30 |      .80 1.11 1.11  |g    15 N1  |
|   1 2|  -.26 |     -.96  .92  .92  |h     6 G6R |
|   1 2|  -.25 |      .11  .78  .78  |i     2 G2  |
|   1 2|  -.19 |    -1.11  .77  .77  |j     4 G4  |
|   1 2|  -.18 |     -.71  .82  .81  |J     7 G7R |
|   1 2|  -.17 |     -.93  .97  .96  |I     5 G5R |
|   1 2|  -.08 |      .43  .88  .88  |H     8 G8R |
------------------------------------------------------
```

Figure 2.7 Standardised residual loadings for Item on first PCAR Contrast

The Items belonging to cluster 1 are circled in blue, while the items belonging to cluster 2 and 3 are circled in red

The same procedure can be repeatedly performed on the split subsets individually, until the eigenvalue of the first PCAR contrast is found being around or smaller than 3.00, which implies that the instrument might be unidimensional.

After conducting a PCAR test, the independent t-test protocol can be performed to further confirm the unidimensionality. It requires the researchers to split the items into subsets according to their loadings on the first extracted PCAR contrast. A series of t-tests would be conducted to compare every person’s measure modelled from the subset of items with highest positive loadings (≥0.30) and that modelled from the subset of items with lowest negative loadings (≤-0.30) after test equating. The rationale behind the method is, if the two subsets of items indeed belong to the same dimension, then the persons’ measures modelled using the two subsets should be same. The proportion of significant t-test results would be counted first, then a binomial 95% confidence interval (CI) would be calculated based on the counts. If less than 5% person-by-person t-tests were significant, or the proportion was overlapped by the lower bound of the binomial 95% CI, then one can draw the conclusion that the instrument is unidimensional.

2.3.2.4 Test of local item independence

Despite the fact that the violation of local item independence may affect the quality of the measurement, the local item independence is only implicitly assumed within the framework of CTT (Lee, 2004). By contrast, it is an essential assumption in Rasch analysis (Christensen *et al.*, 2017). A number of methods have been developed for testing this assumption since 1980s (Cohen, 1988; Haberman, 2007; Van den Wollenberg, 1982; Yen, 1984; Yen, 1993), among which the most popular approach is examining the correlation between item residuals after removing the dominant Rasch dimension[3]. Smith (2000) suggested that an item pair can be diagnosed as LID items if their residual correlation was equal to or greater than 0.30[4].

If local item dependence (LID) was indicated by the residual correlation, one should try to explain the LID issue by reviewing the item statements first. A few sources of LID items have been identified by previous research. For instance, LID may be associated with item presenting order (Royal, 2016). It could also be found if an item is included as part of another item (Wilson *et al.*, 1997). In addition, LID may be observed if a person's response to an item could provide cues for latter items (Marais and Andrich, 2008).

After that, one can eliminate the LID issue by combining the LID items to a super-item. The summate score of the LID items would be used as the score of the super-item. Since the range of the raw scores would change after combining the LID items, the revised data should be refitted to a partial credit model (Masters, 1982) for analysis. In some occasions, the LID issue may be ignored if the residual correlation is only slightly greater than the recommended threshold 0.3, and if the LID items were designed to measure unique and important properties.

2.3.3 Fit

2.3.3.1 Fit statistics

Fit describes how well the observation conforms to the expectation in theory. CTT requires model fit data. In other words, the model with best data fit should be employed to describe the data. In contrast, in order to obtain invariant measures

[3] This method was initially proposed by Yen (1984) for item response theory model.

[4] Linacre (2014c) argued that a residual correlation of 0.4 implies low LID issue. It needs to be greater than 0.7 if one should concern about the LID. However this opinion has not been widely accepted.

across population and items, Rasch analysis does not allow altering model for fitting data.

Despite the fact that there is no data that can perfectly fit the model, the degree of the discrepancy between the model and data matters. Four indices of individual fit for persons, items, or additional facets (when using MFR model) are reported by WINSTEPS and Facet.

Among the four fit statistics, the two mean-square (reported under the name of MNSQ in WINSTEPS and Facets) statistics are computed first. They are the chi-square values divided by the degrees of freedom, reflecting the relation between the observed data and the model. The range of MNSQ statistic is 0 to infinite. It has an expected value of 1 and expected standard deviation of 0. A value less than 1 means the observation is overfit, indicating the item or person may be too predictable. In CTT or any other theory requires model fit data, overfit is good. In Rasch analysis, however, overfitting items or persons are inefficient in the measurement. On the contrary, a MNSQ value greater than 1 implies that the observation is underfit (in other words, unpredictable). Compared to overfit, underfit is more problematic because it reflects the distortion of observed data from model. The higher the MNSQ value, the larger degree of distortion is found. For example, a MNSQ value of 1.2 means the observed variance is 20% higher than the expected variance.

After MNSQ values are reported, the z-standard (reported under the name of “ZSTD” in WINSTEPS and Facets) statistics can be calculated by converting the MNSQ values to their z-scores via Wilson-Hilferty transformation (Wilson and Hilferty, 1931). ZSTD is a t-statistic that shows the statistical significance of MNSQ occurring when data fit the model (Linacre, 2014c). It has a range of negative infinite to positive infinite, with the expected value of 0 and the standard deviation of 1. Negative value suggests overfit, whereas positive value indicates underfit.

Two types of MNSQ and ZSTD statistics can be estimated, which represent outlier-sensitive and inlier-sensitive fit statistics, respectively. The outlier-sensitive fit statistics (namely outfit statistics) are unweighted fit indices. They are more sensitive to extreme unexpected responses compared to the inlier-sensitive fit statistics (Wright and Masters, 1982). For instance, the outfit MNSQ is calculated by taking the average of the squared standardised residuals. So it is not affected by other information such as response patterns.

On the other hand, the infit statistic is information-weighted. It is less influenced by the unexpected responses near the measure rather than those largely apart from the measure. When computing the infit MNSQ statistic, the residuals are weighted by their variances. In the inspection of fit statistics, outfit indices should be scanned first for the purpose of detecting misfitting items or persons. Linacre (2014c) further pointed out that there is no need to report infit statistics unless "*the data are heavily contaminated with irrelevant outliers*".

Using item fit statistics as examples[5], in a measurement involves N respondents, the four indices for item i can be calculated using the equations 2.23~2.26:

(1) Outfit Mean-square(Outfit MNSQ)

$$u_i = \sum_{n=1}^{N} z_{ni}^2 / N \tag{2.23}$$

where
z_{ni} is the standardised residual of person n's response to item i.

(2) Infit Mean-square (Infit MNSQ)

$$v_i = \sum_{n=1}^{N} z_{ni}^2 W_{ni}^2 / \sum_{n=1}^{N} W_{ni} \tag{2.24}$$

where
z_{ni} is the standardised residual of person n's response to item I,
W_{ni} is variance of person n's response to item i.

(3) Outfit Z-Standardised (Outfit ZSTD)

$$t_u = (u_i^{1/3} - 1)/(3/q_i) + (q_i/3) \tag{2.25}$$

where
q_i is the standard deviation of u_i.

(4) Infit Z-standardised (Infit ZSTD)

$$t_w = (v_i^{1/3} - 1)/(3/q_i') + (q_i'/3) \tag{2.26}$$

where
q'_i is the standard deviation of v_i.

Table 2.6 depicts the guideline for interpretation of MNSQ and ZSTD values suggested by Linacre (2002b). It should be noted that ZSTD statistics are sample-

[5] For more information about the calculation of fit statistics, z_{ni}, q_i and q'_i, one can consult Wright and Masters (1982, 1990).

size dependent, thus it is only useful to "*salvage non-significant MNSQ>1.5, when the sample size is small or test length is short*" (Linacre, 2014c).

Table 2.6 Guidelines for the interpretation of MNSQ and ZSTD values, adopted from Linacre (2002b)

Index	Value	Implication for Measurement
MNSQ	> 2.0	Distorts or degrades the measurement system. May be caused by only one or two observations.
	1.5 to 2.0	Unproductive for construction of measurement, but not degrading.
	0.5 to 1.5	Productive for measurement.
	< 0.5	Less productive for measurement, but not degrading. May produce misleadingly high reliability and separation coefficients.
ZSTD	≥ 3.0	Data much unexpected if they fit the model (perfectly), so they probably do not. But, with large sample size, substantive misfit may be small.
	2.0 to 2.9	Data noticeably unpredictable.
	-1.9 to 1.9	Data have reasonable predictability.
	≤ -2.0	Data are too predictable. Other "dimensions" may be constraining the response patterns.

2.3.3.2 Resolving misfitting item issue

A few strategies could be employed if serious misfitting item is found.

(1) Dropping the extreme unexpected responses

A few odd responses may have a significant impact on the item fit. Therefore dropping the extreme unexpected responses associated with the most misfitting items may improve the fit. In practice, one can identify the extreme unexpected responses by inspecting the standardised residuals. The response that has an absolute value of the standardised residual equal to or greater than 2.0 should be removed from the data. After fitting the revised data to the model, one can re-evaluate the item fit to verify the effect of dropping extreme responses.

(2) **Dropping the item**

If dropping the extreme unexpected responses could not bring the item fit back to acceptable range, one should consider dropping the item.

2.3.3.3 Considerations on misfitting persons

The criteria of fit statistic are same for persons and items; however, it is expected that the items should fit the data better than persons. Misfitting person is inevitable, especially in consumer research, where a great number of respondents would be investigated. Wright and Linacre (1994) suggested that a few misfitting persons would have negligible impact on the measurement because the number of persons would be much larger than the number of items. Therefore, although one should still inspect the person fit statistics, the misfitting persons may be retained if only a small proportion of them exhibited serious misfitting issue.

2.3.4 Differential item functioning

Differential item functioning (DIF) refers to the difference in the characteristics of an item between subgroups of people. In early literature, it was also called "item bias" (Lord, 1980). However, this term has been replaced by a more neutral term DIF very soon (Holland and Thayer, 1988). DIF is only a source of item bias. Clauser and Mazor (1998) pointed out that the exhibition of DIF is necessary but not sufficient condition for item bias.

Scheuneman (1975) described her understanding of item without DIF within the CTT context as "*An item is unbiased if, for all individuals having the same score on a homogeneous subtest containing the item*". Lord (1980) provided a definition of DIF in item response theory, which can also be applied to Rasch analysis, that the item is biased if it has a different item response function for one group than another.

Figure 2.8 illustrates an example of DIF using a modified Wright map. Comparing the locations of the items on the scale, one can clearly see that item I03 measured by females locates at a relatively low position on the scale compared to other items in the item set, in contrast to its location at the medium level measured by males. It is suspected of DIF.

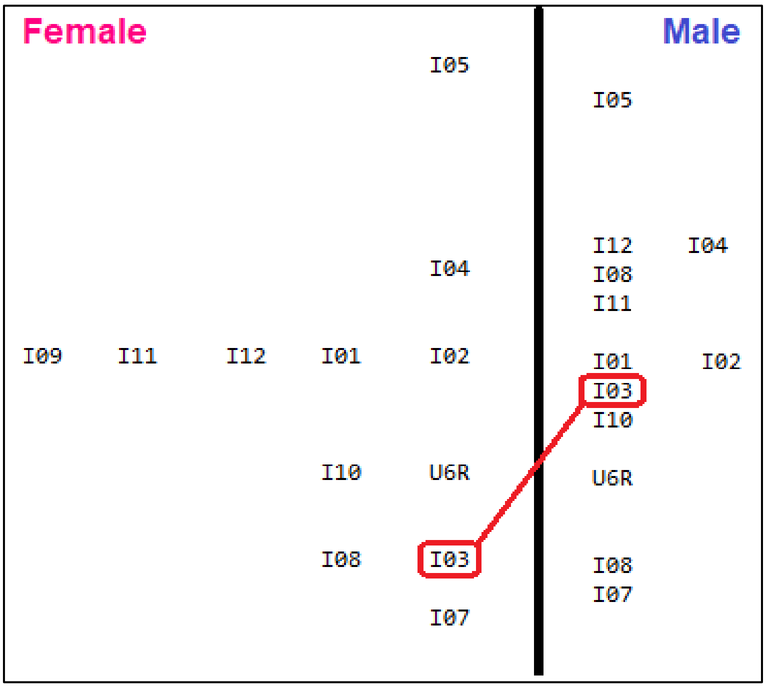


Figure 2.8 An example of DIF item

DIF can be discerned into uniform DIF and non-uniform DIF. The former one refers to the scenario that the item response function deviates consistently across all construct level between the groups. The latter one, means the differences are not constant between groups across the construct levels.

2.3.4.1 Statistical DIF analysis in CTT

Within the framework of CTT, several approaches have been developed as the method for detecting DIF items, such as the Mantel-Haenszel test (M-H test) for dichotomous items (Mantel and Haenszel, 1959) and its extension for polytomous items (Mantel, 1963), Scheuneman's Chi-Square method (Scheuneman, 1979), Logistic regression (Swaminathan and Rogers, 1990), and Simultaneous Item Bias Test (SIBTEST) devised by (Shealy and Stout, 1993).

Among these methods, the M-H test has been implemented in both CTT and Rasch analysis. It is a chi-square test that compares the odds ratios of a series of two-by-two tables. Since Holland and Thayer (1988) proposed to use the M-H test statistic for detecting DIF, it has been widely used for this purpose. Instead of comparing the whole groups, it detects the group difference on an item by comparing persons at similar construct levels in each group. The odds ratios are summated to obtain an estimate of overall DIF.

It should be noted that the CTT-based DIF detection methods shared two common problems (DeVellis, 2006):

(1) CTT is sample-dependent, thus the results of DIF detection may vary between populations.

(2) The scores at the centre of the scale are more sensitive to the change on the variables than those at the two ends of scale.

Therefore, DeVellis (2006) suggested the CTT-based methods may not work well for DIF detection . In this research, the DIF detection was only conducted in Rasch analysis.

2.3.4.2 Statistical DIF analysis in Rasch analysis

In Rasch analysis, the following methods have been developed for detecting DIF:

(1) Sample-based effect size of DIF contrast

The difference between the estimates of individual items by subgroups is provided by software WINSTEPS and Facet under the name of “DIF contrast”. The sample-based effect size of DIF contrast can be estimated using the absolute value of the DIF contrast divided by the pooled standard deviation (Linacre, 2014c):

$$Effect\ size_{DIF} = \frac{|DIF\ contrast|}{\sigma_{pooled}} = \frac{|D_{i1} - D_{i2}|}{\sqrt{\frac{(n_1-1)\sigma_1^2 + (n_2-1)\sigma_2^2 + \cdots + (n_k-1)\sigma_k^2}{n-k}}} \qquad (2.27)$$

where

D_{ij} is the estimated measure of item i for group j,

σ_{pooled} is the pooled standard deviation,

k is the number of the groups,

n_1, n_2, …, n_k are the size of the k groups,

σ_1, σ_2, …, σ_k are the standard deviation of each group.

A sample-based effect size of greater than 0.5[6] can be considered as evidence of the existence of significant DIF.

(2) Mantel-Haenszel (M-H) test

Linacre and Wright (1989b) proved that if the data fit to the Rasch model, then the M-H test can be employed in Rasch analysis using measures as estimators

[6] 0.5 represent a medium effect size (Cohen, 1988)

instead of the raw scores. WINSTEPS and Facets have also integrated the M-H test as a function. It should be noted that the computation of M-H statistics differs in CTT and Rasch analysis if there are missing data. In CTT, the samples are sliced into strata by raw scores after deleting the cases with missing values, whereas in Rasch analysis, there is no deletion. The samples are sliced into strata by the measures estimated with the presence of missing data.

(3) Welch t-test

Rasch analysis estimates the measure and standardised error (SE) of items on an individual basis. Therefore, the group-wise Welch t-test (Welch, 1947) that compares the DIF contrast between subgroups can also be used for DIF detection. The t statistic can be calculated using the equation 2.28:

$$t = \frac{|D_{i1} - D_{i2}|}{\sqrt{SE_{i1}^2 + SE_{i2}^2}} \quad (2.28)$$

where

D_{ij} is the estimated measure of item i for group j,

SE_{ij} is the standard error of D_{ij}.

Software WINSTEPS and Facets have integrated the function of reporting the estimated measure and standard error of individual items for each group. However, when conducting the Welch t-test, one should adjust the statistics for the effect of sample size because the standard error is dependent on the sample size (Linacre and Wright, 1989b).

(4) ANOVA on standardised residuals

For researchers using software RUMM2030, the two-way ANOVA on standardised residuals with an interaction term is usually conducted by the software for DIF detection, where the Bonferroni correction (Dunn, 1961) is applied to reduce type I errors.

In WINSTEPS and Facets, the Welch t-test and M-H test should provide similar results. Linacre (2014c) suggested that M-H test is more accurate when the data is complete and the sample size is large. However, one should not make decisions exclusively on the level of significance of the Welch t-test and M-H test. The sample-based effect size should also be taken into account because it is independent with sample size. Therefore, in this research, DIF items that needed to be reviewed were flagged if evidence was given by both significant M-H test and an effect size greater than 0.5.

2.3.4.3 Resolving DIF

If a significant DIF item is detected, one should try to explain the source of DIF by reviewing the item first. If a number of DIF items exist, one should also search for the common pattern in all DIF items (Hambleton, 2006). The information derived from the item review process would be instructive for future work in design of items.

After that, one of the following strategies can be used for resolving the DIF:

(1) Removing the item

This is the simplest way to eliminate the DIF. However, the removal of items is accompanied with the loss of information. Consequently, the reliability and validity of the measurement may be lower (Hagquist and Andrich, 2017; Hambleton, 2006). Therefore this method is not applicable when there are too many DIF items in a short instrument (Teresi *et al.*, 2008). One should monitor the change of measurement reliability before and after removing the DIF items. If the reliability decreases significantly after removing the DIF items, then the removal method should not be used.

(2) Discarding the responses to a DIF item given by one DIF group.

If the source of DIF could be clearly explained by the misunderstanding of a term by a group, which often happens due to language issues, then discarding the responses to the DIF item given by the whole group may reduce the impact of DIF on the measurement. This method was suggested by Linacre (2014c)

(3) Treating a DIF item as multiple items for each group.

Tennant *et al.* (2004) suggested that the DIF item can be treated as multiple items for each group. For example, if an item displays DIF between females and males, then the item would be split into two items. The raw scores given by all females were assigned to one item, and the responses obtained by all males would be record as another item. The other items are used for linking the two split items.

(4) Estimating persons' measures separately by DIF groups

Boone *et al.* (2013) introduced a two-step method for adjusting the person's estimates after the identification of a DIF item. Firstly, the data set is reanalysed without DIF items. Secondly, the measures of respondents are analysed separately by DIF groups on full instrument using a modified model, where the measures of non-DIF items and the Rasch-Andrich thresholds are anchored at the estimates obtained from first step after removing DIF items. After that, the DIF groups can be compared using the adjusted measures.

2.3.4.4 When should DIF be examined

A DIF item in an instrument is responded to in different ways by different groups. It is a source of bias in the measurement that targets the true difference between groups. Therefore one should always ensure that there is no serious DIF, or the DIF has been minimised, before comparing the designated groups of interest.

However, the detection of DIF items is dependent with how to classify people into groups. There are numerous ways to categorise people into subgroups, for instance, female vs. male, young vs. old, people who grew up in Yorkshire vs London. As a result, one may end up with all items displaying DIF. Therefore, if the research interest is not on comparing the difference between groups, then one may not need to examine DIF.

2.3.5 Reliability/Separation

2.3.5.1 Reliability in CTT

In CTT, reliability refers to the degree of consistency of measurement. The higher the reliability, the more likely a measurement can obtain the same results under identical conditions. It can also be defined as the proportion of total variances that can be explained by the true score (DeVellis, 2011). Under this definition, the reliability is expressed by the equation 2.29:

$$Reliability = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_X^2 - \sigma_E^2}{\sigma_X^2} \qquad (2.29)$$

where

$\boldsymbol{\sigma_T^2}$, $\boldsymbol{\sigma_X^2}$, $\boldsymbol{\sigma_E^2}$ are the variance of true score variable, observed score variable and error variable, respectively.

Unfortunately, neither the true score nor the error can be observed directly. So the "true" reliability cannot be calculated. A variety of approaches have been established within the framework of CTT for estimating the reliability. Table 2.7 tabulates the major types of reliability indices studied in CTT.

Table 2.7 Major types of reliability indices studied in CTT (adapted from DeVellis, 2011)

Types of reliability	**Concerns**	**Estimates of reliability**
Test-retest reliability	Stability of the test	Correlation between test and retest
Parallel forms reliability (alternative form reliability)	Test equivalence	Correlation between the two tests
Inter-rater reliability (inter-observer reliability)	The agreement between the raters/observers	Correlation between the raters/observers
Internal consistency reliability	The homogeneity of items	Average inter-item correlation Average item-total correlation Split-half correlation Cronbach's alpha/KR-20

The most commonly used reliability index is the Cronbach's alpha coefficient (Cronbach, 1951). It is an estimate of internal consistency of items. Internal consistency refers to the homogeneity of the items which measures the same underlying construct.

The equation used for calculating Cronbach's alpha is:

$$\alpha = \frac{K}{K-1}\left(1 - \frac{\sum_{i=1}^{k} \sigma_{y_i}^2}{\sigma_x^2}\right) \tag{2.30}$$

where

K is the number of items,

$\sigma_{y_i}^2$ is the variances associates with item I,

σ_x^2 is the total variances of observed scores.

Cronbach's alpha can also be computed using an alternative equation:

$$\alpha = \frac{K\bar{c}}{\bar{\upsilon}+(K-1)\bar{c}} \tag{2.31}$$

where

K is the number of items,

$\bar{\upsilon}$ is average item variance,

$\bar{c}$ is the average covariance.

It should be noted that the special form of Cronbach's alpha coefficient when the items are all dichotomous is also called Kuder-Richardson formula 20 or KR-20 (Kuder and Richardson, 1937).

Cronbach's alpha is the lower-bound of reliability of the test (Novick and Lewis, 1967). It has a range of 0 to 1. Low alpha values may indicate poor inter-item correlation. When evaluating the reliability of an instrument based on the alpha value, the recommendations from Nunnally and Bernstein (1994) are often followed:

(1) 0.70 is acceptable value for basic research. It is arguable that time and money may be wasteful if the targeted alpha value is higher than 0.80.
(2) 0.90 is the minimum required value for making important decisions about individuals. An alpha value reaching 0.95 or higher is desirable.

However, when using alpha, one should inspect the instrument with extra cautions:

(1) Alpha is not simply a function of internal consistency. It is affected by dimensionality. When multiple dimensions exist, alpha may underestimate the reliability (Green and Thompson, 2005). So it is necessary to evaluate the structure of the instrument before computing the alpha.

(2) Test length has an impact on alpha (Cronbach, 1951). Low alpha values may be caused by an insufficient number of items, whereas high alpha values can be achieved if the test is long enough, even when the instrument is composed of uncorrelated dimensions (Cortina, 1993). This can be deducted from the Spearman-Brown Prophecy formula (Brown, 1910; Spearman, 1910). Guilford (1954) illustrated that the true and error variances both increase if the test length is enlarged, but the former one increases more rapidly than the latter one. The desirable test length can be predicted by a form of the Spearman-Brown Prophecy formula (equation 2.32):

$$j = \frac{R_d(1-R_o)}{R_o(1-R_d)} \tag{2.32}$$

where
j is the predicted ratio of the desirable test length over the current test length,
R_d is the desired reliability,
R_o is the observed reliability of current test.

(3) In practice, if the assumption of unidimensionality could hold, an alpha greater than 0.90 may indicate redundancy of items (Streiner, 2003; Tavakol and

Dennick, 2011). The test can be shortened so that the respondents' burden can be reduced.

(4) The calculation of alpha relies on Pearson correlation coefficient computed on continuous interval data. In reality, however, the raw scores are often collected on discrete ordinal levels. The issue of treating ordinal data as interval data remains.

(5) The alpha is dependent on the true and error variance of a particular population under the framework of CTT. So it may vary between populations.

2.3.5.2 Reliability indices in Rasch analysis

In Rasch analysis, three reliability indices are reported by the software WINSTEPS and Facets.

(1) Separation reliability

Separation reliability is a Cronbach's alpha type of statistic. It is often simply called "reliability" in the literature. The alternative term "Separation Index" can be seen in documentation using the software RUMM2030. The formula for calculating the separation reliability is:

$$R = \frac{\sigma_{True}}{RMSE} = 1 - \frac{\sum(SE)^2/n}{\sigma^2_{Observed}} = 1 - \frac{RMSE^2}{\sigma^2_{Observed}} \tag{2.33}$$

where

R is the Separation reliability,

σ_{True} is the standard deviation of measures corrected for measurement error,

RMSE is the root mean square error,

$\sigma_{\mathrm{Observed}}$ is the observed standard deviation of measures,

SE stands for the model standard error of measurement or the real standard error of measurement reported (see section 2.3.6 for more information).

(2) Separation (separation ratio)

In WINSTEPS and Facets, a ratio-type statistic called "separation" (or "separation ratio") is also reported. It compares the dispersion of the measures with the measurement error. It is the predicted number of statistically distinct levels that can be identified in a sample when the tails of the distributions are considered as merely measurement error (Linacre, 2014b; Wright, 1996b).

$$G = \frac{\sigma_{True}}{RMSE} = \frac{\sqrt{\sigma^2_{Observed} - RMSE^2}}{RMSE} = \sqrt{\frac{R}{1-R}} \tag{2.34}$$

where

G is the separation,

RMSE is the root mean square error,

σ_{True} is the standard deviation of measures adjusted for measurement error,

$\sigma_{\mathrm{Observed}}$ is the observed standard deviation of measures,

R is the Rasch reliability.

(3) Strata

If the tails of sample distribution are treated as extreme levels, then the separation G can be translated into the third reported statistic “strata” (Linacre, 2014b; Wright and Masters, 2002). To cover the tails, the functional range of measures, which is around 4 σ_{True}, is inflated by 1 RMSE. Therefore the strata can be modelled using the equation 2.35, which is used for estimating the number of statistical different levels that can be separated by at least 3 RMSE:

$$H = \frac{\sigma_{True} \times 4 + RMSE}{RMSE \times 3} = \frac{G \times 4 + 1}{3} \tag{2.35}$$

where

σ_{True} is the standard deviation of measures adjusted for measurement error

RMSE is the root mean square error

G is the separation,

H is the strata.

Specifications of the Rasch reliability statistics

In Rasch analysis, the three reliability statistics are all reported in two versions: the “model” and “real” reliability, depending on whether the RMSE in the equations is calculated from the “model” standard error of measurement or the “real” standard error of measurement (see section 2.3.6). According to Linacre (2014c), the “model” reliability statistics are the upper-bound of the estimates of reliability, while the “real” reliability ones are the lower-bound of the estimates. Boone *et al.* (2013) recommended that for the measurement involved in decision making such as market research, the “real” reliability statistics should be used because it is more conservative, although only a slight difference can be found between them.

Rasch analysis estimates the model parameters separately. As a result, in addition to the person reliability which is equivalent to the concept in CTT, the item reliability (and the reliability for other facets using the Many-Facet Rasch model) can also be reported in Rasch analysis.

Interpretation of Rasch reliability statistics

Linacre (2014c) stated that person separation is used to "*classify people*", and the item separation is used to "*verify the item hierarchy*". He argued that low person separation indicates the instrument cannot efficiently distinguish persons at different levels on the scale, thus more items may be needed. Similarly, low item separation reflects that the sample size of the measurement is too small to confirm the item hierarchy of the instrument; therefore more participants are required.

Both separation and strata vary in the range of 0 to infinite. This breaks the restriction of using the reliability that has a value between 0 and 1. A separation statistic smaller than 1 (equivalent to reliability less than 0.5) indicates that the measurement error is the main source of the differences between the measures (Fisher, 1992). The higher the separation or strata, the more reliable results can the measurement produce.

In practice, when the separation statistic is used for evaluating the reliability, one can discriminate the sample into "High" and "Low" level groups with a separation G of 2.00. In addition, a value of 1.50 for separation G, which is equivalent to 2.33 for strata H or 0.69 for reliability was recommended by Tennant and Conaghan (2007) as the minimum requirement for group use. They also suggested that a separation of 2.50 is needed for individual use.

It is worth noting that the rule that longer tests are more reliable than shorter ones may not be always true under some special conditions in Rasch analysis. This is because the standard error of measurement modelled in Rasch analysis is dependent on construct levels (see section 2.3.6). For instance, an adaptive test that minimises these error may have greater reliability than a fixed content test with longer test length (Embretson, 1996).

2.3.6 Standard error of measurement (SE_m)

Standard error of measurement (SE_m) is an indicator of the dispersion of the measurement errors.

2.3.6.1 SE_m in CTT

In CTT, it is assumed that the measurement error is equally and normally distributed in the population, therefore the SE_m is constant across a given population (Embretson and Hershberger, 1999). It is generalised for the whole population by the equation 2.36:

$$SE_{measurement} = \sqrt{1-R} \times \sigma \tag{2.36}$$

where

R is the reliability of the measurement

σ is the standard deviation of the test

The estimate of SE_m in CTT applies to all scores across a given population. This has some disadvantages. Firstly, it reflects imprecision of measurement at the global level, whereas the individual’s imprecision cannot be interpreted from the statistic. Secondly, the reliability within the framework of CTT is dependent on the true and error variance of the population; therefore the reliability coefficients calculated for different populations may differ. Consequently the SE_m may vary between populations. Thirdly, the estimation of SE_m in CTT does not account for the differences between people’s response patterns. Lastly, the assumption of equal SE_m across the population may not hold (Harvill, 1991). Feldt *et al.* (1985) discovered that the SE_m may vary at different score levels within the same population.

2.3.6.2 SE_m in Rasch analysis

Unlike CTT, Rasch analysis focuses on the measurement of individuals. The SE_m conceptualised in Rasch analysis (and also the other item response theory models) is considered as a continuous function of the construct level (*i.e.* it is construct level dependent). To be more specific, it differs between persons at different construct levels, while persons at the same construct level would have the same SE_m (Embretson, 1999). Usually, it is higher for extreme scores than those at the centre of the construct continuum. In addition, since Rasch analysis obtains the sample-free estimates of parameters, the estimate of SE$_m$ does not vary across populations. Moreover, unlike CTT, where only a homogeneous SE$_m$ of the population is estimated, the estimates of SE$_m$ of other parameters such as item and Rasch-Andrich thresholds (for polytomous items) can be obtained in Rasch analysis.

According to the WINSTEPS manual (Linacre, 2014c), the modelled standard error of measurement (reported under the name "Model SE" in WINSTEPS and Facets) is computed using the equations 2.37~2.39:

If the data contains items i=1, L for person B_n, and person n=1, N for item D_i

(1) For dichotomous model

$$Model\ SE_{(B_n,\, D_i)} = 1/\sqrt{\sum(P_{ni}(1-P_{ni}))} \tag{2.37}$$

(2) For polytomous models[7] with categories x=0,m,:

$$Model\ SE_{(B_n,\, D_i)} = 1/\sqrt{\sum_{n\ or\ i}(\sum_{x=0}^{m}(xP_{nix} - \sum_{x=0}^{m} jP_{nix})^2)} \tag{2.38}$$

and for Rasch-Andrich thresholds F_x, where P_{nik} is the probability of observing category k for person n on item i.

$$Model\ SE_{(F_x)} = 1/\sqrt{\sum_{n=1}^{N}\sum_{i=1}^{L}(\sum_{k=0}^{x} P_{nik} \times \sum_{k=x+1}^{m} P_{nik})} \tag{2.39}$$

In addition, a misfit-inflated standard error (named as "Real SE" in WINSTEPS and Facets) can be estimated using equation:

$$Real\ SE = Model\ SE\ \times Maximum[\ 1.0, \sqrt{MNSQ_{Infit}}\ ] \tag{2.40}$$

In practice, the model standard error is usually reported because it is the lower bound of the measurement imprecision (Linacre, 2014c). By contrast, the real standard error shows the upper bound of the measurement imprecision. The actual standard error of measurement lays between them.

[7] Including both rating scale model and partial credit model

2.3.7 Chi-square tests for fixed effect and random effect in MFR models

Two additional statistics obtained from chi-square tests are reported by software Facets (Linacre, 2014a) for the analysis of MFR models.

2.3.7.1 Chi-square test for fixed (all-same) effect

It tests the hypothesis of whether the samples within the particular facet share the same location on the scale after *“accounting for the measurement error”* (*i.e.* the fixed effect*)* (Linacre, 2014b). The chi-square statistic is also named as homogeneity index in other literature (Eckes, 2011). The p-value is the probability of the null hypothesis of the fixed effect is valid with the samples. The formula, using the Panellist facet in a composite consumer sensory liking test as an example[8], can be written as

$$\chi^2 = \frac{\sum\left(w_j \times C_j^2\right) - \sum(w_j \times C_j)^2}{\sum w_j} \quad \text{with degree of freedom J-1} \qquad (2.41)$$

where

C_j is Panellist j’ overall liking level,

J = the number of panellists, and

$w_j = \frac{1}{SE_j^2}$ for j=1, J.

Note: for other facets, C can be replaced by B for products or D for attributes, while $\mathrm{J/j}$ can be replaced by $\mathrm{N/n}$ for product or $\mathrm{L/i}$ for attributes.

2.3.7.2 Chi-square test for random (normal) effect

It tests the hypothesis of whether the data set can be considered as *“randomly sampled from a normal distributed population*” (*i.e.* the random effect) (Linacre, 2014b). The p-value is the probability of the null hypothesis of the random effect is true with the data collected. The formula (using the Panellist facet in a consumer liking test as an example) is:

[8] The model can be seen in section 2.1.2.3.

$$\chi^2 = \frac{\sum\left(w_j \times C_j^2\right) - \left(\sum w_j \times C_j\right)^2}{\sum w_j} \quad \text{with degree of freedom J-2} \qquad (2.42)$$

where

C_j is Panellist j' overall liking level,

J = the number of panellists, and

$w_j = \frac{1}{Variance(C) + SE_j^2} = \frac{1}{\frac{\sum(C_j - C_{mean})^2}{J-1} - \frac{\sum SE_j^2}{J} + SE_j^2}$ for j=1, J.

Note: likewise, for other facets, C can be replaced by B for products or D for attributes, while $\mathrm{J/j}$ can be replaced by $\mathrm{N/n}$ for product or $\mathrm{L/i}$ for attributes.

2.3.8 Remarks

This chapter compared the conceptual differences between CTT and Rasch analysis on the model assumptions, estimation methods, and quality control requirements. It also outlined the procedures of applying CTT and Rasch analysis in research and the criteria to meet for constructing measures. These procedures and criteria would serve as the basis of the data analysis in the case studies.

Chapter 3 Case study I: Comparison between classical test theory approach and Rasch analysis in evaluation of consumer survey – a revisit of the Health and Taste Attitude Scales (HTAS)

3.1 Introduction

This study compared the use of classical test theory (CTT) approach and Rasch analysis in the evaluation of a consumer attitude survey made up of Likert items (Likert, 1932). In order to obtain results from real data rather than simulated data, an existing instrument – the Health and Taste Attitude Scales (HTAS) (Roininen *et al.*, 1999; Roininen *et al.*, 2000) was revisited.

The psychometric properties of the instrument such as dimensionality and reliability were evaluated using both CTT approach and Rasch analysis for comparison. After that, the differences between gender and age groups on each subscale were examined based on the estimates obtained by the two approaches. The effects of improving scale category effectiveness using collapsed scale categories and resolving DIF, which were rooted in the tradition of applying Rasch analysis, were also explored.

3.2 Instrument and sampling procedures

3.2.1 Instrument

This case study revisited the Health and Taste Attitudes Scales (HTAS) instrument, which was developed by Roininen, et al. (1999; 2000) for measuring two important determinants of food choice: health-related attitudes and taste-related attitudes. Tables 3.1 and 3.2 depict the subsidiary subscales and item statements of HTAS. The instrument was made up of thirty-eight items, including twenty items associated with health-related attitudes and eighteen items concerned with the taste-related attitudes. Both the health part and the taste part were further divided into three subscales in the original study based on the results of factor analysis. An equal number of positively-worded and negatively-worded items were composed for each subscale. After its development, it has been used in several research for segmenting people according to their degree of health-related and/or taste-related interests, which were linked to particular eating behaviours (Kowalkowska *et al.*, 2018; Roininen and Tuorila, 1999; Roininen *et al.*, 2001; Zandstra *et al.*, 2001).

In this study, the same item statements and item labels used in HTAS (tables 3.1 and 3.2) were used as the first part of the instrument. The items belonging to different subscales were remixed in presenting order. All participants were asked to rate the items using a 7-point Likert scale from “Strongly Disagree” to “Strongly “Agree”. After that, the respondents were required to provide information about their gender, age and if they had special dietary pattern (*e.g.* restriction to sugar due to diabetes).

3.2.2 Participants and sampling procedures

The study was approved by Faculty Research Ethics Committee (ref: MEEC 14-027). The participants were recruited via email and posters posted in campus of University of Leeds. All participants were instructed to complete an online survey via the Bristol online survey tool at any location they preferred. The participation was fully voluntary.

Table 3.1 Health-related subscales and items in original HTAS instrument

Subscale	Label	Item statement
General Health Interest (G)	G1	I am very particular about the healthiness of food
	G2	I always follow a healthy and balanced diet
	G3	It is important for me that my diet is low in fat
	G4	It is important for me that my daily diet contains a lot of vitamins and minerals
	G5R	I eat what I like and I do not worry about healthiness of food
	G6R	The healthiness of food has little impact on my food choices
	G7R	The healthiness of snacks makes no difference to me
	G8R	I do not avoid any foods, even if they may raise my cholesterol
	Subscale G refers to an interest in eating healthily	
Light product interest (L)	L1	I believe that eating light products keeps one's cholesterol level under control
	L2	I believe that eating light products keeps one's body in good shape
	L3	In my opinion by eating light products one can eat more without getting too many calories
	L4R	In my opinion, the use of light products does not improve one's health
	L5R	In my opinion light products don't help to drop cholesterol levels
	L6R	I do not think that light products are healthier than conventional products
	Subscale L deals with an interest in eating light product	
Natural product interest (N)	N1	I do not eat processed foods, because I do not know what they contain
	N2	I try to eat foods that do not contain additives
	N3	I would like to eat only organically grown vegetables
	N4R	In my opinion, artificially flavoured foods are not harmful for my health
	N5R	In my opinion, organically grown foods are not better for my health than those grown conventionally
	N6R	I do not care about additives in my daily diet
	Subscale N relates to an interest in eating foods that do not contain additives and are unprocessed	

Table 3.2 Taste-related subscales and items in original HTAS instrument

Subscale	Label	Item statement
Craving for sweet foods (C)	C1	I often have cravings for sweets
	C2	I often have cravings for chocolate
	C3	I often have cravings for ice-cream
	C4R	In my opinion it is strange that some people have cravings for sweets
	C5R	In my opinion it is strange that some people have cravings for chocolate
	C6R	In my opinion it is strange that some people have cravings for ice-cream
	Subscale C describes the strength of cravings for chocolate, sweets, and ice-cream	
Using food as a reward (U)	U1	I reward myself by buying something really tasty
	U2	I indulge myself by buying something really delicious
	U3	When I am feeling down I want to treat myself with something really delicious
	U4R	I avoid rewarding myself with food
	U5R	In my opinion, comforting oneself by eating is self-deception
	U6R	I try to avoid eating delicious food when I am feeling down
	Subscale U considers food as a reward	
Pleasure (P)	P1R	The appearance of food makes no difference to me
	P2	When I eat, I concentrate on enjoying the taste of food
	P3R	I do not believe that food should always be source of pleasure
	P4	It is important for me to eat delicious food on weekdays as well as weekends
	P5	An essential part of my weekend is eating delicious food
	P6R	I finish my meal even when I do not like the taste of a food
	Subscale P reflects the importance of obtaining pleasure from food	

3.2.3 Data collection and pre-treatment

The online survey was opened to the public for three months. 817 persons visited the survey webpage. However, 515 people left before completing the survey, including 396 persons who stopped browsing at the participant information and consent form section before having a chance to read and respond to any survey item. Only 302 respondents completed the survey (completion rate: 36.96%). It implies that there is a need to improve the recruitment of participants and the design of online survey in future study.

Data from two participants were removed because they were the only two people who selected option "*Prefer not to Answer*" on the item of gender. It is too small to be compared with the other gender groups (*i.e.* Female and Male). Twenty-eight Respondents who did not respond to all items were dropped, too. After reviewing the information about dietary patterns, three more respondents were removed because they presented extreme responses resulting from special dietary pattern caused by disease (*e.g.* respondent who has type II diabetes). The final sample size was 269.

The distribution of respondents by gender and age was reviewed. The sizes of some age groups were too small to be used for comparison (for example, there were only two respondents who belonged to the age group "65-74"). To solve this issue, all respondents aged 35 and above were combined to a new age group "35+" after taking both the group size and nature of the groups into consideration. The final distribution of respondents by gender and age is tabulated in table 3.3.

Table 3.3 Distribution of respondents by gender and age

	Number of respondents
Gender	
Female	194
Male	75
Age	
16-24	114
25-34	68
35+	87

3.3 Analysis by CTT approach

3.3.1 Methods

In this study, in order to verify whether the original structure of the Health and Taste scales and their subscales that was found by Roininen *et al.* (1999;2000) could be repeatedly found in CTT approach, an exploratory factor analysis (EFA) was used to examine the dimensionality of the HTAS using the "psych" package (Revelle, 2015). Two preliminary tests were conducted prior to analysing the data with EFA to examine the suitability of using all 38 items from the HTAS. Sampling adequacy was examined with the Kaiser-Meyer-Olkin test for (KMO test), which should exceed a value greater than 0.6. The Bartlett's test of sphericity should also show a statistical significance ($p<0.05$). A parallel analysis (Horn, 1965) was then applied to provide an initial prediction of the number of factors to be extracted from health-related items and taste-related items. EFA was performed on the data using the Principal Axis Factoring (PAF) procedure as the extraction method, with varimax rotation (Kaiser, 1958) to improve the interpretation of the factors [9]. The reliability of the instrument was evaluated by computing Cronbach's alpha for each subscale, where an alpha greater than 0.7 is considered as acceptable (Nunnally and Bernstein, 1994). After the structure of the instrument was verified, the respondents' summated scores of each subscale were calculated as their total score for each underlying factor.

3.3.2 Results

3.3.2.1 Structure of the instrument

(1) KMO test and Bartlett's test of sphericity

The results of KMO test for sampling adequacy and Bartlett's test of sphericity on health-related items and taste-related items are summarised in table 3.4. Figures 3.1 and 3.2 further present the individual measure of sample adequacy (MSA) for each item.

According to the results, the overall KMO measures greater than recommended 0.6 and the significant Bartlett's test for sphericity statistics indicated that the data sets were suitable for factor analysis. However, the individual MSAs of items P1R

[9] More detailed information about EFA can be found in section 2.3.2.2.

(0.54) and P6R (0.56) were slightly off criteria (figure 3.2), implying that they were not strongly correlated with the other items.

Table 3.4 Results of KMO test and Bartlett’s test of sphericity on raw scores of HTAS

Scale	**KMO**[1]	**Bartlett's test of sphericity**		
		χ^2	***d.f.***	**p-value**
Health-related items	0.81	1783	190	0.000
Taste-related items	0.77	1373	153	0.000
Criteria	>0.60			<0.05

[1] Kaiser–Meyer–Olkin overall measure of sampling adequacy

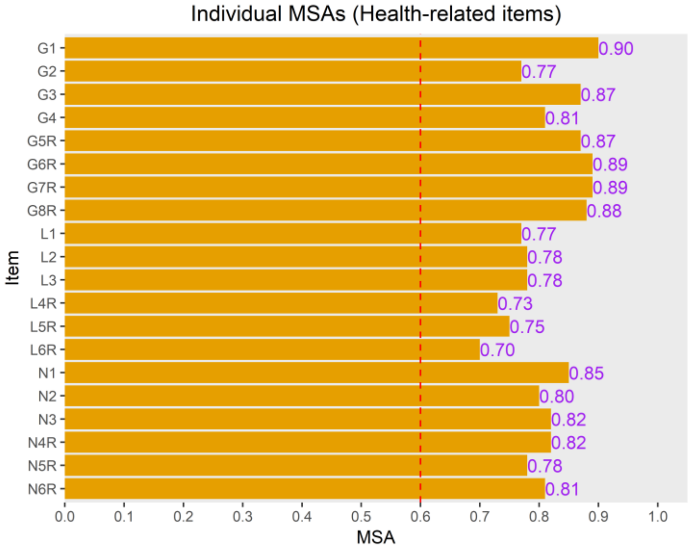


Figure 3.1 Individual MSA of health-related items

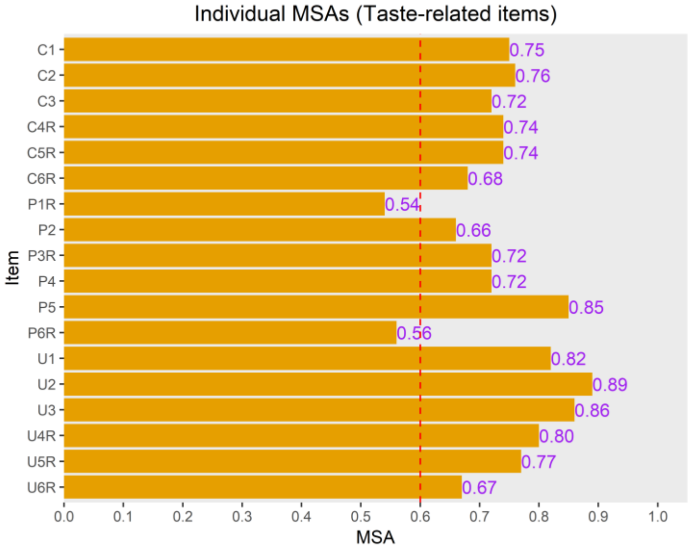


Figure 3.2 Individual MSA of taste-related items

(2) Parallel analysis

The parallel analysis with principal axis factoring implied that the items could be grouped into 4 and 5 factors for health-related part and taste-related attitudes, respectively. The scree plots obtained from parallel analysis can be seen in Appendix C.

(3) Factor analysis

Table 3.5 tabulates the factor loadings of health-related items with the 4-factor solution. Items G2 and G3 were both cross-loaded on more than one factor when using 0.30 as the cut-off point. They were allocated to factor 2 based on the comparison of the loadings on those factors. As a result, the factor 4 was removed because there was no other item that had a loading greater than 0.30 on it. The data were re-analysed with a 3-factor solution using the same factor analysis procedure. The extracted loadings (table 3.6) revealed an underlying structure that was consistent with the original HTAS study (Roininen *et al.*, 1999). Therefore the factors (*i.e.* subscales) were named using the original names. The same labels “G”, “L” and “N” were assigned to them accordingly.

The factor loadings of the taste-related items with the 5-factor solution are depicted in table 3.7. Items P1R and P6R did not have a loading greater than 0.30 on any factor. This was not surprising because both items had unsatisfying individual MSA values. Item U6R was the only item loaded on factor 5 with a loading higher than 0.3. For a practical perspective, single item subscale should be avoided (Raubenheimer, 2004). So the factor analysis was repeated with a 4-factor solution. The factor loadings with the 4-factor solution are listed in table 3.8. The items U6R, P1R and P6R were eventually removed because they did not have loadings above 0.30 on any of the factors. The factors 1-4 were labelled as subscales U, CA, CB and P based on their connections with the original subscales. The items in subscales CA and CB were both from the original subscale C (Craving for sweet foods). The items in subscale CA (item C1~C3) referred to respondents’ attitudes towards their own craving for sweet foods, while the items in subscale CB (C4R~C6R) concerned respondents’ attitudes towards other people’s craving for sweet foods.

Table 3.5 Factor loadings of health-related items with the 4-factor solution

Items	**Factors**			
	1	**2**	**3**	**4**
G1	-0.008	**0.565**	0.227	-0.025
G2	-0.073	**0.610**	0.034	**~~0.342~~**
G3	**~~0.314~~**	**0.385**	0.228	**~~0.304~~**
G4	-0.006	**0.650**	0.061	0.154
G5R	0.097	**0.583**	0.152	-0.051
G6R	0.091	**0.612**	0.207	-0.205
G7R	0.149	**0.616**	0.219	-0.132
G8R	0.178	**0.351**	0.270	0.107
L1	**0.699**	0.119	-0.009	0.114
L2	**0.658**	-0.031	0.074	0.272
L3	**0.474**	-0.068	0.014	0.261
L4R	**0.813**	0.123	-0.078	-0.244
L5R	**0.616**	0.173	-0.020	-0.110
L6R	**0.638**	0.005	-0.176	-0.162
N1	-0.096	0.196	**0.430**	0.155
N2	0.020	0.181	**0.728**	0.138
N3	-0.096	0.157	**0.598**	-0.186
N4R	-0.067	0.030	**0.541**	0.010
N5R	0.098	0.129	**0.577**	-0.083
N6R	-0.060	**~~0.321~~**	**0.648**	0.114

The factor loadings greater than 0.30 are in **bold.**

Table 3.6 Factor loadings of health-related items with the 3-factor solution

Items	**Factors**		
	1	**2**	**3**
G1	**0.572**	-0.013	0.216
G2	**0.587**	-0.060	0.076
G3	**0.395**	**~~0.312~~**	0.240
G4	**0.658**	-0.009	0.068
G5R	**0.589**	0.092	0.138
G6R	**0.594**	0.086	0.185
G7R	**0.611**	0.144	0.200
G8R	**0.357**	0.181	0.276
L1	0.125	**0.708**	-0.005
L2	-0.007	**0.644**	0.079
L3	-0.050	**0.469**	0.026
L4R	0.117	**0.786**	-0.102
L5R	0.167	**0.619**	-0.031
L6R	-0.002	**0.634**	-0.190
N1	0.201	-0.090	**0.443**
N2	0.187	0.028	**0.739**
N3	0.159	-0.096	**0.559**
N4R	0.028	-0.061	**0.547**
N5R	0.132	0.100	**0.559**
N6R	**~~0.325~~**	-0.055	**0.659**

The factor loadings greater than 0.30 are in **bold.**

Table 3.7 Factor loadings of taste-related items under the 5-factor solution

Items	**Factors**				
	1	**2**	**3**	**4**	**5**
C1	0.261	**0.477**	-0.103	-0.191	-0.355
C2	0.263	**0.552**	0.027	-0.044	-0.398
C3	0.273	**0.438**	0.214	-0.197	-0.326
C4R	-0.006	**0.818**	-0.057	-0.011	0.236
C5R	0.075	**0.781**	0.015	0.183	0.138
C6R	0.059	**0.614**	0.028	-0.028	0.135
U1	**0.787**	0.047	0.122	0.107	-0.143
U2	**0.682**	0.126	0.169	0.090	-0.020
U3	**0.610**	0.171	0.119	0.012	0.014
U4R	**0.723**	0.072	0.094	0.286	0.201
U5R	~~**0.309**~~	-0.019	-0.002	**0.498**	-0.074
U6R	0.180	0.022	0.153	0.087	**0.332**
P1R	0.005	0.111	-0.034	0.001	0.290
P2	0.108	0.023	**0.730**	0.008	0.019
P3R	0.088	0.025	~~**0.338**~~	**0.587**	0.064
P4	0.139	-0.033	**0.474**	0.108	0.047
P5	**0.455**	0.037	**0.449**	0.088	0.005
P6R	-0.049	0.024	0.042	-0.188	0.279

The factor loadings greater than 0.30 are in **bold**

Table 3.8 Factor loadings of taste-related items under the 4-factor solution

Items	**Factors**			
	1	**2**	**3**	**4**
C1	0.100	**0.669**	0.073	-0.144
C2	0.152	**0.675**	0.141	-0.009
C3	0.093	**0.649**	0.049	0.183
C4R	-0.016	~~**0.374**~~	**0.771**	-0.052
C5R	0.123	~~**0.359**~~	**0.700**	0.031
C6R	0.025	~~**0.337**~~	**0.535**	0.034
U1	**0.732**	~~**0.306**~~	-0.140	0.137
U2	**0.636**	0.260	0.003	0.188
U3	**0.537**	0.275	0.050	0.136
U4R	**0.791**	0.021	0.136	0.143
U5R	**0.441**	-0.055	0.012	0.051
U6R	0.225	-0.155	0.195	0.182
P1R	0.034	-0.099	0.246	-0.018
P2	0.055	0.077	-0.012	**0.752**
P3R	0.293	-0.165	0.137	**0.331**
P4	0.156	-0.023	-0.010	**0.478**
P5	~~**0.423**~~	0.143	-0.029	**0.463**
P6R	-0.091	-0.083	0.129	0.040

The factor loadings greater than 0.30 are in **bold**

3.3.2.2 Reliability

The Cronbach's alpha coefficient of each subscale and its 95% CI were computed (see table 3.9). The results showed that most of the subscales have acceptable reliability except the subscale P (pleasure), which had a questionable alpha value of 0.61, less than the recommended value 0.70 (Nunnally and Bernstein, 1994). This finding is consistent with the original HTAS study (Roininen *et al.*, 1999), where the subscale P (pleasure) had the smallest alpha coefficient (0.67) among the subscales. This implies that the composition of subscale P was not ideal.

Table 3.9 The Cronbach's alpha of each subscale (and its 95%CI)

Subscales	**Cronbach's alpha**	**95%CI Lower bound**	**95%CI Upper bound**
G	0.80	0.76	0.83
L	0.81	0.77	0.84
N	0.78	0.73	0.82
CA	0.72	0.67	0.78
CB	0.80	0.76	0.84
U	0.79	0.75	0.83
P	0.61	0.53	0.68

3.3.2.3 Respondents' total score on each subscale

After verifying the structure of the instrument, the respondents' scores of the items in each scale were summed as their total scores for each subscale. Table 3.10 provides an overview of the total scores.

Table 3.10 Summary of respondents' total score on each subscale

Subscale	**Range of possible total score**	**Mean**	**Median**	**Min**	**Max**	**SE**[1]
G (8 items)	8~56	37.87	39	11	55	0.482
N (6 items)	6~42	26.07	27	7	41	0.436
L (6 items)	6~42	22.52	23	6	39	0.437
CA (3 items)	3~21	12.52	13	3	21	0.276
CB (3 items)	3~21	16.97	18	6	21	0.204
U (5 items)	5~35	22.12	23	5	34	0.368
P (4 items)	4~28	19.41	20	7	28	0.264

[1] Standard error of mean

3.4 Rasch analysis

3.4.1 Methods

The data were fitted to the unidimensional Rasch Rating Scale (RS) model (Andrich, 1978a) using WINSTEPS (Linacre, 2014d). After that, a series of tests were conducted in the following order:

(1) Evaluating rating scale category effectiveness

Rasch analysis provides an additional quality control procedure in order to ensure the rating scale categories are functioning effectively. In this study, the category effectiveness of health-related and taste-related scales were evaluated using the method and criteria described in section 2.3.1.2. If the criteria were not satisfied, attempts at collapsing scale categories would be made. Further analysis would be conducted with both original and collapsed rating scales for comparison purposes.

(2) Verifying the assumption of unidimensionality and local item independence

A principal component analysis on standardised model residuals (PCAR) followed by independent t-test protocol were performed to examine the assumption of unidimensionality for the health-related and taste-related scales separately, following the procedure described in section 2.3.2.3.

In addition to the tests of unidimensionality, the assumption of local item independence was also inspected by computing the residual correlation between the items. Item pairs that had a residual correlation coefficient equal to or greater than 0.30 were reviewed as they may violate the assumption of local item independence (Smith, 2000).

After the underlying structure of the instrument was identified, each subscale that represented a sub-dimension was refitted to the model for further analysis.

(3) Assessing differential item functioning (DIF)

The examination of the differences between gender and age groups on HTAS subscales was one of the target of this study. Therefore the influence of DIF had to be minimised before the comparison. DIF by gender and age groups was evaluated according to the statistics obtained from Mantel-Haenszel test (M-H test) and sample-based effect size of DIF contrast between groups, following the procedure illustrated in section 2.3.4.2.

(4) **Assessing fit**

The outfit MNSQ statistics were evaluated for item fit and person fit using the criteria outlined in section 2.3.3.1.

(5) Creating Wright maps

The estimated person and item measures on each subscale were visualised using the Wright maps. An introduction of Wright map had been provided in section 1.3.1.

(6) Inspecting Rasch reliability statistics.

The Rasch reliability statistics were obtained from WINSTEPS.

3.4.2 Results

3.4.2.1 Category effectiveness

The statistics for the rating scale categories are tabulated in table 3.11. The results showed that all essential criteria suggested by Linacre (2002a) were satisfied. For example, the outfit MNSQ of each category were all around 1.0, suggesting good fit. The average measure advanced monotonically, implying that the higher category can represent higher degree of agreeability.

However, disordered Rasch-Andrich thresholds were observed in both health-related and taste-related scales, indicating that the respondents may not have used the categories in a consistent manner.

Attempts at collapsing categories had been made. The best solution was described in table 3.12, where the 7-point rating scale was collapsed to a 4-point rating scale. The revised data were fitted to the Rating scale Rasch model. Further analysis was conducted on both original model and revised model.

Table 3.11 Statistics for the original 7-point rating scale

		Health-related items					**Taste-related items**				
Scale category	**Raw Score**	**Counts[1]**	**Obs[2]**	**Exp[3]**	**Rasch-Andrich Threshold ± SE**	**Outfit MNSQ**	**Counts[1]**	**Obs[2]**	**Exp[3]**	**Rasch-Andrich Threshold ± SE**	**Outfit MNSQ**
Strongly disagree	1	322 (5.99%)	-0.39	-0.39	NA	1.06	189 (3.90%)	-0.23	-0.23	NA	1.31
Disagree	2	752(13.98%)	-0.20	-0.23	-1.16±0.06	1.06	585 (12.08%)	-0.07	-0.09	-1.29±0.08	1.11
Slightly disagree	3	826(15.35%)	-0.07	-0.09	**-0.25±0.04**	1.01	507 (10.47%)	0.09	0.05	**0.12±0.04**	1.11
Neither disagree nor agree	4	547(10.17%)	-0.02	0.05	**0.39±0.03**	0.89	330 (6.82%)	0.17	0.2	**0.55±0.04**	1.10
Slightly agree	5	1286(23.90%)	0.17	0.20	**-0.73±0.03**	1.04	1004 (20.74%)	0.30	0.35	**-0.84±0.04**	0.8
Agree	6	1196(22.23%)	0.35	0.35	0.34±0.03	1.04	1562 (32.26%)	0.52	0.51	-0.01±0.03	0.95
Strongly agree	7	451(8.38%)	0.57	0.51	1.41±0.05	0.97	665 (13.73%)	0.74	0.71	1.46±0.04	1.02

[1] The counts of each category. The percentages of the counts are displayed in brackets.

[2] Modelled average measure in logits.

[3] Expected average measure if data fitted the model.

Disordered Rasch-Andrich thresholds are in **bold**.

Table 3.12 Statistics for the collapsed 4-point rating scale

		Health-related items					**Taste-related items**				
Scale category	**Raw Score**	**Counts**[1]	**Obs**[2]	**Exp**[3]	**Rasch-Andrich Threshold ± SE**	**Outfit MNSQ**	**Counts**[1]	**Obs**[2]	**Exp**[3]	**Rasch-Andrich Threshold ± SE**	**Outfit MNSQ**
Strongly disagree	1	322 (5.99%)	-0.91	-1.05	NA	1.11	189 (3.90%)	-0.59	-0.77	NA	1.20
Somewhat disagree (Disagree+ Slightly disagree+ Neither disagree nor agree)	2	2125 (39.50%)	-0.30	-0.26	-2.53±0.06	0.92	1422 (29.37%)	0.00	0.05	-2.36±0.08	0.95
Somewhat disagree (Slightly agree+ Agree)	3	2482 (46.13%)	0.50	0.49	-0.04±0.03	0.98	2566 (52.99%)	0.81	0.80	-0.16±0.03	0.89
Strongly agree	4	451 (8.38%)	1.27	1.26	2.57±0.05	1.01	665 (17.73%)	1.56	1.56	2.52±0.05	1.02

[1] The counts of each category. The percentages of the counts are displayed in brackets.

[2] Modelled average measure in logits.

[3] Expected average measure if data fitted the model.

3.4.2.2 Tests of unidimensionality and local item independence

The results (figures 3.3 and 3.4) suggested that three sub-dimensions could be identified from the 20 health-related items for both models (*i.e.* initial model with original rating scale and revised model with collapsed rating scale). The items belonging to each subscale were exactly the same with the original research (Roininen *et al.*, 1999), as well as the results obtained from EFA evaluated on the raw scores of current data.

For the twenty taste-related items, however, a few differences were found:

(1) Two sub-dimensions were found from the taste-related items, less than three and four found in original research (Roininen *et al.*, 1999) and this research using EFA approach.

(2) The original subscales U and P were considered as a unidimensional subscale UP in Rasch analysis.

(3) All items under the original subscale C (craving for sweet foods) were extracted as a unidimensional subscale, which was consistent with the original research (Roininen *et al.*, 1999), but in contrast with the result from EFA in this study where these items were further differentiated into two individual subscales CA and CB.

(4) Unlike the CTT approach where a decision of dropping three items was made[1], none of the items needed to be removed at this stage in Rasch analysis.

(5) The independent t-test protocol failed to support the assumption of unidimensionality for subscale UP with collapsed rating scale, as the lower bound of 95% binomial CI was slighter greater than 5% (6.82%). However it was still considered as a unidimensional subscale in this study because the eigenvalue of the first PCAR contrast was only 2.32. Another reason is its unidimensionality had been supported by t-test protocol on the original model.

In addition, the residual correlations between the item pairs were all less than 0.30 in each subscale, which indicated that the assumption of local item independence could hold for all subscales.

[1] More methods that can be used for resolving DIF can be found in section 2.3.4.3

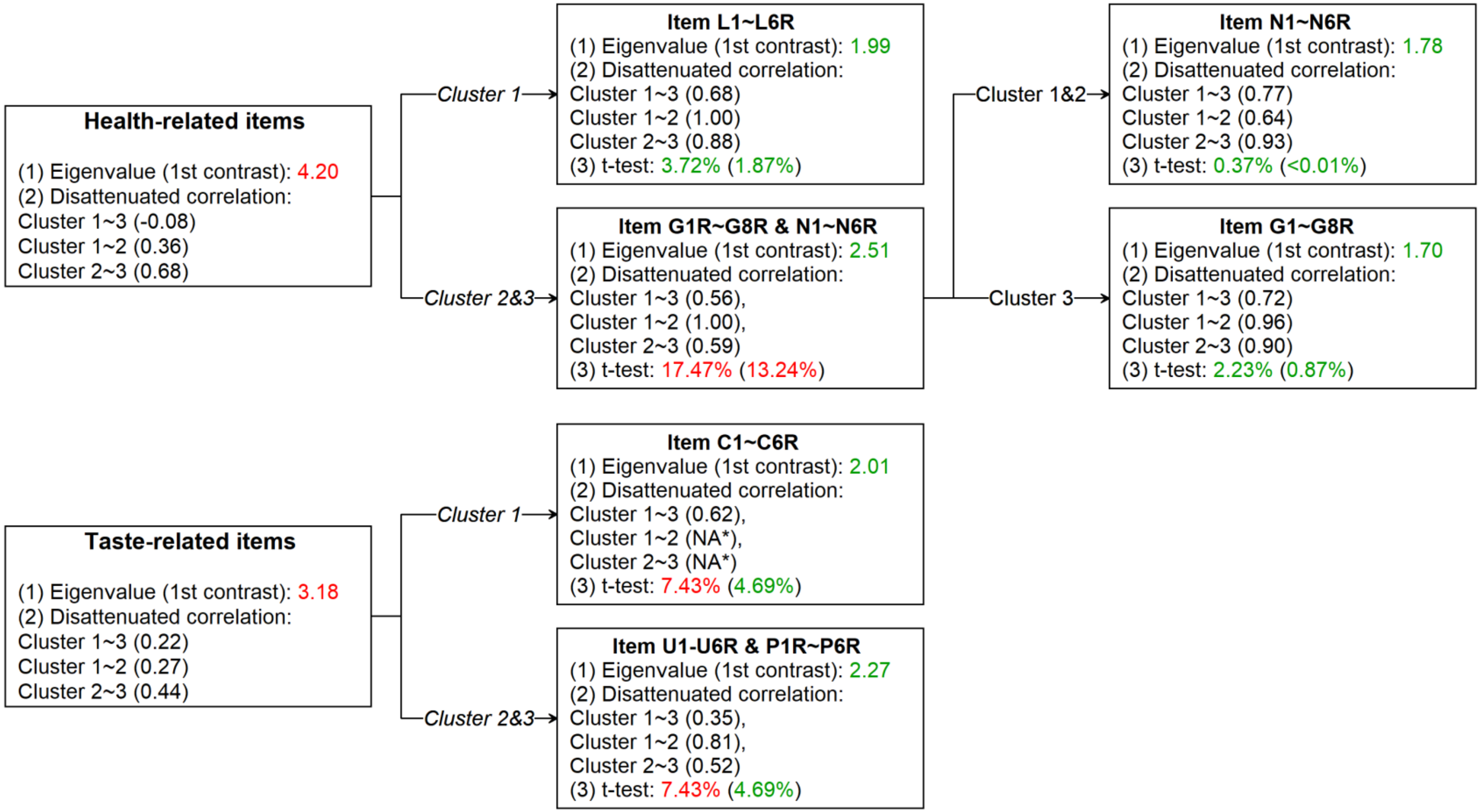
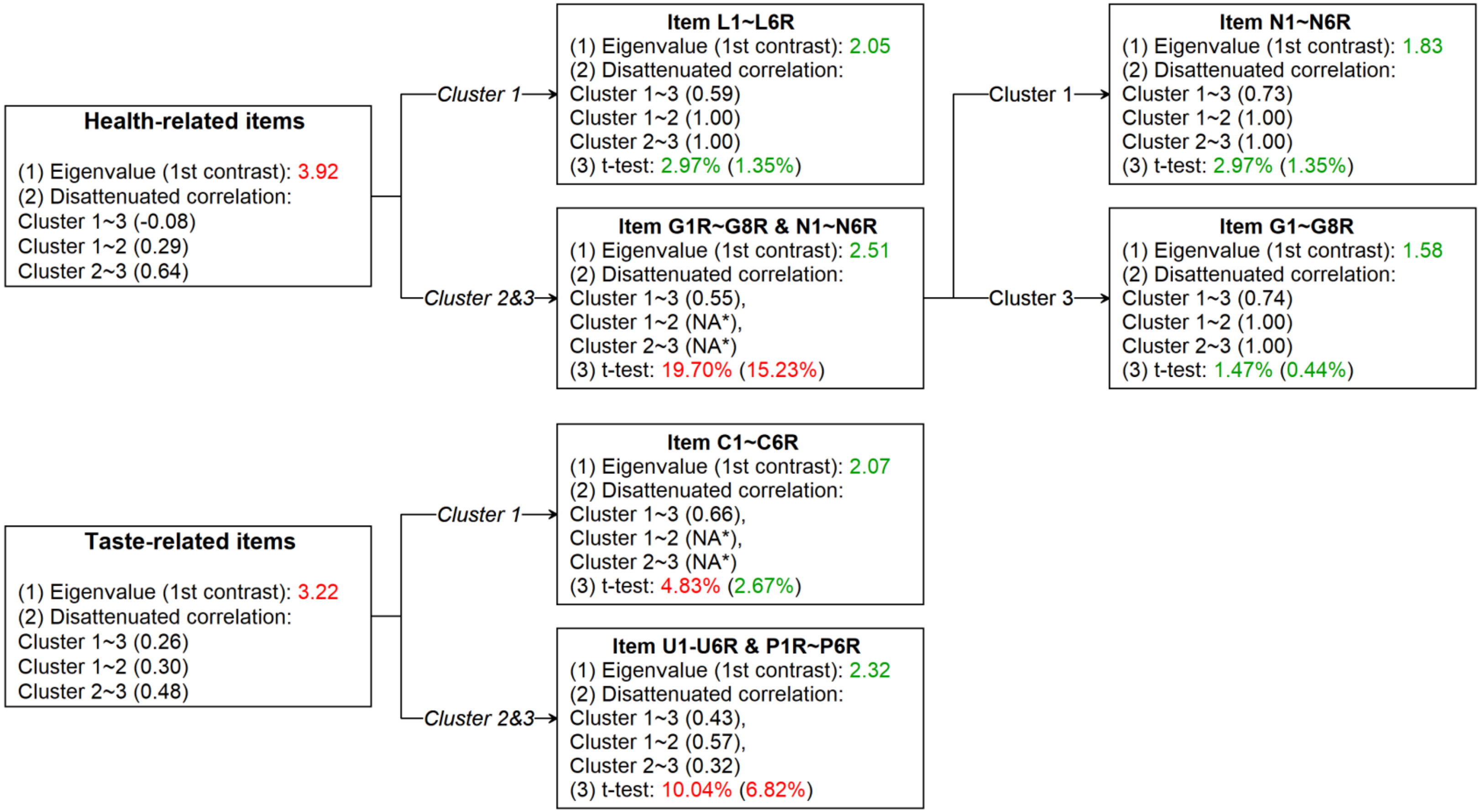

Health-related items
(1) Eigenvalue (1st contrast): 4.20
(2) Disattenuated correlation:
Cluster 1~3 (-0.08)
Cluster 1~2 (0.36)
Cluster 2~3 (0.68)
Cluster 1
Cluster 2&3
Item L1~L6R
(1) Eigenvalue (1st contrast): 1.99
(2) Disattenuated correlation:
Cluster 1~3 (0.68)
Cluster 1~2 (1.00)
Cluster 2~3 (0.88)
(3) t-test: 3.72% (1.87%)
Item G1R~G8R & N1~N6R
(1) Eigenvalue (1st contrast): 2.51
(2) Disattenuated correlation:
Cluster 1~3 (0.56),
Cluster 1~2 (1.00),
Cluster 2~3 (0.59)
(3) t-test: 17.47% (13.24%)
Cluster 1&2
Cluster 3
Item N1~N6R
(1) Eigenvalue (1st contrast): 1.78
(2) Disattenuated correlation:
Cluster 1~3 (0.77)
Cluster 1~2 (0.64)
Cluster 2~3 (0.93)
(3) t-test: 0.37% (<0.01%)
Item G1~G8R
(1) Eigenvalue (1st contrast): 1.70
(2) Disattenuated correlation:
Cluster 1~3 (0.72)
Cluster 1~2 (0.96)
Cluster 2~3 (0.90)
(3) t-test: 2.23% (0.87%)
Taste-related items
(1) Eigenvalue (1st contrast): 3.18
(2) Disattenuated correlation:
Cluster 1~3 (0.22)
Cluster 1~2 (0.27)
Cluster 2~3 (0.44)
Cluster 1
Cluster 2&3
Item C1~C6R
(1) Eigenvalue (1st contrast): 2.01
(2) Disattenuated correlation:
Cluster 1~3 (0.62),
Cluster 1~2 (NA*),
Cluster 2~3 (NA*)
(3) t-test: 7.43% (4.69%)
Item U1-U6R & P1R~P6R
(1) Eigenvalue (1st contrast): 2.27
(2) Disattenuated correlation:
Cluster 1~3 (0.35),
Cluster 1~2 (0.81),
Cluster 2~3 (0.52)
(3) t-test: 7.43% (4.69%)

Figure 3.3 Tests of unidimensionality of HTAS (original 7-point rating scale)

Values in brackets for t-test are the estimated lower bound of 95% binomial CI.

The satisfied statistics are marked in green, while the unsatisfied statistics are marked in red.

* Only two item clusters were identified.



Health-related items

(1) Eigenvalue (1st contrast): 3.92
(2) Disattenuated correlation:
Cluster 1~3 (-0.08)
Cluster 1~2 (0.29)
Cluster 2~3 (0.64)

Cluster 1

Item L1~L6R
(1) Eigenvalue (1st contrast): 2.05
(2) Disattenuated correlation:
Cluster 1~3 (0.59)
Cluster 1~2 (1.00)
Cluster 2~3 (1.00)
(3) t-test: 2.97% (1.35%)

Cluster 2&3

Item G1R~G8R & N1~N6R
(1) Eigenvalue (1st contrast): 2.51
(2) Disattenuated correlation:
Cluster 1~3 (0.55),
Cluster 1~2 (NA*),
Cluster 2~3 (NA*)
(3) t-test: 19.70% (15.23%)

Cluster 1

Item N1~N6R
(1) Eigenvalue (1st contrast): 1.83
(2) Disattenuated correlation:
Cluster 1~3 (0.73)
Cluster 1~2 (1.00)
Cluster 2~3 (1.00)
(3) t-test: 2.97% (1.35%)

Cluster 3

Item G1~G8R
(1) Eigenvalue (1st contrast): 1.58
(2) Disattenuated correlation:
Cluster 1~3 (0.74)
Cluster 1~2 (1.00)
Cluster 2~3 (1.00)
(3) t-test: 1.47% (0.44%)

Taste-related items

(1) Eigenvalue (1st contrast): 3.22
(2) Disattenuated correlation:
Cluster 1~3 (0.26)
Cluster 1~2 (0.30)
Cluster 2~3 (0.48)

Cluster 1

Item C1~C6R
(1) Eigenvalue (1st contrast): 2.07
(2) Disattenuated correlation:
Cluster 1~3 (0.66),
Cluster 1~2 (NA*),
Cluster 2~3 (NA*)
(3) t-test: 4.83% (2.67%)

Cluster 2&3

Item U1-U6R & P1R~P6R
(1) Eigenvalue (1st contrast): 2.32
(2) Disattenuated correlation:
Cluster 1~3 (0.43),
Cluster 1~2 (0.57),
Cluster 2~3 (0.32)
(3) t-test: 10.04% (6.82%)

Figure 3.4 Tests of unidimensionality of HTAS (collapsed 4-point rating scale)

Values in brackets for t-test are the estimated lower bound of 95% binomial CI.

The satisfied statistics are marked in green, while the unsatisfied statistics are marked in red.

* Only two item clusters were identified.

3.4.2.3 Test of individual fit

The item fit was evaluated by inspecting the outfit MNSQ, which can be seen in table 3.13 for health-related subscales and table 3.14 for taste-related subscales. The outfit MNSQ of most items were located in the suggested range between 0.5~1.5. The exceptional items were P6R and P1R within the frame of reference of both original and collapsed rating scales. After removing the extreme unexpected responses (table 3.14) that had absolute standardised residuals greater than 2.0, the outfit MNSQ statistics of these two items were adjusted to acceptable levels.

Some differences of these statistics were found between the original and the collapsed rating scales. The range of the measures and model SE increased after the rating scale was collapsed. The outfit MNSQ statistics estimated with the scale-collapsed data, on the other hand, were closer to the expected value 1.0 than those modelled with the original rating scale.

Table 3.13 Measure, model SE and outfit MNSQ statistics of items in health-related subscales

Subscales and Items	**Original rating scale**		**Collapsed rating scale**	
	Measure ±SE	**Outfit MNSQ**	**Measure ±SE**	**Outfit MNSQ**
General Health interest (G)				
G8R	0.59±0.05	1.21	1.31±0.12	1.15
G2	0.42±0.05	1.03	0.85±0.12	0.86
G3	0.25±0.05	1.09	0.42±0.12	1.10
G1	0.07±0.05	0.90	0.14±0.12	0.94
G7R	-0.19±0.05	0.95	-0.36±0.12	0.92
G5R	-0.31±0.06	1.11	-0.68±0.12	1.00
G6R	-0.34±0.06	1.26	-0.73±0.12	1.06
G4	-0.49±0.06	0.89	-0.95±0.12	0.90
Natural product interest (N)				
N1	0.50±0.05	1.20	1.10±0.11	1.12
N3	0.20±0.05	1.20	0.37±0.11	1.15
N5R	-0.11±0.05	1.18	-0.16±0.11	1.22
N2	-0.14±0.05	0.64	-0.32±0.11	0.63
N4R	-0.20±0.05	1.10	-0.38±0.11	0.99
N6R	-0.26±0.05	0.86	-0.61±0.11	0.82
Light product interest (L)				
L3	0.39±0.05	1.46	0.72±0.12	1.25
L2	0.14±0.05	1.03	0.13±0.12	1.05
L1	-0.06±0.05	0.70	0.05±0.12	0.67
L6R	-0.06±0.05	1.22	-0.20±0.12	1.19
L4R	-0.09±0.05	0.80	-0.28±0.12	0.87
L5R	-0.32±0.05	1.01	-0.41±0.12	0.89

Table 3.14 Measure, model SE and outfit MNSQ statistics of items in taste-related subscales

Subscales and Items	Original rating scale		Collapsed rating scale	
	Measure ±SE	Outfit MNSQ	Measure ±SE	Outfit MNSQ
Cravings for sweet foods (C)				
C3	1.05±0.05	1.23	1.94±0.11	1.13
C2	0.34±0.05	1.02	0.46±0.12	1.19
C1	0.32±0.05	1.10	0.43±0.12	1.07
C6R	-0.45±0.06	1.14	-0.71±0.12	0.98
C5R	-0.62±0.07	0.87	-1.05±0.13	0.79
C4R	-0.64±0.07	0.94	-1.07±0.13	0.78
Using food as rewards and pleasure (UP)				
U5R	0.71±0.04	1.31	1.75±0.11	1.01
U4R	0.27±0.04	0.73	0.60±0.11	0.92
P6R[1]	0.09±0.05	1.29	0.37±0.12	1.06
P3R	0.21±0.04	1.37	0.35±0.11	1.44
P5	0.20±0.04	0.92	0.28±0.11	0.90
U1	0.14±0.04	0.90	0.13±0.11	0.86
U2	0.12±0.04	0.80	0.13±0.11	0.84
P4	-0.13±0.05	1.16	-0.39±0.11	1.13
U6R	-0.26±0.05	1.20	-0.46±0.11	1.12
U3	-0.16±0.05	1.17	-0.64±0.11	1.14
P2	-0.38±0.05	0.89	-0.88±0.11	0.75
P1R[2]	-0.81±0.07	0.88	-1.24±0.12	0.79
Using food as rewards and pleasure (without DIF items)[3] (UPDIF)				
U5R	-	-	1.79±0.11	0.96
U4R	-	-	0.58±0.11	0.91
P3R	-	-	0.31±0.11	1.45
P5	-	-	0.23±0.11	0.94
U1	-	-	0.08±0.11	0.93
U2	-	-	0.08±0.11	0.86
U6R	-	-	-0.56±0.12	1.23
P2	-	-	-1.01±0.12	0.84
P1R[4]	-	-	-1.50±0.13	0.83

[1] After removing thirty-two extreme unexpected responses with both the original and collapsed scale.

[2] After removing twenty-three and twenty-nine extreme unexpected responses with the original and collapsed scale, respectively.

[3] After correcting for differential item functioning (section 3.4.2.4).

[4] After removing thirty-two extreme unexpected responses.

For person fit, table 3.15 tabulates the proportion of misfitting respondents for all models according to their outfit MNSQ statistics. A large proportion of respondents exhibited overfit, suggesting they were too predictable. Overfitting is not desired, but it would not degrade the measurement. For underfitting, the outfit MNSQ statistics of around 10% of respondents were great than 2.0. However, as discussed in 2.3.3.3, a small amount of misfitting respondents are inevitable in consumer research, but their impact to the measurement would be less than misfitting items. Therefore they were all retained in the data set.

Table 3.15 Counts and proportion of misfitting respondents

	Original rating scale			**Collapsed rating scale**		
	Overfit	Underfit		Overfit	Underfit	
	<0.5	1.5~2.0	>2.0	<0.5	1.5~2.0	>2.0
G	24.91%	22.30%	11.15%	29.37%	20.27%	8.92%
N	24.54%	19.70%	11.90%	26.39%	17.10%	8.92%
L	28.62%	21.19%	11.52%	21.19%	17.84%	8.92%
C	36.06%	18.96%	12.64%	43.87%	19.70%	12.64%
UP*	18.96%	17.84%	8.92%	17.10%	14.87%	8.92%
UPDIF*				25.65%	17.47%	10.04%

*After removing the extreme unexpected responses associated with misfit items

3.4.2.4 Tests for differential item functioning (DIF)

Table 3.16 highlights the potential DIF items which was significant in M-H test and also had a sample-based effect size greater than 0.5. No DIF item was found in subscale G, L, N and C. 6 items belonging to subscale UP were identified as significant DIF items with the original model. No action was done to them for comparison purpose. With the revised model using collapsed rating scale, the considerable DIF were found only in 3 items:

(1) U3 between both gender and age groups, and
(2) P4 and P6R between age groups

A modified subscale UPDIF was created, in which these three items were removed from subscale UP. The data of UPDIF was fitted to model. The main statistics of the items are tabulated in table 3.14. No misfitting items were found in this subscale.

Table 3.16 DIF detection by gender group

Group	Item	Model[1]	Absolute DIF Contrast (logit)[2]	M-H test[3]	Effect size
Gender					
Female-Male	U3	Original	0.35	**0.002**	**0.56**
Female-Male	U3	Collapsed	0.78	**0.024**	**0.64**
Age[4]					
1-2	U3	Original	0.34	**0.023**	**0.55**
1-3	U4R	Original	0.31	**0.034**	**0.50**
1-3	U5R	Original	0.36	**0.002**	**0.58**
1-3	P1R	Original	0.36	**0.008**	**0.58**
2-3	P1R	Original	0.37	**0.032**	**0.60**
1-3	P4	Original	0.43	**0.010**	**0.69**
1-3	P6R	Original	0.44	**0.010**	**0.71**
1-2	U3	Collapsed	0.70	**0.003**	**0.58**
1-3	P4	Collapsed	1.16	**<0.001**	**0.97**
1-3	P6R	Collapsed	0.87	**0.006**	**0.72**

[1] Original=scaled with original 7-point rating scale, Collapsed=scaled collapsed 4-point rating scale

[2] The absolute value of the difference on the estimates of the measures of individual items between subgroups. [3] Rasch-based Mantel-Haenszel test

[4] 1=16~24, 2=25~34, 3=35+

3.4.2.5 Wright maps

The locations of respondents and items on each subscale were presented using the Wright maps, for initial inspection (figures 3.5~3.8).

It should be noted that the person measures and item measures are interpreted in different way in two-facet Rasch analysis. For example, in this study, the person measures represented respondents' agreeability on the scale. The greater the measure, the more likely it is that the respondent would agree with the items. The respondents who had the highest agreeability on the scale were plotted at the top of the Wright map. On the contrary, the item measures were the indicators of how difficult an item can be agreed with. The direction of item measures was opposite to that of person measures. The higher the measure, the less endorsability an item would show in the measurement. The items were placed on the Wright map from most agreed items at the bottom to the least agreed items at the top. For example, in subscale G, it was hardest to agree with item G8R and easiest to agree with item G4.

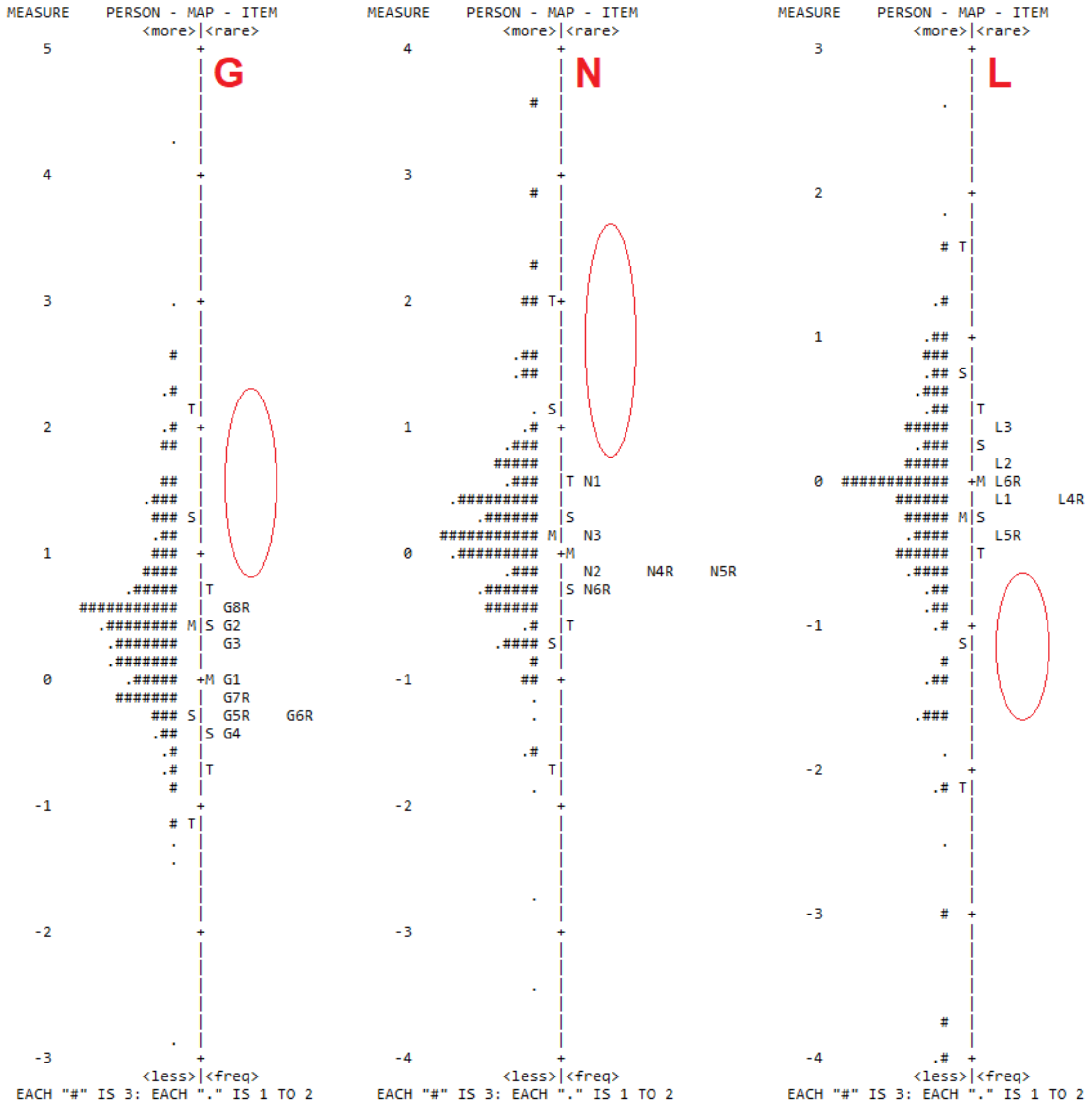
Some information can be extracted from the Wright maps.

(1) To differentiate people, an ideal instrument should be an assembly of items that covers all people at different construct levels. However, the ceiling effect and the flooring effect were found as the items in some subscales spread in relatively narrow ranges on the scale than persons, which can be observed on the Wright maps. For example, in figure 3.5, some gaps between or beyond current items were identified. The respondents located in these gaps cannot be well differentiated.

(2) The distributions of both persons and items were similar in both models using original 7-point rating scale and 4-point collapsed rating scale.

(3) The respondents are approximately normal distributed. Therefore, the separation statistics are more suitable for describing the dispersion of respondents than strata[1], which are both Rasch reliability statistics.

[1] The explanation of the difference between separation and strata can be seen in section 2.3.5.2.



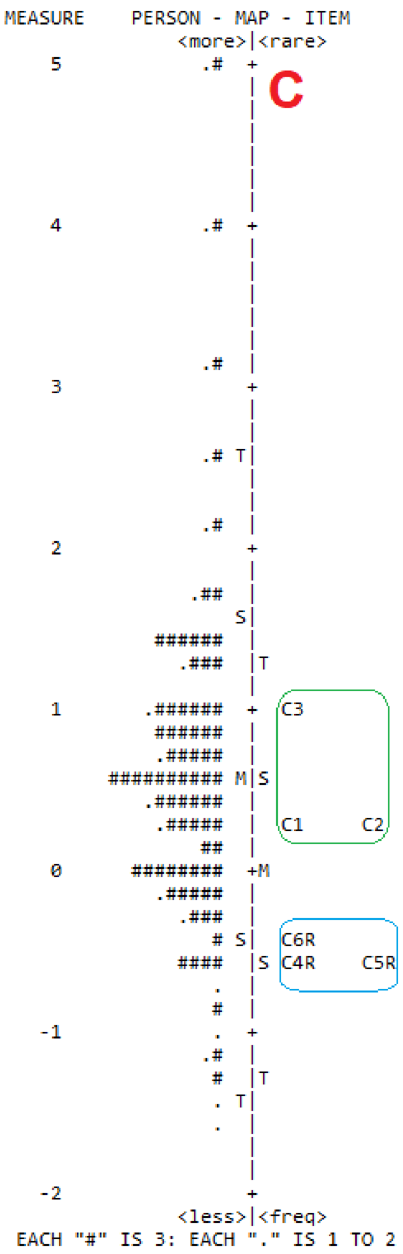
```
MEASURE    PERSON - MAP - ITEM
               <more>|<rare>
    5                +
                     |
                     |
                     |
                     |
                  .  |
                     |
    4                +
                     |
                     |
                     |
                     |
                     |
                     |
    3             .  +
                     |
                     |
                  #  |
                     |
                 .#  |
                    T|
    2            .#  +
                 ##  |
                     |
                 ##  |
               .###  |
                ### S|
                .##  |
    1           ###  +
               ####  |
             .#####  |T
        ###########  |   G8R
          .######## M|S G2
           .#######  |   G3
           .#######  |
    0        .#####  +M G1
            #######  |   G7R
                ### S|   G5R    G6R
                .##  |S G4
                 .#  |
                 .#  |T
                  #  |
   -1                +
                  # T|
                  .  |
                  .  |
                     |
                     |
                     |
   -2                +
                     |
                     |
                     |
                     |
                     |
                  .  |
   -3                +
               <less>|<freq>
 EACH "#" IS 3: EACH "." IS 1 TO 2
```

```
MEASURE    PERSON - MAP - ITEM
               <more>|<rare>
    4                +
                     |
                     |
                  #  |
                     |
                     |
                     |
    3                +
                  #  |
                     |
                     |
                     |
                  #  |
                     |
    2            ## T+
                     |
                     |
                .##  |
                .##  |
                     |
                  . S|
    1            .#  +
               .###  |
              #####  |
               .###  |T N1
         .#########  |
            .######  |S
        ########### M|   N3
    0    .#########  +M
               .###  |   N2     N4R    N5R
            .######  |S N6R
             ######  |
                 .#  |T
              .#### S|
                  #  |
   -1            ##  +
                  .  |
                  .  |
                     |
                 .#  |
                    T|
                  .  |
   -2                +
                     |
                     |
                     |
                     |
                  .  |
                     |
   -3                +
                     |
                     |
                  .  |
                     |
                     |
                     |
   -4                +
               <less>|<freq>
 EACH "#" IS 3: EACH "." IS 1 TO 2
```

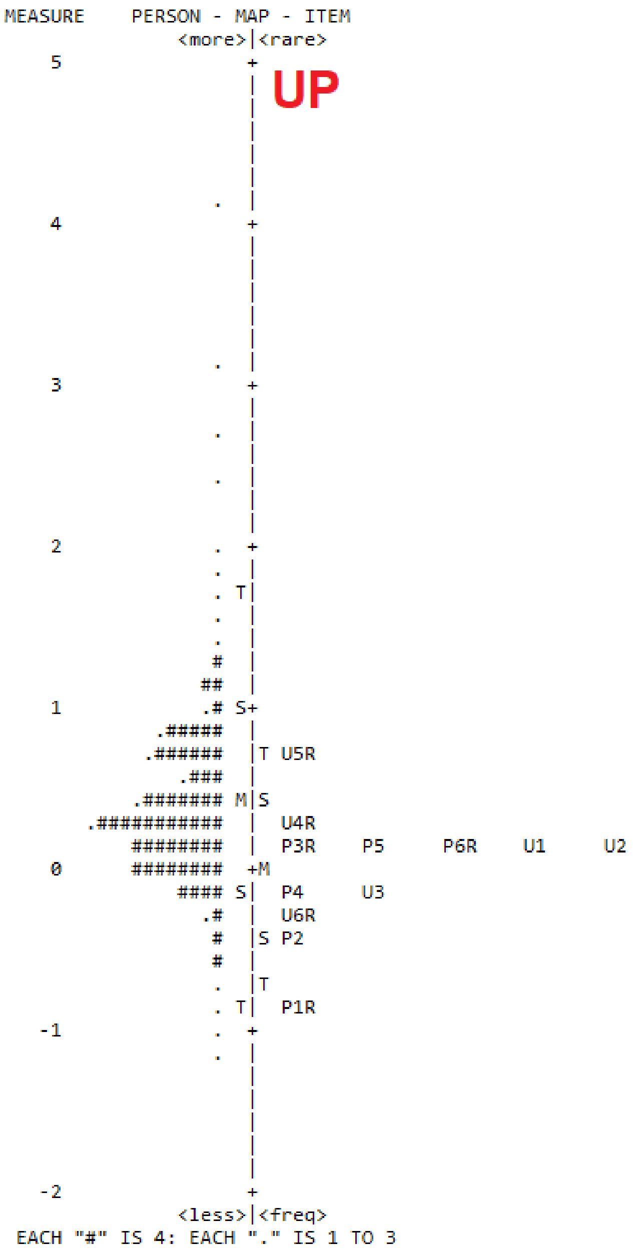
```
MEASURE    PERSON - MAP - ITEM
               <more>|<rare>
    3                +
                     |
                     |
                  .  |
                     |
                     |
                     |
                     |
    2                +
                  .  |
                     |
                  # T|
                     |
                     |
                 .#  |
                     |
    1           .##  +
                ###  |
                .## S|
               .###  |
                .##  |T
              #####  |   L3
               .###  |S
              #####  |   L2
    0  ############  +M L6R
             ######  |   L1     L4R
              ##### M|S
              .####  |   L5R
             ######  |T
              .####  |
                .##  |
                .##  |
   -1            .#  +
                    S|
                  #  |
                .##  |
                     |
               .###  |
                     |
                  .  |
   -2                +
                 .# T|
                     |
                     |
                  .  |
                     |
                     |
                     |
   -3             #  +
                     |
                     |
                     |
                     |
                     |
                  #  |
                     |
   -4            .#  +
               <less>|<freq>
 EACH "#" IS 3: EACH "." IS 1 TO 2
```

Figure 3.5 Wright map based on health-related subscales with original rating scale

The red ellipses shows the gaps in item facet.



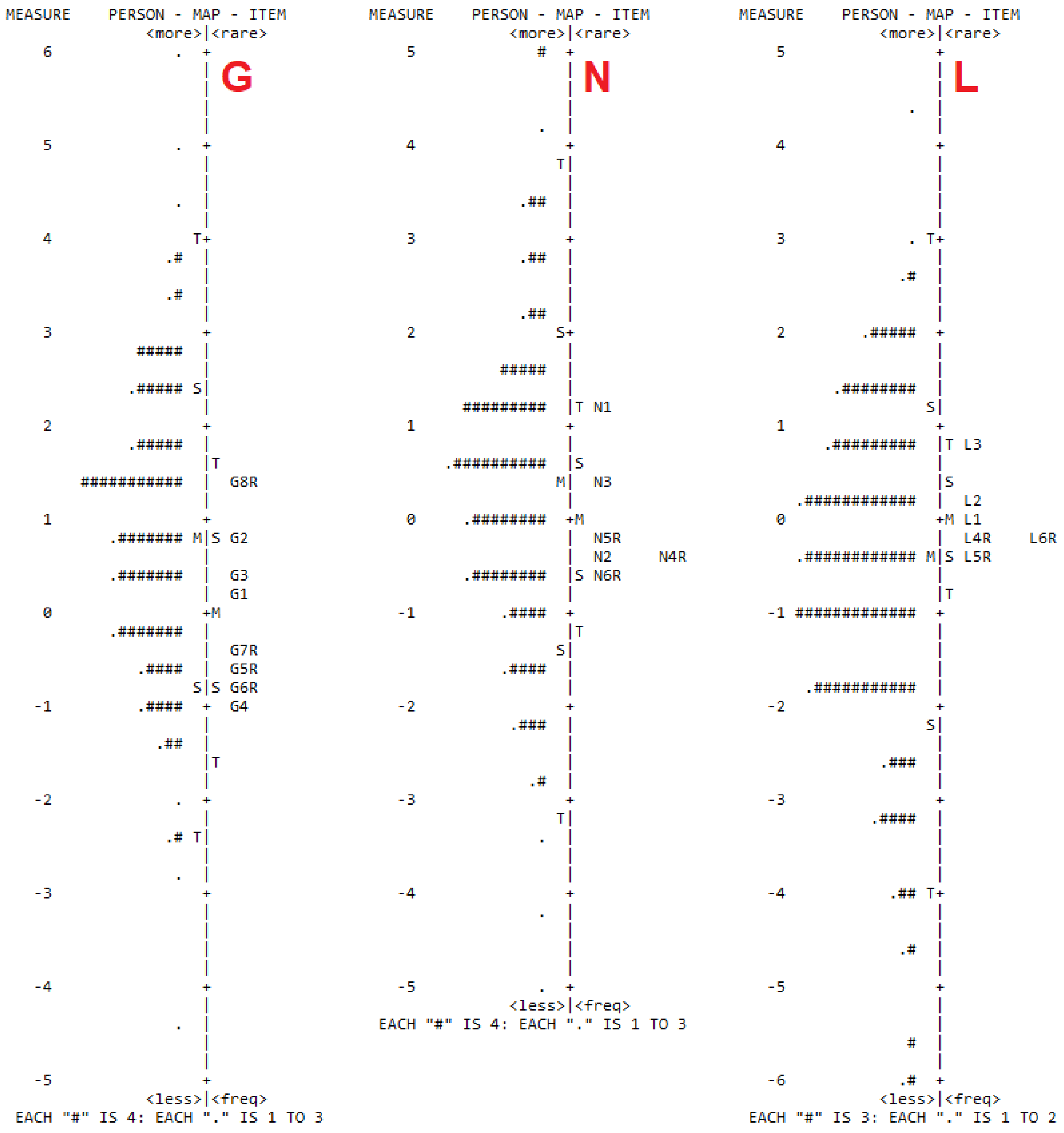
```
MEASURE    PERSON - MAP - ITEM
               <more>|<rare>
    5            .#  +
                     |
                     |
                     |
                     |
                     |
                     |
    4            .#  +
                     |
                     |
                     |
                     |
                     |
                 .#  |
    3                +
                     |
                     |
                 .# T|
                     |
                     |
                 .#  |
    2                +
                     |
                .##  |
                    S|
             ######  |
               .###  |T
                     |
    1       .######  +  C3
             ######  |
             .#####  |
         ########## M|S
            .######  |
             .#####  |  C1     C2
                 ##  |
    0      ########  +M
             .#####  |
               .###  |
                  # S|  C6R
               ####  |S C4R    C5R
                  .  |
                  #  |
   -1             .  +
                 .#  |
                  #  |T
                  . T|
                  .  |
                     |
                     |
   -2                +
               <less>|<freq>
 EACH "#" IS 3: EACH "." IS 1 TO 2
```



```
MEASURE    PERSON - MAP - ITEM
               <more>|<rare>
    5                +
                     |
                     |
                     |
                     |
                     |
                  .  |
    4                +
                     |
                     |
                     |
                     |
                     |
                  .  |
    3                +
                     |
                  .  |
                     |
                  .  |
                     |
                     |
    2             .  +
                  .  |
                  . T|
                  .  |
                  .  |
                  #  |
                 ##  |
    1            .# S+
             .#####  |
            .######  |T U5R
               .###  |
           .####### M|S
       .###########  |  U4R
           ########  |  P3R    P5     P6R    U1     U2
    0      ########  +M
               #### S|  P4     U3
                 .#  |  U6R
                  #  |S P2
                  #  |
                  .  |T
                  . T|  P1R
   -1             .  +
                  .  |
                     |
                     |
                     |
                     |
                     |
   -2                +
               <less>|<freq>
 EACH "#" IS 4: EACH "." IS 1 TO 3
```

Figure 3.6 Wright map based on health-related subscales with original rating scale

C1~C3 were differentiated with C4R~C6R on their endorsability.

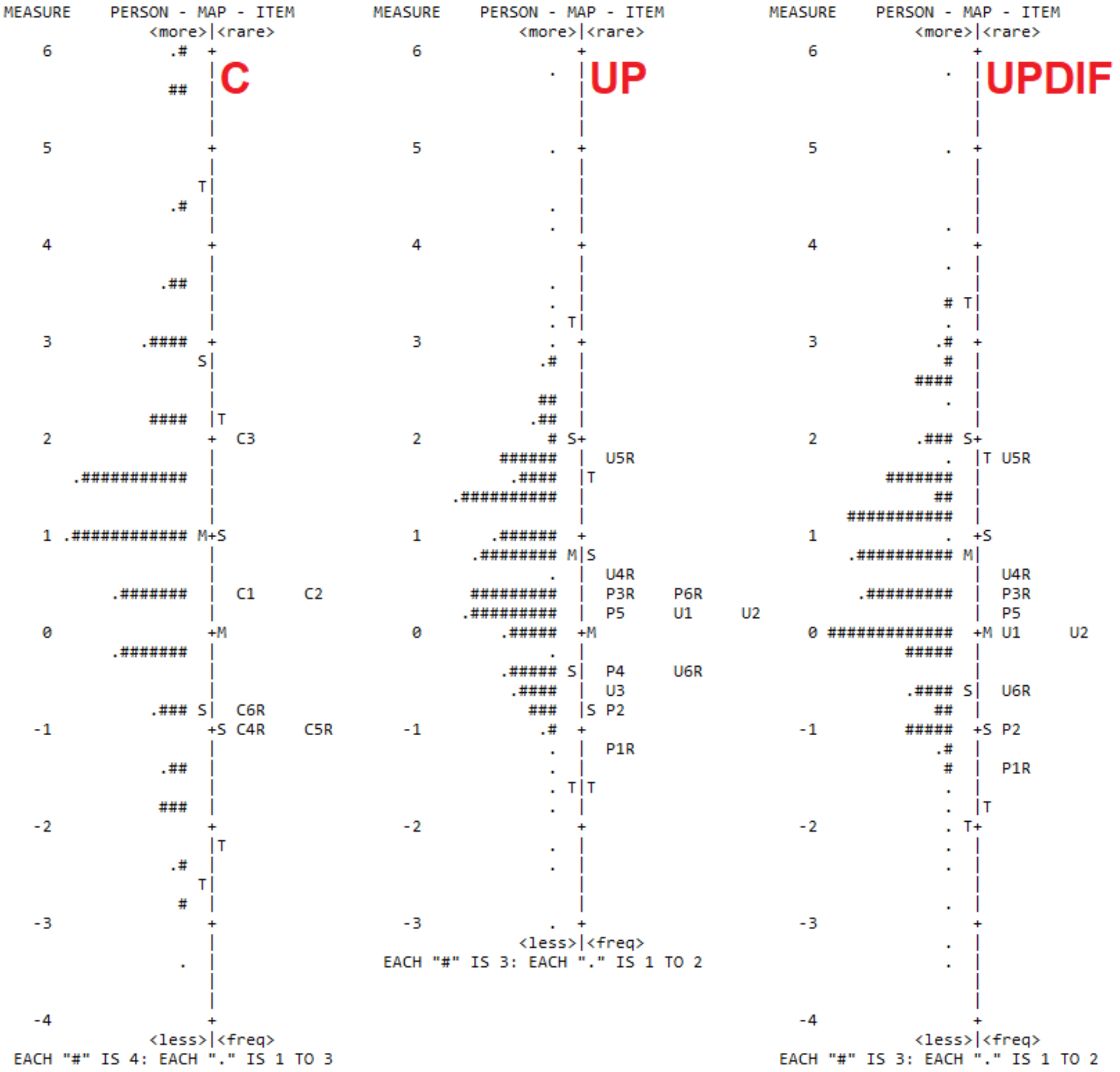


```
MEASURE    PERSON - MAP - ITEM
               <more>|<rare>
    6             .  +
                     |
                     |
                     |
                     |
    5             .  +
                     |
                     |
                  .  |
                     |
    4               T+
                 .#  |
                     |
                 .#  |
                     |
    3                +
              #####  |
                     |
             .##### S|
                     |
    2                +
             .#####  |
                     |T
        ###########  |  G8R
                     |
    1                +
           .####### M|S G2
                     |
           .#######  |  G3
                     |  G1
    0                +M
           .#######  |
                     |  G7R
              .####  |  G5R
                    S|S G6R
   -1         .####  +  G4
                     |
                .##  |
                     |T
                     |
   -2             .  +
                     |
                 .# T|
                     |
                  .  |
   -3                +
                     |
                     |
                     |
                     |
   -4                +
                     |
                  .  |
                     |
                     |
   -5                +
               <less>|<freq>
 EACH "#" IS 4: EACH "." IS 1 TO 3
```

```
MEASURE    PERSON - MAP - ITEM
               <more>|<rare>
    5             #  +
                     |
                     |
                     |
                  .  |
    4                +
                    T|
                     |
                .##  |
                     |
    3                +
                .##  |
                     |
                     |
                .##  |
    2               S+
                     |
              #####  |
                     |
         ##########  |T N1
    1                +
                     |
       .###########  |S
                    M|  N3
                     |
    0     .#########  +M
                     |  N5R
                     |  N2     N4R
         .#########  |S N6R
                     |
   -1         .####  +
                     |T
                    S|
              .####  |
                     |
   -2                +
               .###  |
                     |
                     |
                 .#  |
   -3                +
                    T|
                  .  |
                     |
                     |
   -4                +
                  .  |
                     |
                     |
                     |
   -5             .  +
               <less>|<freq>
 EACH "#" IS 4: EACH "." IS 1 TO 3
```

```
MEASURE    PERSON - MAP - ITEM
               <more>|<rare>
    5                +
                     |
                     |
                  .  |
                     |
    4                +
                     |
                     |
                     |
                     |
    3             . T+
                     |
                 .#  |
                     |
                     |
    2        .#####  +
                     |
                     |
          .########  |
                    S|
    1                +
         .#########  |T L3
                     |
                     |S
      .############  |  L2
    0                +M L1
                     |  L4R    L6R
      .############ M|S L5R
                     |
                     |T
   -1 #############  +
                     |
                     |
                     |
       .###########  |
   -2                +
                    S|
                     |
               .###  |
                     |
   -3                +
              .####  |
                     |
                     |
                     |
   -4           .## T+
                     |
                     |
                 .#  |
                     |
   -5                +
                     |
                     |
                  #  |
                     |
   -6            .#  +
               <less>|<freq>
 EACH "#" IS 3: EACH "." IS 1 TO 2
```

Figure 3.7 Wright map based on health-related subscales with collapsed rating scale



```
MEASURE    PERSON - MAP - ITEM
                <more>|<rare>
   6              .#  +
                      |
                  ##  |C
                      |
                      |
   5                  +
                      |
                     T|
                  .#  |
                      |
   4                  +
                      |
                 .##  |
                      |
                      |
   3           .####  +
                     S|
                      |
                      |
                ####  |T
   2                  +  C3
                      |
       .############  |
                      |
                      |
   1 .############# M+S
                      |
                      |
           .#######   |  C1      C2
                      |
   0                  +M
           .#######   |
                      |
                      |
              .### S|  C6R
  -1                  +S C4R     C5R
                      |
                 .##  |
                      |
                 ###  |
  -2                  +
                      |T
                  .#  |
                     T|
                   #  |
  -3                  +
                      |
                   .  |
                      |
                      |
  -4                  +
                <less>|<freq>
 EACH "#" IS 4: EACH "." IS 1 TO 3
```

```
MEASURE    PERSON - MAP - ITEM
                <more>|<rare>
   6                  +
                   .  |UP
                      |
                      |
                      |
   5               .  +
                      |
                      |
                   .  |
                   .  |
   4                  +
                      |
                   .  |
                   .  |
                   . T|
   3               .  +
                  .#  |
                      |
                  ##  |
                 .##  |
   2               # S+
              ######  |  U5R
               .####  |T
         .##########  |
                      |
   1         .######  +
           .######## M|S
                   .  |  U4R
           #########  |  P3R     P6R
          .#########  |  P5      U1      U2
   0          .#####  +M
                   .  |
              .##### S|  P4      U6R
               .####  |  U3
                 ###  |S P2
  -1              .#  +
                   .  |  P1R
                   .  |
                   . T|T
                   .  |
  -2                  +
                   .  |
                   .  |
                      |
                      |
  -3               .  +
                <less>|<freq>
 EACH "#" IS 3: EACH "." IS 1 TO 2
```

```
MEASURE    PERSON - MAP - ITEM
                <more>|<rare>
   6                  +
                   .  |UPDIF
                      |
                      |
                      |
   5               .  +
                      |
                      |
                      |
                   .  |
   4                  +
                   .  |
                      |
                   # T|
                   .  |
   3              .#  +
                   #  |
                ####  |
                   .  |
                      |
   2            .### S+
                   .  |T U5R
             #######  |
                  ##  |
         ###########  |
   1               .  +S
         .########## M|
                      |  U4R
          .#########  |  P3R
                      |  P5
   0 #############    +M U1      U2
               #####  |
                      |
               .#### S|  U6R
                  ##  |
  -1           #####  +S P2
                  .#  |
                   #  |  P1R
                   .  |
                   .  |T
  -2               . T+
                   .  |
                   .  |
                      |
                   .  |
  -3                  +
                   .  |
                   .  |
                      |
                      |
  -4                  +
                <less>|<freq>
 EACH "#" IS 3: EACH "." IS 1 TO 2
```

Figure 3.8 Wright map based on taste-related subscales with collapsed rating scale

3.4.2.6 Reliability

Table 3.17 lists the information regarding the Rasch separation, strata and reliability statistics of each subscale. As stated in section 2.3.5, the “real” reliability statistics calculated with misfit-inflated standard error were reported here because they are more conservative (Boone *et al.*, 2013).

According to the results, the person-related separations were similar in all subscales (around 2), indicating that two statistically distinct levels can be discerned under each subscale individually. This was acceptable but not ideal. Better discrimination power may be achieved if more items can be added to the subscales. In addition, a noticeable decrease in reliability and separation statistics was found with subscale UP after three items were removed due to DIF. This was expected because the length of this subscale was reduced from twelve items to nine items, although the separation of 1.57 was still above the minimum requirement of 1.50 suggested by Tennant and Conaghan (2007).

Compared to the respondent facet, the reliability statistics were high in the item facet, implying that the sample size of the survey (269) was adequately large.

Table 3.17 Rasch reliability statistics of each subscales

Instrument	**Respondents**			**Items**		
	Separation	Strata	Reliability	Separation	Strata	Reliability
Original rating scale						
G	1.88	2.84	0.78	6.61	9.17	0.98
L	1.89	2.85	0.78	5.18	7.24	0.96
N	2.03	3.04	0.81	3.86	5.48	0.94
C	1.90	2.87	0.78	10.60	14.47	0.99
UP	1.55	2.40	0.70	6.00	8.33	0.97
Collapsed rating scale						
G	1.76	2.68	0.76	6.13	8.51	0.97
L	1.75	2.67	0.75	4.78	6.71	0.96
N	1.78	2.71	0.76	2.77	4.03	0.88
C	1.77	2.69	0.76	8.66	11.88	0.99
UP	1.77	2.69	0.76	6.43	8.91	0.98
UPDIF	1.57	2.43	0.71	7.44	10.25	0.98

3.5 Differences on the estimation of person parameters between CTT and Rasch analysis

Since the same underlying structure was identified by both EFA and PCAR from the 20 health-related items, the stability of estimation and the estimates of each respondent obtained from different methods (*i.e.* the CTT and Rasch analysis) on the health-related subscales G, N and L[2] can be compared after rescaling both the raw scores and the Rasch person measures to the same range of 0~100.

[2] G=General health interest; N=Natural product interest; L=Light product interest.

3.5.1 Methods

3.5.1.1 Comparing the stability of estimation

The stability of estimation can be compared using the standard error of measurement (SE_m) as indicator. The CTT assumes that the SE_m is constant across the given population, which can be calculated using the equation 2.36. On the contrary, Rasch analysis generalises the SE_m values across the population, which is dependent on construct levels. In this study, the SE_m values estimated by CTT on subscales G, N and L were compared to both “model SE” and “real SE” statistics reported by Rasch analysis on the same subscales. The “model SE” sets up the lower bound of the measurement imprecision, while the real SE stands for the misfit-inflated standard error, which is the upper bound of the measurement imprecision. They were estimated using the equations 2.38 and 2.40, respectively[3].

3.5.1.2 Comparing the respondents’ estimates

The respondents’ estimates on each subscale were computed by three methods:

(1) The CTT-based raw total scores summated from the raw scores of items belonging to each subscale;

(2) The person measures modelled within the reference of original 7-point rating scale by Rasch analysis;

(3) The person measures modelled within the reference of collapsed 4-point rating scale by Rasch analysis.

In order to obtain statistical inference, a series of independent t-tests were conducted using R program (R Core Team, 2018) to compare each person’s location on the subscales G, N and L estimated by different approaches. The homogenised SE_m provided by CTT and the “model SE” values associated with individual Rasch measures were used in the t-tests. The “model SE” values were selected for the comparison because they are equal to or smaller than the “real SE”[4] values. Therefore the critical distance computed in the t-test based on the “model SE” would be smaller than that based on the “real SE”. After the t-tests were conducted, the number of significant pairs were recorded.

[3] See section 2.3.6 for more details about the SE_m.

[4] See section 2.3.6.2

3.5.2 Results

3.5.2.1 The stability of estimation of person parameters within different frameworks

Figures 3.9 and 3.10 compare the SE_m estimated using both CTT and Rasch analysis.

Firstly, most of the model SE values obtained by Rasch analysis with both original and collapsed rating scales were smaller than the homogenised SE_m estimated by CTT (figure 3.9), indicating that the lower bound of the measurement error produced by Rasch analysis was consistently smaller than the measurement error associated with the CTT in this study.

Secondly, most of the real SE values estimated by Rasch analysis were also smaller than the SE_m obtained by CTT (figure 3.10). Since the real SE presents the upper bound of the measurement error, the results further confirm that the measurement errors are smaller when using Rasch analysis than using CTT for estimating the person parameters with current data.

Lastly, the SE_m values (both model SE and real SE) estimated by Rasch analysis with the original rating scale distributed across the whole scale range in a “U” shape. The SE_m values were smallest at the centre of the scale and biggest at the two ends. However, this U-shaped distribution was not observed when using the collapsed scale. The SE_m values increased significantly at the centre of the scale after collapsing the scale categories.

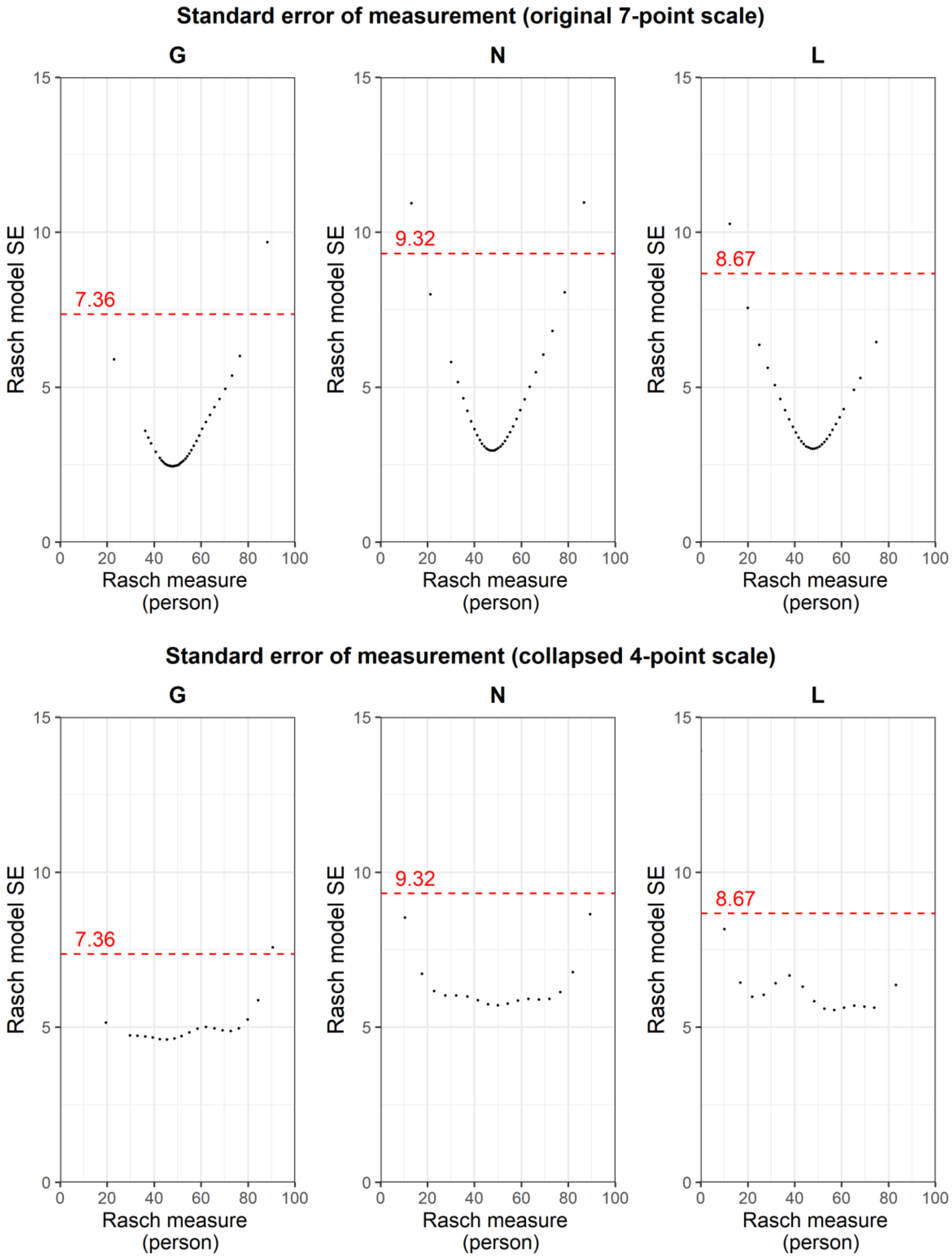


Figure 3.9 Comparisons between the SE_m estimated by CTT (in red) and the model SE estimated by Rasch analysis on the three health-related subscales.

(G=General health interest; N=Natural product interest; L=Light products interest)

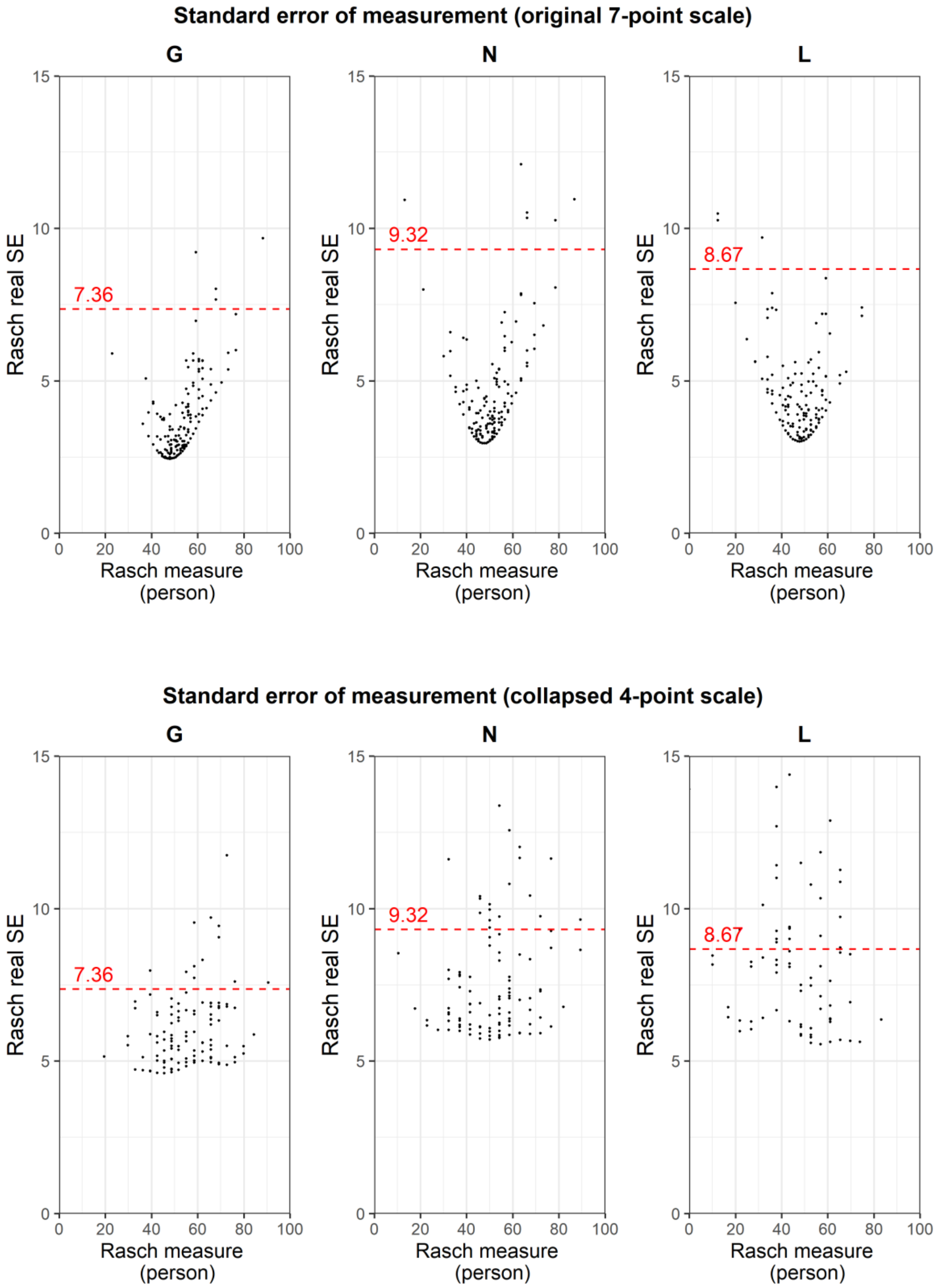


Figure 3.10 Comparisons between the SE$_m$ estimated by CTT (in red) and the real SE estimated by Rasch analysis on the three health-related subscales

3.5.2.2 Comparisons between the person estimates obtained by CTT and Rasch analysis

(1) Direct comparison

A direct comparison between the estimates of the same respondents on the health-related subscales G, N and L obtained by different approaches is visualised in figure 3.11. On one hand, compared to the Rasch measures modelled using the original 7-point rating scale, the raw scores of the respondents who had relatively low agreeability on the subscales were underestimated, while the raw scores of the others with relatively high agreeability on the subscales were overestimated. This is due to the non-linearity of the rating scale. On the other hand, the values of Rasch measures produced within the reference of original and collapsed scales are close.

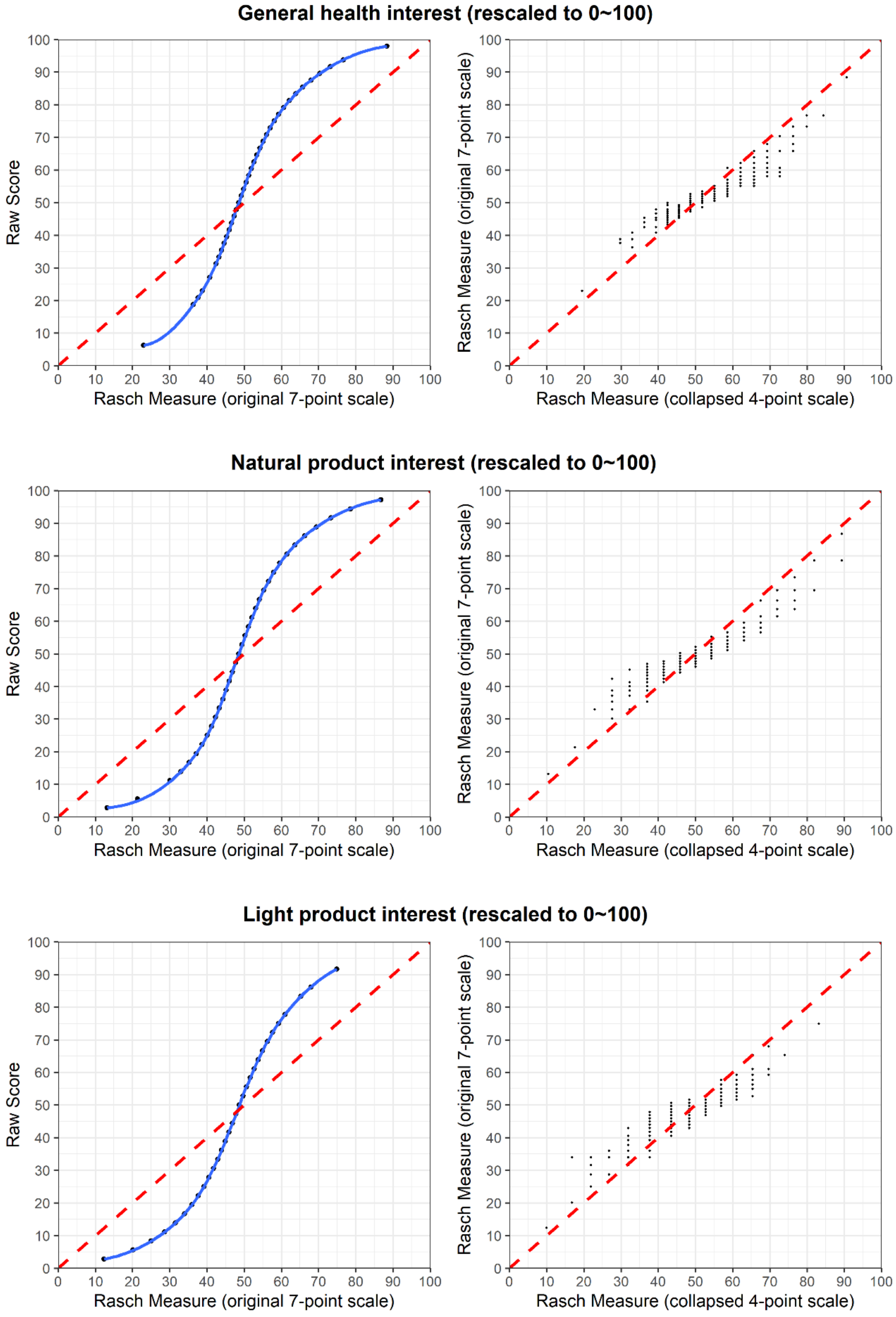
(2) Independent t-tests

Table 3.18 summarises the results of the independent t-tests which compared every respondent's raw scores and Rasch measures on subscales G, N and L.

Firstly, no significant result was identified between the Rasch measures modelled using original and collapsed scale, indicating that collapsing scale categories did not change the accuracy of the measurement.

Secondly, no significant difference was found in the estimates of individual respondents when comparing their raw scores and Rasch measures modelled using either the original scale or the collapsed scale on subscales N and L.

Lastly, 36 significant cases were reported from the t-tests in the comparison between the raw scores and Rasch measures modelled using the original 7-point scale on subscale G, which accounts for 13.38% of the respondents.


General health interest (rescaled to 0~100)

Natural product interest (rescaled to 0~100)

Light product interest (rescaled to 0~100)

Figure 3.11 Comparisons between the raw scores and Rasch measures on the three health-related subscales

Table 3.18 Significant cases reported from independent t-tests between the respondents' estimates obtained by CTT and Rasch analysis (total number=269)

Independent t-tests on person estimates	**Subscales**		
between	**G**	**N**	**L**
Raw score and Rasch measure (original 7-point scale)	36 (13.38%)	0	0
Raw score and Rasch measure (collapsed scale)	2 (0.74%)	0	0
Rasch measure (original 7-point scale) and Rasch measure (collapsed 4-point scale)	2 (0.74%)	0	0

3.6 Differences on health and taste attitudes between groups by gender and age

3.6.1 Methods

The effects of gender and age on were examined by comparing the difference of raw total scores or measures of each subscale, between the respondents grouped by gender or age. For the comparison between gender groups, the Levene's test (Levene, 1960) was employed first for examining the homogeneity of variance. After that, the independent t-test with or without equal variance was conducted to compare the difference. Moreover, the effect size was estimated in the form of Cohen's d (Cohen, 1988).

For the comparison between age groups, the ANOVA and non-parametric Kruskal-Wallis test (Kruskal and Wallis, 1952) were performed. In addition, the assumptions of ANOVA were also examined via residual analysis. For residual analysis, firstly, the Shapiro-Wilk test (Shapiro and Wilk, 1965) was applied for examining the assumption of normality. Next, if the assumption of normality can hold, then the Levene's test was used for investigating the homogeneity of variance. Otherwise, the non-parametric Brown-Forsythe test (Brown and Forsythe, 1974) was used for this purpose. After that, the Bonferroni outlier test was conducted. If the assumptions of ANOVA could hold within the sampled data set, then Tukey's HSD test (Tukey, 1949) would be employed for multiple comparisons. If they could not hold, then Dunn's test (Dunn, 1964) with Hochberg correction (Hochberg, 1988) would be used for multiple comparisons. In addition, the Hays' omega-squared (ω^2) (Hays, 1963) was computed as the measure of effect size for ANOVA.

The R program (R Core Team, 2018) was used for independent t-test, ANOVA, Kruskal-Wallis test, Tukey HSD test and Shapiro-Wilk test, while the Levene's test, and Bonferroni outlier test were conducted using R package car (Fox *et al.*, 2018). The R packages vGWAS (Shen, 2015), PMCMRplus (Pohlert, 2018), effsize (Torchiano, 2018) and sjstats (Lüdecke, 2019) were used for Brown-Forsythe test, Dunn's test with Hochberg correction, and the estimation of Cohen's d and Hays' omega-squared (ω^2), respectively.

3.6.2 Results

The differences between gender and age groups on health and taste attitudes were compared. Tables 3.19 and 3.20 tabulate the information about the test statistics and assumptions. Figures 3.12 and 3.13 show the results of multiple comparisons by groups, where the rating scales were all rescaled to a range of 0-100 for illustration purposes.

3.6.2.1 Effect of gender

A significant difference between female and male on health-related subscales was only found on subscale G (general health interest) using the measures modelled with scale-collapsed data by Rasch analysis. The result revealed that females are more concerned about general health issue than males.

The differences between gender groups were not significant on most of taste-related subscales, except subscale CB (attitudes towards the other people's craving for sweet foods) within CTT approach. Within Rasch analysis, the p-values of t-test for gender difference on subscale C, which was split into two subscales CA and CB in CTT approach, were 0.121 and 0.060 with raw data and scale-collapsed data, respectively. They were just above the significance level of 0.05. A possible reason is when items belonging to subscale CA and CB defined by CTT approach were modelled together, the difference between gender groups were cancelled.

In addition, the values of Cohen's d of the comparisons were all below 0.5, suggesting only small effect size could be identified from the difference between females and males (Cohen, 1988)

3.6.2.2 Effect of age

For health-related attitudes, a conclusion that the respondents of 35+ group are more concerned about the G (general health issues) and N (the use of natural product) than the other two age groups can be drawn for total score or measure obtained from all methods.

For taste-related attitudes, the respondents aged below 35 showed significantly greater interests on U (using food as reward) compared to those aged 35 and above, according to the comparison of raw total scores on subscale U (see figure 3.13). Within Rasch analysis, although the null-hypothesis of Kruskal-Wallis test was rejected in comparing the difference between ages groups on the person measures of subscale UP modelled with original rating scale and collapsed rating scale, the results of multiple comparisons suggested that with scale-collapsed data, if the DIF was not resolved, the significance of the difference between ages groups on the same subscale UP may not be confirmed.

Moreover, only three of the omega squared ω^2 statistics were slightly greater than 0.06, which represents the lower bound of the small effect size (Kirk, 1996).

Table 3.19 Comparisons of raw total scores or Rasch measures between gender groups on each subscale

Subscale		Gender		
		Levene's test	t-test	Effect size d
Exploratory factor analysis: Original rating scale	**G**	0.347	0.065	0.251
	N	0.721	0.288	0.145
	L	0.873	0.266	0.152
	CA	0.458	0.218	0.168
	CB	**<0.001**	**0.005**	0.467
	U	0.321	0.618	0.068
	P	0.830	0.639	0.064
Rasch analysis: Original rating scale	**G**	0.844	0.072	0.245
	N	0.314	0.180	0.182
	L	0.462	0.268	0.151
	C	0.328	0.121	0.212
	UP	**0.048**	0.414	0.111
Rasch analysis: Collapsed rating scale	**G**	0.523	**0.042**	0.278
	N	0.371	0.164	0.190
	L	0.567	0.212	0.170
	C	0.451	0.060	0.257
	UP	0.072	0.477	0.097
	UPDIF	0.252	0.861	0.024

Significant p-values (<0.05) are in **bold**

Table 3.20 Comparisons of raw total scores or Rasch measures between age groups on each subscale

Subscale		**Age**						
		Shapiro-Wilk test	**Levene's test**	**Brown-Forsythe test**	**Bonferroni outlier test**	**ANOVA**	**ANOVA Effect size ω²**	**Kruskal-Wallis test**
Exploratory factor analysis: Original rating scale	**G**	**0.010**	0.065	0.053	NA	**<0.001**	0.046	**<0.001**
	N	0.129	0.292	0.292	NA	**<0.001**	0.061	**<0.001**
	L	**0.008**	0.986	0.973	NA	0.697	0.005	0.693
	CA	**<0.001**	0.071	0.074	NA	0.078	0.012	0.084
	CB	**<0.001**	0.058	0.224	NA	0.636	0.004	0.924
	U	**<0.001**	**0.003**	**0.005**	NA	**<0.001**	0.070	**<0.001**
	P	**0.008**	0.123	0.136	NA	0.684	0.005	0.657
Rasch analysis: Original rating scale	**G**	**<0.001**	0.851	0.871	3 extreme outliers	**0.005**	0.032	**<0.001**
	N	**<0.001**	0.559	0.694	4 extreme outliers	**<0.001**	0.052	**<0.001**
	L	**<0.001**	0.172	0.310	5 extreme outliers	0.709	0.005	0.693
	C	**<0.001**	0.040	0.067	5 extreme outliers	0.740	0.005	0.187
	UP	**<0.001**	0.144	0.168	2 extreme outliers	**0.042**	0.016	**0.028**
Rasch analysis: Collapsed rating scale	**G**	0.375	0.252	0.292	NA	**0.004**	0.033	**0.003**
	N	**0.019**	0.135	0.138	NA	**<0.001**	0.065	**<0.001**
	L	**<0.001**	0.747	0.807	NA	0.889	0.007	0.812
	C	**<0.001**	0.264	0.246	NA	0.575	0.003	0.340
	UP	**<0.001**	0.218	0.235	1 extreme outlier	**0.033**	0.018	**0.036**
	UPDIF	**0.010**	0.197	0.208	1 extreme outlier	**0.002**	0.039	**0.002**

Significant p-values (<0.05) are in **bold**

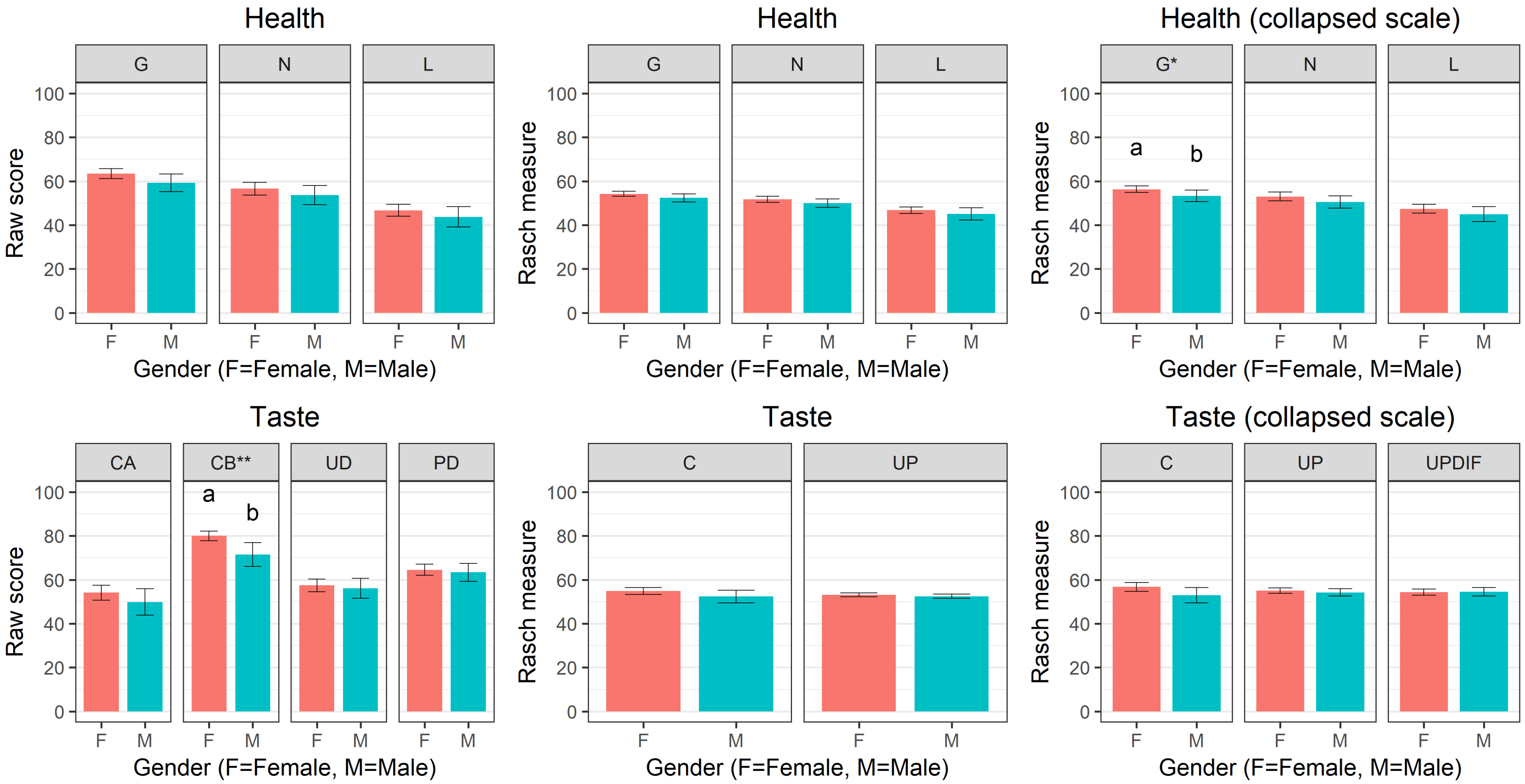


Figure 3.12 Difference between Females (n=194) and Males (n=75) groups for Health and Taste related subscales

* and ** represents significance levels at 0.05 and 0.01, respectively; error bars = 95% CI.

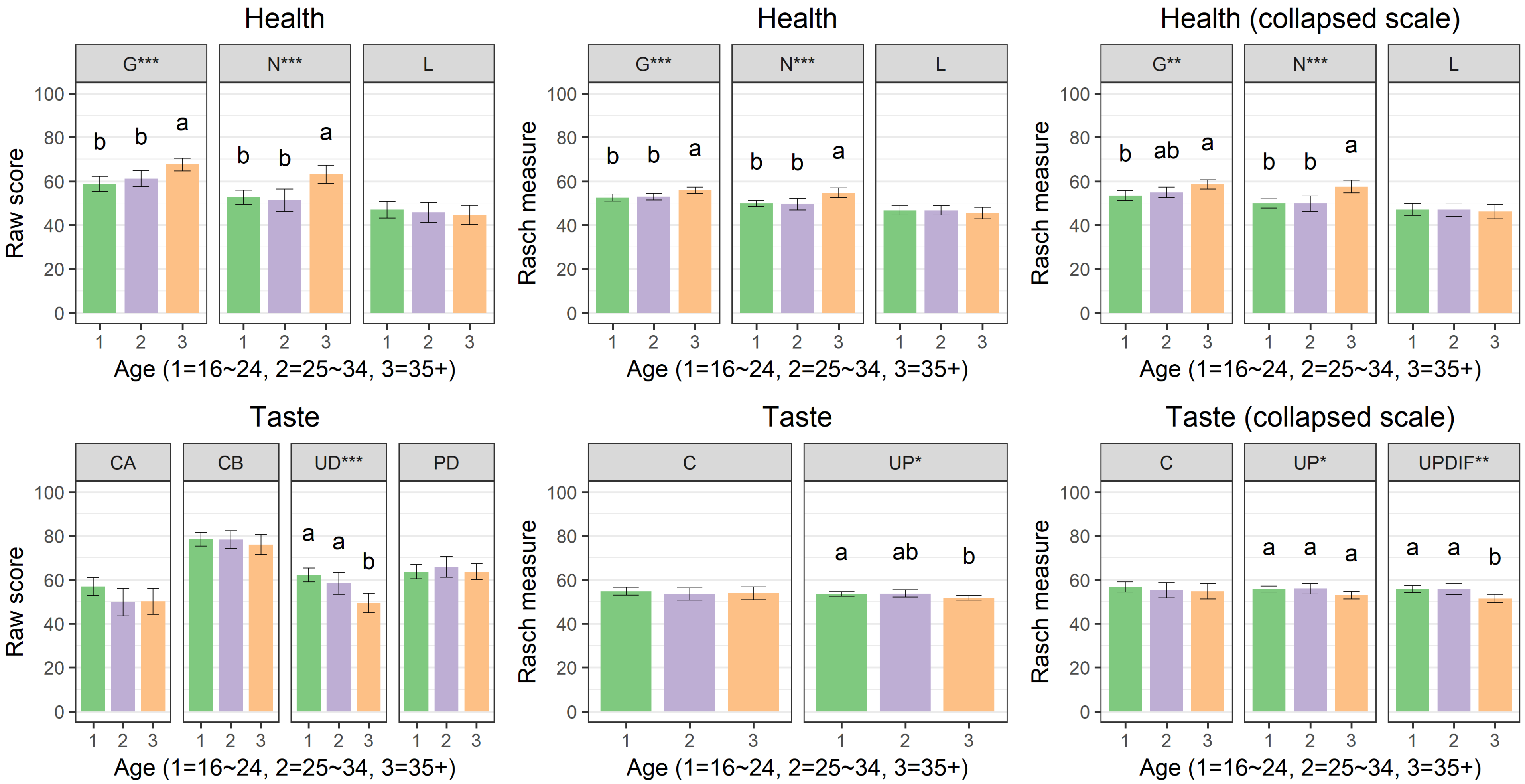


Figure 3.13 Difference between 16~24 (n=114), 25~34 (n=68) and 35+ (n=87) age groups for health and taste related subscales

*, ** and *** represents significance levels at 0.05, 0.01 and 0.001, respectively; error bars = 95% CI.

3.7 Discussion

3.7.1 CTT approach vs. Rasch analysis

3.7.1.1 Dimensionality of the instrument

In this study, both CTT approach using factor analysis on raw scores and Rasch analysis using PCA on model residuals followed by t-test identified three sub-dimensions from the twenty health-related items. The same result was given by original research (Roininen *et al.*, 1999). However, the number of sub-dimensions predicted by the initial parallel analysis on raw scores was four for health-related items.

For taste-related items, parallel analysis on raw scores implied that there might be five underlying factors, but only four factors were confirmed by factor analysis. The subscale U and P suggested by original research were retained. However, items U6R, P1R and P6R were dropped in this study. The original subscale C was split into two subscales in this study by CTT approach, including subscale CA that consisted of three positive-worded items, and subscale CB made up of three reversed-worded items. In contrast with the suggestion of four factors, Rasch analysis discerned only two sub-dimensions with same items. The results indicated that all items belonging to subscale C identified by original research remained together, whereas the other two pre-defined subscale U and P should be combined to form a new subscale UP.

The factor analysis relies on the computation of Pearson correlation matrix, which requires interval level of data. However, CTT approach usually misuses the raw scores of the ordinal items such as the Likert items used in this study, as interval data. It has been found that the Pearson correlation would underestimate the relationship between ordinal items (Olsson, 1979). Consequently, factor analysis would extract an excessive number of factors including artificial factors (Bernstein and Teng, 1989). By contrast, Rasch analysis can overcome the limitation of factor analysis. It starts from assuming the instrument is unidimensional, while the residuals are merely random noises. It tests whether this assumption can hold by inspecting the amount of common variance shared by the linear model residuals (*i.e.* to what extent the secondary dimensions account for the unexplained variance). It is not affected by the biased correlation on ordinal observations. This is a possible reason that, in this study, there were four factors from taste-related items reported by factor analysis, but only two sub-dimensions identified using Rasch analysis. Each Rasch sub-dimension was in fact a combination of two factors extracted by factor analysis.

Moreover, Ferguson (1941) argued that factor analysis cannot clearly distinguish the difference from the nature of the underlying construct and the difference in the scale levels of the same construct. Duncan (1984) further pointed out that the inter-item correlation and the item loadings are influenced by the scale levels of the items. To be more specific, in a unidimensional instrument, the items that are difficult to be endorsed by respondents may not strongly correlate with the items that has higher endorsability. As a result, they may not load together. Thus factor analysis may report spurious factors that represent the same underlying construct but at different scale levels separately. For example, in this study, according to the meanings of the statements of the six items that came from original subscale “Craving for sweet foods”, the three items C1~C3 referred to respondents' attitudes about their own cravings for sweet foods, while the other three items C4~C6 concerned respondents' attitudes on the other people's cravings for sweet foods. They were discerned into two factors in factor analysis, even if they were designed to discuss the same phenomenon. According to Rasch analysis, the items C1~C3 showed lower endorsability than C4~C6 (see figure 3.6), possibly because the respondents are more strict to themselves on diet. They can be interpreted as two different levels on the same construct. On the contrary, Rasch analysis can avoid this issue because it constructs a hierarchical structure from the beginning by taking the difference of item response pattern into account. In this research, Rasch analysis indicated that there was only one sub-dimensions that can be explained by all six items related to “craving for sweet foods”. This result is more reasonable.

In addition, after factor analysis, the factors would be interpreted according to the meanings of correlated items reported under the same factor, where disputed name may be entitled to the factor (Wright, 1991). While in Rasch analysis, the assumed unidimensional construct would be defined before testing the unidimensionality. Only the items belonging to the secondary dimension, if there is one, need to be defined.

In summary, factor analysis on raw scores and Rasch PCA on standardised model residuals can obtain different results about the dimensionality of the instrument. Rasch analysis can overcome the limitations of the Factor analysis, providing nuisance-free interpretation on the underlying structure of the instrument.

3.7.1.2 Reliability

The Cronbach's alpha coefficients estimated based on ordinal raw scores were slightly higher than the Rasch separation reliability statistics calculated with interval measures on the same subscales G, N and L, which is expected. Cronbach’s alpha was estimated with the whole data set including the extreme scores. On the contrary, the calculation of Rasch separation reliability excluded the extreme measures because they are indefinite located with infinite standard error. Rasch separation reliability is more conservative and less miss leading (Clauser and Linacre, 1999; Linacre, 1997).

The reliability statistics were not comparable between the other subscales because they contain different items within different approach. However, it is noticeable that the Cronbach’s alpha of subscale P within CTT approach (0.61) was below than recommended minimum value of 0.70. According to Spearman-Brown prophecy formula (Brown, 1910; Spearman, 1910), to achieve an acceptable alpha of 0.70 on subscale P, when assuming all items are equally correlated, the ratio of desirable test length over current test length for this sampled population should be

$$j = \frac{0.70 \times (1 - 0.61)}{0.61 \times (1 - 0.70)} \approx 1.49$$

Then the subscale should be increased from current 4 items to at least 4x1.49≈6 items (*i.e.* at least 2 additional items are needed).

This, however, can be considered as an extension of the consequence of using deficient factor analysis for dimensionality test. This can be avoided, in this case, if Rasch analysis was applied. The original subscale U and P formed one sub-dimension in Rasch analysis, which can provide more reliable result. In fact, the items in original subscale U and P share common meanings. Logically speaking, “using food as a reward” is for “pleasure”.

In addition, the reliability can be increased by introducing more items to the scale, which can be guided by Rasch analysis. To differentiate people, an ideal instrument should be an assembly of items that covers different scale levels. The Wright map can provide support to this task. For example, as can be seen in figure 3.5, one can identify some gaps between or beyond current items. The respondents located in the gap cannot be well differentiated. Therefore the new developed items should be designed to fill in the gaps.

3.7.1.3 Estimation of person parameter

Although the respondents’ estimates obtained by CTT and Rasch analysis were almost statistically equivalent on the three health-related subscales, Rasch analysis produced less measurement errors than CTT, indicating that replacing CTT by Rasch analysis can improve the precision of the measurement.

3.7.1.4 Group effects

Similar results regarding the difference between gender and age groups on HTAS were given with the ordinal total scores in CTT approach and the continuous measures modelled by Rasch analysis. Rasch analysis showed slightly high sensitivity on detecting group effects. The significant difference between genders on subscale G was only suggested by Rasch analysis with the collapsed rating scale. However, this cannot be considered as a strong evidence of preferring Rasch analysis over CTT approach.

3.7.2 Effect of collapsing rating scale (Rasch analysis)

In this study, the analysis of category effectiveness indicated that the differences between some adjacent scale categories on their definition were too small to be distinguished by respondents. Therefore the original 7-point rating scale was collapsed to a 4-point rating scale.

The item fits were slightly improved with the collapsed rating scale, which was expected (see section 2.5.1 for explanation). This showed a benefit of resolving disordered Rasch-Andrich thresholds by collapsing categories.

The improvements of item fits were followed by reductions of reliability statistics, although the decreases were not big enough to harm the discrimination power of the measurement. Collapsing scale categories can result in a loss of information. When this happened, little variance would be yielded, which has a negative effect on the reliability (Daher *et al.*, 2015). However, Linacre (2014c) suggested that this is not an issue unless the person reliability statistics become too small to separate people into statistical distinct levels due to categories being collapsed. This contradiction between improved fit and decreased reliability reflected from this study is instructive for future practice. One should always monitor the change of reliability when revising the rating scale by reducing the number of categories during instrument development.

3.7.3 Resolving DIF

A few differences were found in this study when comparing subscale UP with collapsed rating scale before and after removing the three items with DIF.

Firstly, the length of the subscale was reduced from twelve items to nine items, which had a negative impact on the reliability of the measurement. The Rasch reliability statistics were maintained above acceptable levels after the removal of DIF items, thus no other action was done. However, if the separation dropped below the recommended value of 1.5 due to excluding DIF items, one should rethink the method used for resolving DIF. For instance, one can treat the item as separated items rated by different groups (Tennant *et al.*, 2004), although this method is more complex than directly removing items in practice.

Secondly, resolving DIF can minimise the influence of item bias between groups on the whole scale, so that accurate difference between groups can be evaluated. In this study, a multiple comparison procedure found no significant difference between age groups on subscale UP with collapsed rating scale. But, a different conclusion was drawn from the revised data after dropping the three DIF items. This indicated that, to obtain meaningful results of group effect, one should always identify and resolve DIF prior to comparison.

3.7.4 Conclusion

In conclusion, Rasch analysis can break through the restrictions of CTT approach by providing meaningful interpretation on dimensionality of an instrument, less misleading reliability indicators, and more stable estimation of person parameters. Furthermore, it can optimise the rating scale structure by collapsing the categories that represent narrow intervals on a scale, and provide bias-minimised group comparison after resolving DIF. It should be applied to consumer survey if ordinal rating scale such as Likert scale is used.

Chapter 4 Case study II: Measuring the overall liking in sensory acceptability test using Many-Facet Rasch model – a comparison between composite measurement and individual measurement

4.1 Introduction

4.1.1 Aim

This study explores the application of Rasch analysis in a sensory acceptability test. A three-facet Many-Facet Rasch Rating scale (MFR-RS) model was used to model the estimates of panellists, products and attributes. Twenty-four products across nine food and drink categories were tested by 192 participants using an incomplete design. The consumers' overall liking on products was determined using two methods:

(1) A composite measure that was modelled using a set of attribute ratings;
(2) A holistic measure modelled using a single overall acceptability attribute.

4.1.2 Research hypothesis

This study tested the hypothesis that a composite measure of consumers' overall liking of a food product modelled using sensory attribute ratings has a greater power to discriminate products than a holistic measure modelled using a single overall acceptability item.

4.2 Participants and sampling procedures

The experimental work was approved by Faculty Research Ethics Committee (MEEC14-027). Participants were recruited from the campus of the University of Leeds via email, poster and personal contact.

During the pre-screening session, people were excluded for safety reasons if:

(1) they were allergic or intolerant to any sample used in this study;
(2) they were pregnant or lactating.

To ensure the accuracy of the sensory study data, people were also excluded if:

(1) they were ill or suffering from any underlying health condition that could affect their ability to taste, smell, chew, digest or expectorate samples;
(2) they were taking any medication at the panel date.

Eventually a total number of 192 panellists attended the study.

4.2.1 Sample selection

Twenty-four commercial products across nine food and drink categories, as shown in table 4.1, were purchased from local supermarkets as samples. The storage instructions were strictly followed during the experimental period to ensure food safety and quality.

4.2.2 Sensory attributes and rating scale

The panellists were asked to rate their perception of the samples via a questionnaire that consisted of four sensory modalities, four individual attributes and the overall acceptability, using an 11-point hedonic scale labelled from "Greatest Imaginable Dislike" to "Greatest Imaginable Like". It consisted of the eleven hedonic descriptors taken from the labelled affective magnitude scale (Schutz and Cardello, 2001). The 11-point hedonic scale had been proved to perform "*equally well*" with the more commonly used 9-point hedonic scale by Lawless *et al.* (2010). Other researchers (Lim and Fujimaru, 2010; Schutz and Cardello, 2001) argued that, compared to the 9-point scale, the 11-point scale can provide more options for panellists to rate the attributes, thus increasing the ability of discriminating the extreme liked/disliked product affected by the celling effect. The list of the attributes and their labels are tabulated in Table 4.2.

Table 4.1 24 Products evaluated in this study.

Food and drink categories	Product	Sample label	Sample Code
Chocolate	Morrisons' Dark	Chocolate (DA)	368
	Morrisons' Milk	Chocolate (MI)	455
Crisps	Walker's light ready salted	Crisps (LRS)	950
	Walker's Ready salted	Crisps (RS)	807
Digestive Biscuit	McVitie's	Digestive Biscuit (MC)	313
	WeightWatchers	Digestive Biscuit (WC)	527
Fizzy Drink	Coca Cola	Fizzy Drink (CC)	123
	Diet Coke	Fizzy Drink (DC)	344
Juice	Morrisons' Grapefruit	Juice (GR)	376
	Morrisons' Orange (smooth)	Juice (OR)	211
Milk	Morrisons' Full fat	Milk (FF)	798
	Morrisons' semi-skimmed	Milk (SS)	741
	Morrisons' Skimmed	Milk (SK)	152
	Alpro' soya milk	Milk (SO)	918
Pickles	Morrisons' pickled black olive	Pickles (BO)	304
	Morrisons' pickled cucumber	Pickles (CU)	838
	Morrisons' pickled green olive	Pickles (GO)	940
	Morrisons' pickled onion	Pickles (ON)	300
Spread	Marmite	Spread (MA)	859
	Morrisons' Strawberry jam	Spread (SJ)	661
Yoghurt	Morrisons' Fat free	Yoghurt (FF)	245
	Tesco's Greek	Yoghurt (GR)	759
	Morrisons' Nature	Yoghurt (NA)	419
	Alpro' Soya	Yoghurt (SO)	980

Table 4.2 Sensory attributes evaluated in this study

Order in questionnaire	Type of attribute	Attribute	Label
1st	Modality	Overall appearance	AP
2nd	Modality	Overall aroma	AR
3rd	Modality	Overall texture	TE
4th	Individual attribute	Sweetness	SW
5th	Individual attribute	Sourness	SO
6th	Modality	Overall taste	TA
7th	Individual attribute	Aftertaste	AF
8th	Individual attribute	Persistence	PE
9th	Overall perception	Overall acceptability	OA

4.2.3 Experimental design

Each panellist was provided with eight samples for evaluation in one session that lasted for around twenty minutes. To reduce the carryover effect and respondent burden, a combination of a Williams Design and incomplete block design was used (Patterson, 1951; Wakeling and MacFie, 1995; Williams, 1949), which was produced by R package crossdes version 1.1-1 (Sailer, 2013). The same 8 products were assigned to every 8 panellists in a crossover order. Table 4.3 shows an example of the basic design for a group of 8 panellists. Every product was evaluated 64 times in total.

Table 4.3 Example of the crossover design

Panellist	**Sample presenting order**							
	1st Sample	2nd Sample	3rd Sample	4th Sample	5th Sample	6th Sample	7th Sample	8th Sample
1	A	B	C	D	E	F	G	H
2	B	D	A	F	C	H	E	G
3	D	F	B	H	A	G	C	E
4	F	H	D	G	B	E	A	C
5	H	G	F	E	D	C	B	A
6	G	E	H	C	F	A	D	B
7	E	C	G	A	H	B	F	D
8	C	A	E	B	G	D	H	F

4.2.4 Administration

The experimental work was conducted in centrally located sensory lab (see figure 4.1 and 4.2) equipped with a Compusense 5 computer system (Compusense Inc., 2013). All samples were served at proper temperatures.

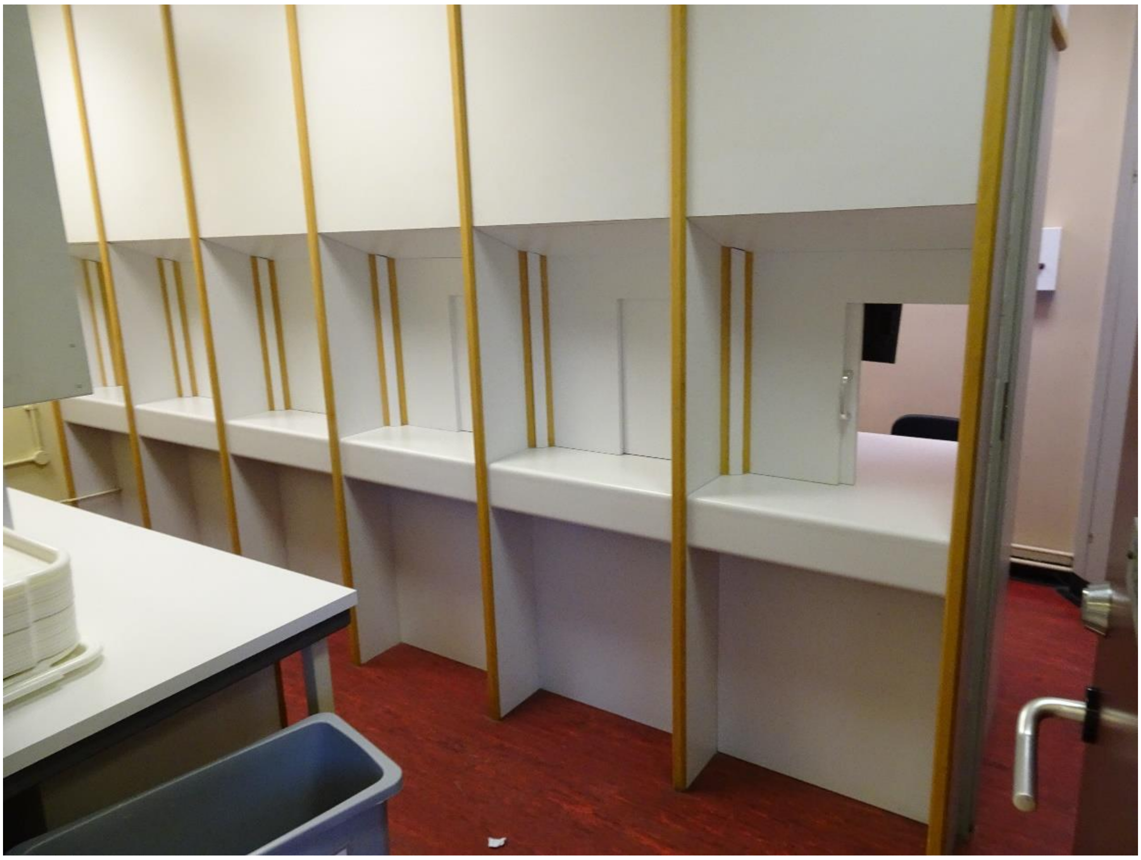


Figure 4.1 Sensory booth

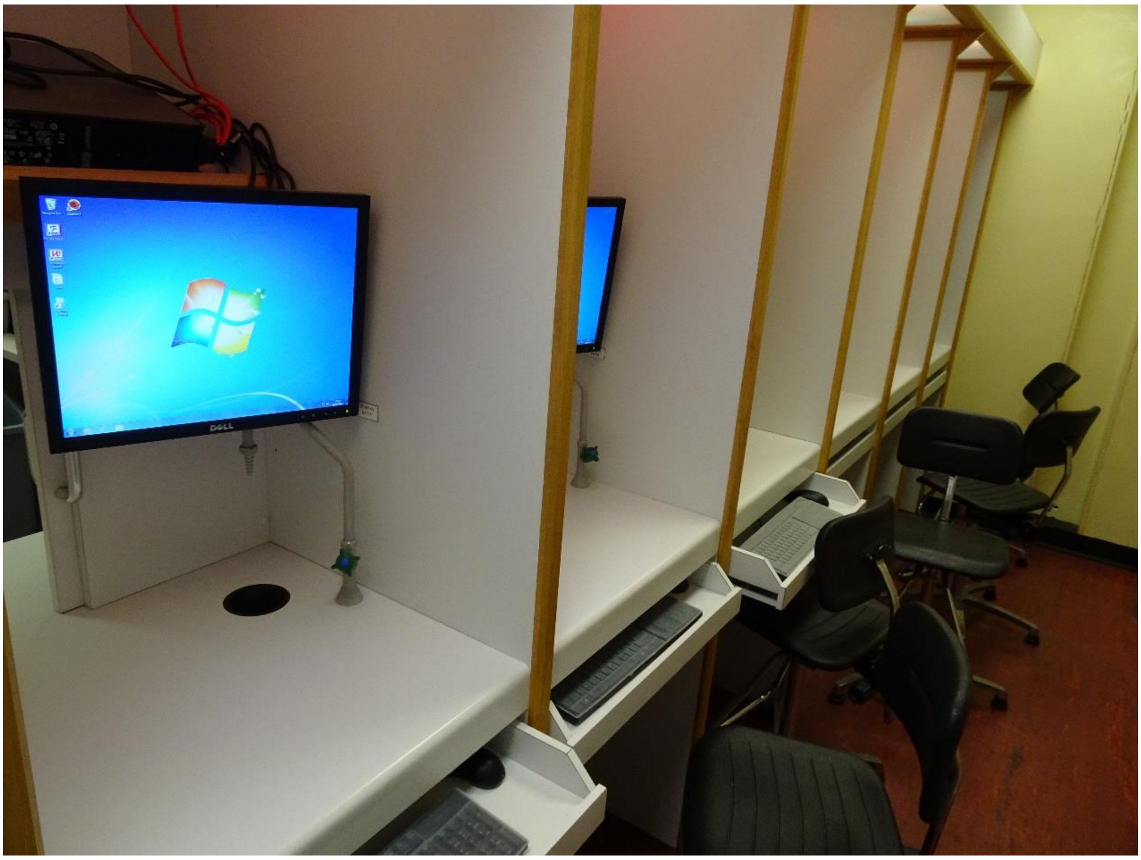


Figure 4.2 Sensory booth

4.3 Rasch analysis procedures

The raw data were split into two data sets, one that contained the ratings for the sensory modalities and attributes (labelled as Rasch ATTRIBUTES) and the other consisted of the ratings for the single overall acceptability variable (labelled as Rasch OA). The two data sets were fitted to the Many-Facet Rasch Rating scale model (MFR-RS model) using Facets (Linacre, 2014a) individually. The MFR-RS model consisted of three facets including the Panellist facet, the Product facet, and the Attribute or Overall acceptability facet. All facets were parametrised to positive orientation. A more detailed explanation of the MRF-RS model can be found in section 2.1.2.3. Several procedures described in section 2.3 were applied to examine the rating scale category effectiveness, unidimensionality, local item dependence, and fit.

4.3.1 Rating scale category effectiveness

The rating scale categories effectiveness was evaluated using the criteria outlined in section 2.3.1.2. Attempts at collapsing categories were tried out when the criteria could not be fulfilled.

4.3.2 Tests of unidimensionality and local item independence

The assumption of unidimensionality for the Attribute facet in the model using the ratings of sensory modalities and other attributes was evaluated by conducting a principal component analysis on standardised residuals (PCAR) followed by independent t-tests protocol. The procedures and the criteria of the tests illustrated in section 2.3.2.3 were followed. To do that, the Panellist facet and the Product facet were combined to form a single Panellist-Product facet at first. Then the data were fitted to WINSTEPS (Linacre, 2014d) using the two-facet Rating Scale model (Andrich, 1978a).

In addition, the local item independence was examined by checking the correlation between the residuals. The violation of the requirement of local item independence would be implied if the residual correlation between two items was equal to or greater than 0.30 (Smith, 2000). A decision on how to handle this issue would be made after further inspection on the meaning of the attribute items.

4.3.3 Global model fit

The global model fit was checked by inspecting the distribution of the standardised model residuals. According to Linacre (2014a), the global model fit of a MFR model would be satisfied if no more than 5% of the absolute standardised residuals were equal to greater than 2 and no more than 1% of that were equal to or greater than 3.

4.3.4 Test of individual fit

The individual fit was assessed according to the estimates of outfit MNSQ statistics. The detailed criteria can be found in section 2.3.3.1.

4.3.5 Wright maps

The estimates of all elements in three facets were visualised for initial inspection using Wright maps.

4.4 Statistical analysis procedures

4.4.1 Reliability

The Rasch reliability statistics of all facets were computed by Facets.

4.4.2 Chi-square tests for fixed effect and random effect

The results of chi-square tests for fixed effect and random effect were reported by Facets. As described in section 2.3.7, the chi-square statistic for fixed effect (namely the homogeneity index) is an indicator of whether the elements of a facet differed significantly, while the chi-square statistic for random effect is used for determining whether the elements of a facet were randomly sampled from a normally distributed population.

4.4.3 Modelling of the replicate measures

In order to obtain the estimates of the sixty-four replicate measures of each product, the procedure proposed by Ho (2019) was employed. In this procedure, the data were refitted to a modified MFR-RS model that consisted of Product-by-Panellist facet, Product facet and attribute/overall acceptability facet, while all elements in Product facet and the Rasch-Andrich thresholds of the rating scale were anchored at their estimated values obtained before the refitting. In addition, the group anchoring was made to the Product-by-Panellist facet, where the mean estimate of the 64 replicate measures of every product was anchored to the estimated measure of each product computed before the refitting.

4.4.4 Multiple comparisons

Firstly, both ANOVA and non-parametric Kruskal-Wallis test (Kruskal and Wallis, 1952) were applied to evaluate if there was a significant difference between the product overall liking. The comparison was conducted using the following types of data sets, individually:

(1) The composite measures of the overall liking modelled using the attribute ratings by Rasch analysis (labelled as Rasch ATTRIBUTES)

(2) The holistic measures of overall liking modelled using the single overall acceptability item by Rasch analysis (labelled as Rasch OA).

Secondly, the residual analysis was conducted to check the assumptions of ANOVA. The same procedure used in case study I was applied (see section 3.6.1). The Shapiro-Wilk test (Shapiro and Wilk, 1965) was used to determine the normality. If the assumption of normality could hold, then the Levene's test (Levene, 1960) would be applied for evaluating the homogeneity of variances, otherwise the Brown-Forsythe test (Brown and Forsythe, 1974) would be performed for the same purpose. In addition, the Bonferroni outlier test was conducted to check the extreme outliers.

Thirdly, if the ANOVA assumptions could hold, then the multiple comparisons would be conducted using Tukey HSD test (Tukey, 1949). Otherwise the non-parametric method would be applied using Dunn's test (Dunn, 1964) with Benjamini-Hochberg correction (Benjamini and Hochberg, 1995), depending on the results of Kruskal-Wallis test.

The ANOVA, Kruskal-Wallis test, Tukey HSD test and Shapiro-Wilk test were conducted using R program (R Core Team, 2018), while the Levene's test, Brown-Forsythe test and the Dunn's test with Benjamini-Hochberg correction were performed using R package car (Fox *et al.*, 2018), vGWAS (Shen, 2015), and PMCMRplus (Pohlert, 2018), respectively.

4.5 Results

4.5.1 Rating scale category effectiveness

The category statistics for the original 11-point hedonic scale are tabulated in table 4.4. The main findings are:

(1) The distribution of category frequency was skewed

Although the criterion of "*at least 10 observations of each category*" proposed by Linacre (2002a) were satisfied, it was noticeable that the categories represented low levels of liking were less frequently used by the panellists than the categories associated with high levels of liking in both models with attributes ratings and overall acceptability ratings. This means the sampled products were generally liked by the panellists. However, the skewed distribution of category frequency might result in disordered Rasch-Andrich thresholds because some of the categories were not observed frequently enough (Linacre, 2001).

(2) The fit of each category was acceptable

The outfit MSNQ statistics for each category were all around 1.0, indicating good fit.

(3) The mean measures of categories advanced monotonically

This indicated that the higher categories on the scale represented higher level of liking as they should be.

(4) Disordered Rasch-Andrich thresholds were observed

Disordered Andrich-thresholds were found around the centre of the scale, indicating that the central categories might only cover a narrow range of interval on the scale.

4.5.1.1 Optimising rating scale by collapsing categories

Attempts at collapsing categories were made in order to improve the effectiveness of scale categories. Two solutions (see table 4.5) that can solve the disordered Rasch-Andrich thresholds issues were found. They both required for the original 11-point scale to be collapsed to a 5-point scales. However only one of them fulfilled the recommendation of minimum advancing distance for Rasch-Andrich thresholds (Linacre, 2006) in both models with attribute ratings and overall acceptability ratings. This was selected as the final collapsing method. The details of the final collapsing method is demonstrated in table 4.6. Figure 4.3 and 4.4 illustrate the improvement of the Rasch-Andrich thresholds on the category probability plot. Further analyses were done on models based on the original rating scale and collapsed rating scale in parallel.

Table 4.4 Statistics for the original 11-point hedonic scale for different MFR-RS models

		Rasch ATTRIBUTES					Rasch OA				
Scale category	Raw Score	Counts[1]	Obs[2]	Exp[3]	Rasch-Andrich Threshold ± SE	Outfit MNSQ	Counts[1]	Obs[2]	Exp[3]	Rasch-Andrich Threshold ± SE	Outfit MNSQ
Dislike greatest imaginable	1	174 (1.52%)	-0.37	-0.39	NA	1.0	35 (2.28%)	-0.47	-0.50	NA	1.0
Dislike extremely	2	342 (2.98%)	-0.33	-0.31	-1.03±0.08	1.1	66 (4.30%)	-0.39	-0.39	-1.08±0.18	1.2
Dislike very much	3	541 (4.72%)	-0.25	-0.23	-0.73±0.05	1.0	97 (6.32%)	-0.34	-0.29	-0.73±0.11	0.8
Dislike moderately	4	661 (5.76%)	-0.15	-0.15	-0.39±0.04*	1.0	94 (6.12%)	-0.22	-0.19	-0.21±0.09*	0.8
Dislike slightly	5	1048 (9.14%)	-0.05	-0.06	-0.57±0.03*	1.0	126 (8.20%)	-0.08	-0.08	-0.43±0.08*	1.0
Neither like nor dislike	6	1106 (9.64%)	0.02	0.03	-0.07±0.03*	1.0	73 (4.75%)	0.03	0.02	0.51±0.07*	0.9
Like slightly	7	1770 (15.43%)	0.12	0.12	-0.40±0.02*	1.0	198 (12.89%)	0.15	0.12	-0.93±0.07*	1.0
Like moderately	8	2155 (18.79%)	0.23	0.22	-0.03±0.02	0.9	306 (19.92%)	0.25	0.23	-0.26±0.06*	0.8
Like very much	9	2266 (19.76%)	0.33	0.32	0.22±0.02	1.0	313 (20.38%)	0.37	0.35	0.27±0.06	1.0
Like extremely	10	1138 (9.92%)	0.42	0.42	1.06±0.03	1.0	179 (11.65%)	0.47	0.49	0.98±0.08	1.1
Like greatest imaginable	11	267 (2.23%)	0.47	0.54	1.93±0.06	1.1	49 (3.19%)	0.52	0.65	1.87±0.15	1.2

[1] The counts of each category. The percentages of the counts are displayed in brackets.

[2]. Modelled average measure in logits.

[3]. Expected average measure if data fitted the model.

* Disordered Rasch-Andrich thresholds.

Table 4.5 Effects of collapsing categories to Rasch-Andrich thresholds

	Original Rating scale	**Collapsing 2-3-4-5, 6-7-8, and 9-10**	[1]**Collapsing 2-3-4, 5-6-7, and 8-9-10**
Number of categories	11	5	5
Recommendation for minimum advances (Linacre, 2006)	0.36	0.81	0.81
Rasch ATTRIBUTES			
Rasch-Andrich Threshold	Disordered	Ordered	Ordered
Actual minimum advancing distance	-	1.65	1.24
Rasch OA			
Rasch-Andrich Threshold	Disordered	Ordered	Ordered
Actual minimum advancing distance	-	1.30	0.53

[1] This collapsing method was not adapted because the actual minimum advancing distance of Rasch-Andrich thresholds were less than the recommended value.

Table 4.6 Statistics for the collapsed 5-point rating scale categories for different MFR-RS models

		Rasch ATTRIBUTES					**Rasch OA**				
Scale category	Raw Score	Counts[1]	Obs[2]	Exp[3]	Rasch-Andrich Threshold ± SE	Outfit MNSQ	Counts[1]	Obs[2]	Exp[3]	Rasch-Andrich Threshold ± SE	Outfit MNSQ
Dislike greatest imaginable	1	174 (1.52%)	-0.98	-1.10	NA	1.0	35 (2.28%)	-1.34	-1.49	NA	1.1
Dislike *(Dislike extremely +* *Dislike very much +* *Dislike moderately +* *Dislike slightly)*	2	2592 (22.60%)	-0.55	-0.49	-3.50±0.08	1.0	383 (24.93%)	-0.77	-0.65	-3.45±0.18	0.9
Like somewhat *(Neither like nor dislike +* *Like slightly +* *Like moderately)*	3	5031 (43.87%)	0.15	0.11	-0.85±0.02	1.0	97 (37.57%)	0.20	0.10	-0.68±0.07	0.9
Like very much *(Like very much +* *Like extremely)*	4	3404 (29.68%)	0.70	0.70	0.80±0.02	1.0	94 (32.03%)	0.81	0.82	0.62±0.06	1.1
Like greatest imaginable	11	267 (2.23%)	0.47	0.54	1.93±0.06	1.1	49 (3.19%)	0.52	0.65	1.87±0.15	1.2

[1] The counts of each category. The percentages of the counts are shown in brackets.

[2]. Modelled average measure in logits.

[3]. Expected average measure if data fitted the model.

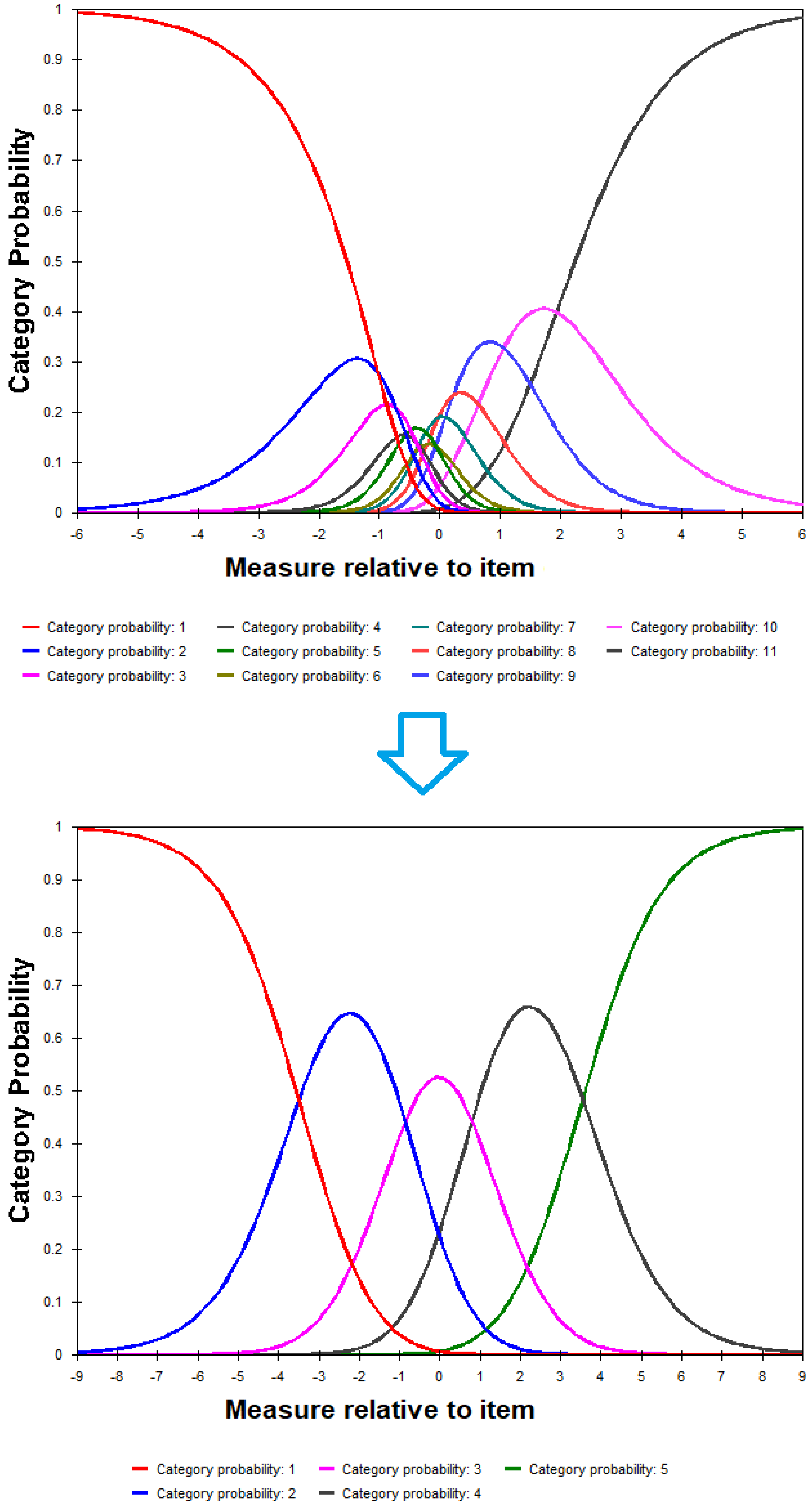


Figure 4.3 Category probability plot - Rasch ATTRIBUTES

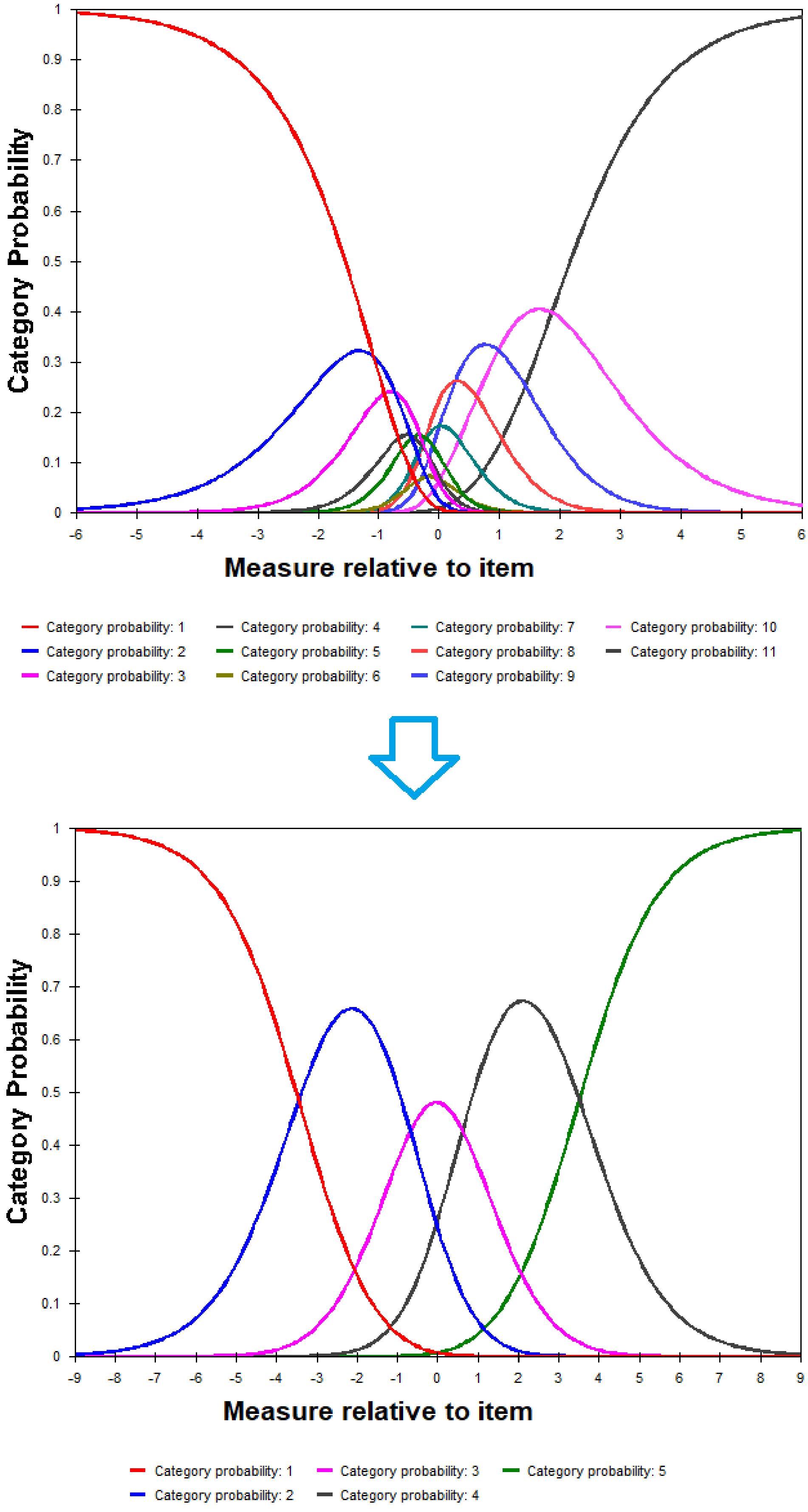


Figure 4.4 Category probability plot - Rasch OA

4.5.2 Tests of unidimensionality and local item independence of Attribute facet

Table 4.7 tabulates the results of the tests of unidimensionality of the Attribute facet for both models using original rating scale and collapsed rating scale. The results suggested that the assumption of unidimensionality could hold in both models. In addition, table 4.8 depicts the item pairs which might violate the assumption of local item independence.

4.5.2.1 Tests of unidimensionality - original rating scale

For the model using the original rating scale, both PCAR and t-tests suggested that the assumption of unidimensionality could hold in the Attribute facet. The following indicators were observed:

(1) Two misfitting items were observed when using the original rating scale. Their outfit MNSQ statistics were just above 1.50, which indicated that they were less productive but would not degrade the measurement (Wright and Linacre, 1994). This implied a potential violation of the assumption of unidimensionality, but it was not a decisive indicator.

(2) In the results of PCA on standardised residuals (PCAR) test, the eigenvalue of raw unexplained variance on the first PCAR contrast (2.24) was less than 3.00, indicating that the Attribute facet might be unidimensional because the strength of the potential additional dimension was below three items (Linacre, 2014c).

(3) The proportion of significant cases identified in the independent t-tests was 5.22%, which was slightly greater than the recommended cut-off point of 5% (Smith, 2002). However, the estimated lower bound of the binomial CI for the t-tests was only 4.18%, less than the 5% threshold (Tennant and Conaghan, 2007). This strongly suggested that the assumption of unidimensionality could hold.

(4) The disattenuated correlations between the three item clusters were quite high. All of them were greater than 0.71, above which implies that the clusters are more dependent than independent (Linacre, 2014c).

4.5.2.2 Tests of unidimensionality - collapsed rating scale

The tests of unidimensionality for the model using the collapsed rating scale obtained even more prominent results to support the assumption of unidimensionality in the Attribute facet:

(1) No misfitting item was found in the data with the collapsed rating scale.

(2) The eigenvalue of raw unexplained variance on the first PCAR contrast was reduced to 2.07 in the model with collapsed rating scale.

(3) The proportion of significant t-tests was decreased to 1.43%, while the estimate of lower bound of 95% binomial CI was only 0.92%.

(4) The disattenuated correlation between the item clusters were even larger than those estimated using the model with the original rating scale.

Table 4.7 Tests of unidimensionality

		Original scale	Collapsed scale
PCA on model standardised residuals			
Eigenvalue of raw unexplained variance in 1st PCAR contrast		2.24	2.07
Disattenuated correlations between item clusters in 1st PCAR contrast	1~3	0.73	0.75
	1~2	0.72	0.80
	2~3	0.98	1.00
Item fit[1]	Range	0.66-1.62	0.69-1.50
(Outfit MNSQ)	Misfitting item	AR (1.52), and AP (1.62)	NA
Independent t-tests			
Proportion of significant t statistics		5.22%	1.43%
Lower bound of 95% binomial CI		4.18%	0.92%

[1] The outfit MNSQ statistics used in this table were estimated with the two-facet model for the purpose of testing the assumption of unidimensionality, which were different with those estimated with the three-facet model.

4.5.2.3 Test of local item independence

The results implied that the attribute item aftertaste (AF) and persistence (PE) exhibited potential local item dependence (table 4.8), as the residual correlation between them reached 0.47 with original rating scale and 0.37 with collapsed rating scale, which were greater than the diagnostic boundary of 0.30 (Smith, 2000). Aftertaste and persistence were both rated after the product was swallowed, therefore they may share more common features than with the other attributes. However, no action was done to them because:

(1) The unidimensionality was not compromised.

(2) The aftertaste and persistence are associated with the panellists' perceptions of taste intensity and taste duration after swallowing the samples, respectively. They represent different sensory characteristics.

Table 4.8 The item pairs that exhibited potential local item dependence

Item 1	**Item 2**	**Residual correlation**[1]	
		Original rating scale	Collapsed rating scale
Aftertaste (AF)	Persistence (PE)	0.47	0.37

[1] Only the item pairs with a residual correlation greater than 0.30 are presented here.

4.5.3 Global model fit

Table 4.9 depicts the proportion of the extreme residuals that had an absolute value equal or greater than 2 or 3, respectively. The results satisfied the criteria proposed by Linacre (2014b), implying that the global model fit was accepted for all models. The proportions of extreme residuals in the model using the attribute ratings (Rasch ATTRIBUTES) were less than those in the model using single overall acceptability ratings (Rasch OA).

Table 4.9 Counts and proportion of extreme residuals

	Absolute standardised residuals ≥2		Absolute standardised residuals ≥3	
	Counts	Proportion	Counts	Proportion
Rasch ATTRIBUTES[1]				
Original scale	457	3.99%	73	0.64%
Collapsed scale	542	4.73%	36	0.31%
Rasch OA[2]				
Original scale	64	4.17%	12	0.78%
Collapsed scale	75	4.88%	8	0.52%

[1] For Rasch ATTRIBUTES, the number of total responses is 11468 (missing value=820)

[2] For Rasch OA, the number of total responses is 1536 (no missing value)

4.5.4 Test of individual fit

The individual fit of each element was evaluated according to the outfit MNSQ statistics.

4.5.4.1 Product fit

A couple of underfitting products were observed when fitting the attribute ratings to the model using the original rating scale and fitting the overall acceptability ratings to the model using both the collapsed and original rating scales (table 4.10). However, their outfit MNSQ statistics were only slightly above the recommended range of 0.5~1.5. They were all less than 2.0, beyond which the distortion of measurement could be indicated (Wright and Linacre, 1994). So the degree of underfit was not likely to have significant impact on the measurement. Moreover, they were the main targets for comparison in this study. Therefore they were all retained. In addition, no evidence of misfit issue was found when modelling the attribute ratings with collapsed rating scale.

Table 4.10 Overview of product fit for all models

Product	**Rasch ATTRIBUTES**		**Rasch OA**	
	Original rating scale	Collapsed rating scale	Original rating scale	Collapsed rating scale
Outfit MNSQ (range)	0.72-1.56	0.71-1.38	0.54-1.54	0.55-1.63
Misfitting product	Chocolate (DA): 1.56	NA	Chocolate (DA): 1.54 Pickle (GO): 1.54	Pickle (GO): 1.63

4.5.4.2 Attribute fit

No misfit issue was found in Attribute facet, where the range of the outfit MNSQ statistics were similar between the estimates based on the original and collapsed rating scales (table 4.11).

Table 4.11 Measure, standard error (SE) and outfit MNSQ statistics of Attribute facet for all models

Attributes	Original rating scale		Collapsed rating scale	
	Measure±SE	Outfit MNSQ	Measure±SE	Outfit MNSQ
AP	0.12±0.01	0.95	0.25±0.04	0.94
TE	0.11±0.01	1.03	0.22±0.04	1.01
AR	0.09±0.01	1.00	0.23±0.04	1.01
TA	0.01±0.01	1.16	0.04±0.04	0.97
SW	0.00±0.02	1.03	0.05±0.04	0.96
SO	-0.10±0.02	1.00	-0.18±0.04	1.13
AF	-0.12±0.01	0.98	-0.30±0.04	1.01
PE	-0.12±0.01	0.92	-0.31±0.04	0.93

4.5.4.3 Panellist fit

Table 4.12 shows the numbers of misfitting panellists and their proportion in the whole sample. Firstly, several panellists were diagnosed as overfit panellists (outfit MNSQ<0.5). According to Linacre (2002b), overfit persons in a measurement do not degrade measurement system, although they may yield to a raise of reliability misleadingly. Secondly, around 11% of the panellists were slightly underfit with outfit MNSQ statistics between 1.5 and 2.0. They did not deteriorate the measurement, even if they were less productive for it. Thirdly, a few panellists (4.69%~8.33% for different models) with serious underfit issue were identified (outfit MNSQ>2.0). However, compared to the Product facet and Attribute facet, the misfitting panellists may be less an issue due to the relatively large quantity of them compared to the items (Wright and Linacre, 1994). Since their proportion was quite small across all models, and the fits of the products and the attributes were all acceptable, none of them were dropped. Lastly, the proportions of misfitting panellists modelled using attributes ratings were less than those estimated based on overall acceptability ratings. The scale collapsing also slightly improved the fit statistics.

Table 4.12 Counts and proportion of misfitting panellists for all models

Panellist[1]	Rasch ATTRIBUTES		Rasch OA	
	Original rating scale	Collapsed rating scale	Original rating scale	Collapsed rating scale
Overfit				
<0.5	35 (18.23%)	18 (10.94%)	53 (27.60%)	46 (25.00%)
Underfit				
1.5~2.0	21 (10.94%)	19 (9.90%)	22 (11.46%)	22 (11.46%)
>2	10 (5.21%)	9 (4.69%)	16 (8.33%)	12 (6.25%)

[1] Outfit MNSQ statistics

4.5.5 Wright maps

The distributions of the elements for each facet on the scale were visualised using Wright maps (figure 4.5~4.8), on which all elements were presented in a positive direction according to the measures (from bottom to top). The following could be observed from the Wright maps:

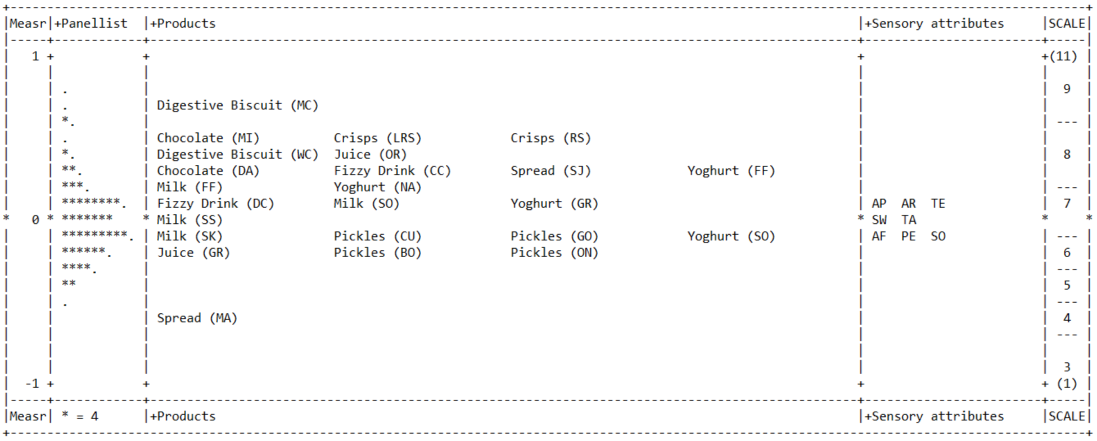
Firstly, the hierarchical rank of the products on the scale (figure 4.5~4.8) can be seen on the Wright maps, which were similar among all four models. The higher the position of the product on the scale, the more likely that it was liked by the panellists. The digestive biscuit (MC) was the most liked product, followed by the two crisps products, whereas the spread (MA) was the least liked product[1].

Secondly, for the composite measurement using attribute ratings (figure 4.5 and 4.6), the contributions of the attributes to the composite overall liking measure were different. The attributes AP (appearance), AR (aroma) and TE (texture and mouthfeel) were most liked attributes by the panellists, in contrast with the attributes AF (aftertaste) and PE (persistent), which were the least liked attributes.

Thirdly, the locations of the panellists showed their tendencies to the use of rating scale categories. Those at top of the scale (figure 4.5~4.8) were the most lenient panellists who tend to give relatively high ratings on the same product compare to the others, while the panellists at the bottom of the scale were more likely to use the categories that represent low levels of liking.

Lastly, the distributions of the panellists and products seemed to follow the normal distribution, therefore when inspecting the Rasch reliability statistics, separation is more suitable than strata for describing the degree of their dispersion on the scale, as there were not many extreme outliers.

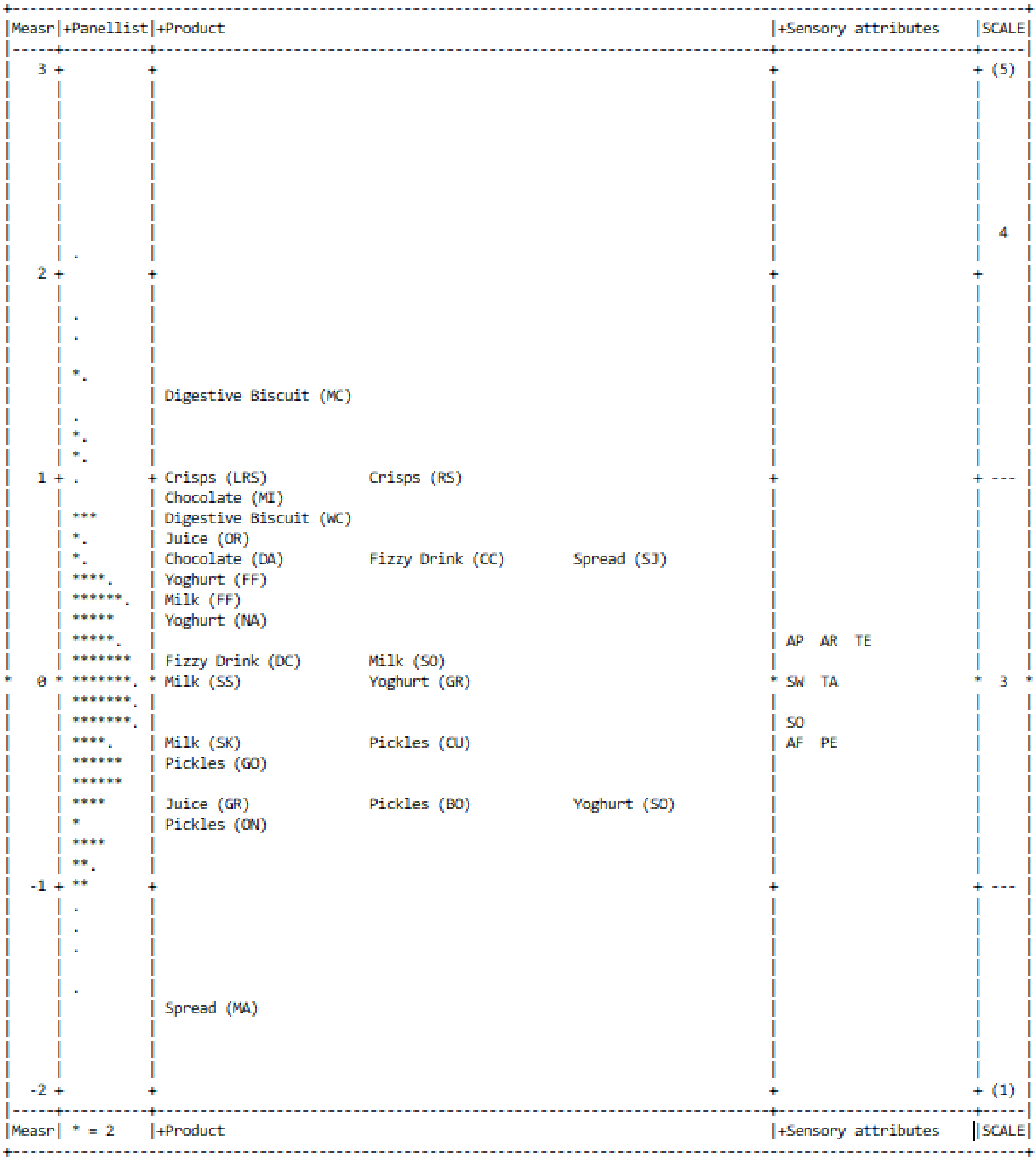
[1] The Digestive biscuit (MC) was the McVitie’s product, while the Spread (MA) was marmite. The full list of the product names are shown in table 4.2.



```
+----------------------------------------------------------------------------------------------------------------------------------------------------+
|Measr|+Panellist  |+Products                                                                                         |+Sensory attributes     |SCALE|
|-----+------------+--------------------------------------------------------------------------------------------------+------------------------+-----|
|   1 +            +                                                                                                  +                        +(11) |
|     |            |                                                                                                  |                        |     |
|     | .          |                                                                                                  |                        |  9  |
|     | .          | Digestive Biscuit (MC)                                                                           |                        |     |
|     | *.         |                                                                                                  |                        | --- |
|     | .          | Chocolate (MI)          Crisps (LRS)            Crisps (RS)                                      |                        |     |
|     | *.         | Digestive Biscuit (WC)  Juice (OR)                                                               |                        |  8  |
|     | **.        | Chocolate (DA)          Fizzy Drink (CC)        Spread (SJ)             Yoghurt (FF)             |                        |     |
|     | ***.       | Milk (FF)               Yoghurt (NA)                                                             |                        | --- |
|     | ********.  | Fizzy Drink (DC)        Milk (SO)               Yoghurt (GR)                                     | AP  AR  TE             |  7  |
*   0 * *******    * Milk (SS)                                                                                        * SW  TA                 *     *
|     | *********. | Milk (SK)               Pickles (CU)            Pickles (GO)            Yoghurt (SO)             | AF  PE  SO             | --- |
|     | ******.    | Juice (GR)              Pickles (BO)            Pickles (ON)                                     |                        |  6  |
|     | ****.      |                                                                                                  |                        | --- |
|     | **         |                                                                                                  |                        |  5  |
|     | .          |                                                                                                  |                        | --- |
|     |            | Spread (MA)                                                                                      |                        |  4  |
|     |            |                                                                                                  |                        | --- |
|     |            |                                                                                                  |                        |     |
|     |            |                                                                                                  |                        |  3  |
|  -1 +            +                                                                                                  +                        + (1) |
|-----+------------+--------------------------------------------------------------------------------------------------+------------------------+-----|
|Measr| * = 4      |+Products                                                                                         |+Sensory attributes     |SCALE|
+----------------------------------------------------------------------------------------------------------------------------------------------------+
```

Figure 4.5 Wright maps for Rasch ATTRIBUTES with original rating scale

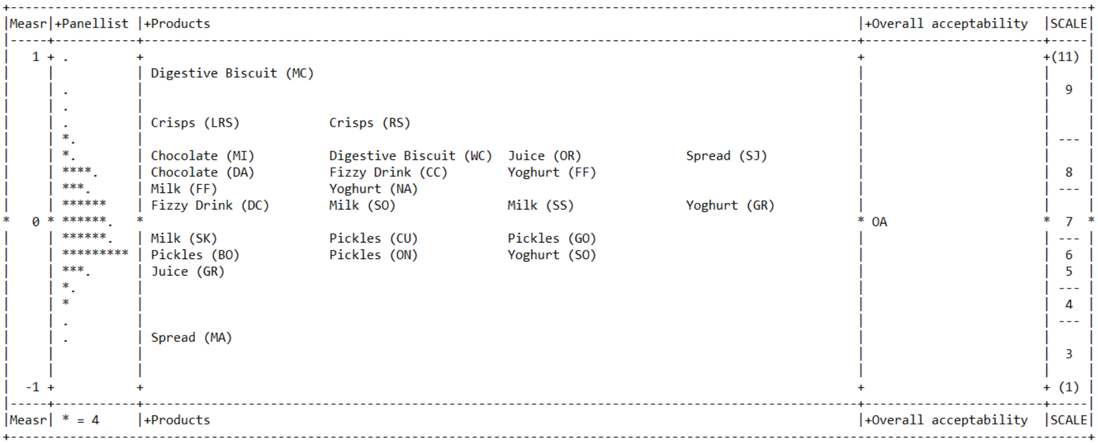
The names of the products and attributes are listed in table 4.1 and 4.2, respectively.



```
+-----------------------------------------------------------------------------------------------------------------------+
|Measr|+Panellist|+Product                                                                 |+Sensory attributes    |SCALE|
|-----+----------+------------------------------------------------------------------------+-----------------------+-----|
|   3 +          +                                                                        +                       + (5) |
|     |          |                                                                        |                       |     |
|     |          |                                                                        |                       |     |
|     |          |                                                                        |                       |     |
|     |          |                                                                        |                       |     |
|     |          |                                                                        |                       |     |
|     |          |                                                                        |                       |     |
|     |          |                                                                        |                       |     |
|     |          |                                                                        |                       |  4  |
|     | .        |                                                                        |                       |     |
|   2 +          +                                                                        +                       +     |
|     |          |                                                                        |                       |     |
|     | .        |                                                                        |                       |     |
|     | .        |                                                                        |                       |     |
|     |          |                                                                        |                       |     |
|     | *.       |                                                                        |                       |     |
|     |          | Digestive Biscuit (MC)                                                 |                       |     |
|     | .        |                                                                        |                       |     |
|     | *.       |                                                                        |                       |     |
|     | *.       |                                                                        |                       |     |
|   1 + .        + Crisps (LRS)            Crisps (RS)                                    +                       + --- |
|     |          | Chocolate (MI)                                                         |                       |     |
|     | ***      | Digestive Biscuit (WC)                                                 |                       |     |
|     | *.       | Juice (OR)                                                             |                       |     |
|     | *.       | Chocolate (DA)          Fizzy Drink (CC)        Spread (SJ)            |                       |     |
|     | ****.    | Yoghurt (FF)                                                           |                       |     |
|     | ******.  | Milk (FF)                                                              |                       |     |
|     | *****    | Yoghurt (NA)                                                           |                       |     |
|     | *****.   |                                                                        | AP  AR  TE            |     |
|     | *******  | Fizzy Drink (DC)        Milk (SO)                                      |                       |     |
*   0 * *******. * Milk (SS)               Yoghurt (GR)                                   * SW  TA                *  3  *
|     | *******. |                                                                        |                       |     |
|     | *******. |                                                                        | SO                    |     |
|     | ****.    | Milk (SK)               Pickles (CU)                                   | AF  PE                |     |
|     | ******   | Pickles (GO)                                                           |                       |     |
|     | ******   |                                                                        |                       |     |
|     | ****     | Juice (GR)              Pickles (BO)            Yoghurt (SO)           |                       |     |
|     | *        | Pickles (ON)                                                           |                       |     |
|     | ****     |                                                                        |                       |     |
|     | **.      |                                                                        |                       |     |
|  -1 + **       +                                                                        +                       + --- |
|     | .        |                                                                        |                       |     |
|     | .        |                                                                        |                       |     |
|     | .        |                                                                        |                       |     |
|     |          |                                                                        |                       |     |
|     | .        |                                                                        |                       |     |
|     |          | Spread (MA)                                                            |                       |     |
|     |          |                                                                        |                       |     |
|     |          |                                                                        |                       |     |
|     |          |                                                                        |                       |     |
|  -2 +          +                                                                        +                       + (1) |
|-----+----------+------------------------------------------------------------------------+-----------------------+-----|
|Measr| * = 2    |+Product                                                                |+Sensory attributes    |SCALE|
+-----------------------------------------------------------------------------------------------------------------------+
```

Figure 4.6 Wright maps for Rasch ATTRIBUTES with collapsed rating scale

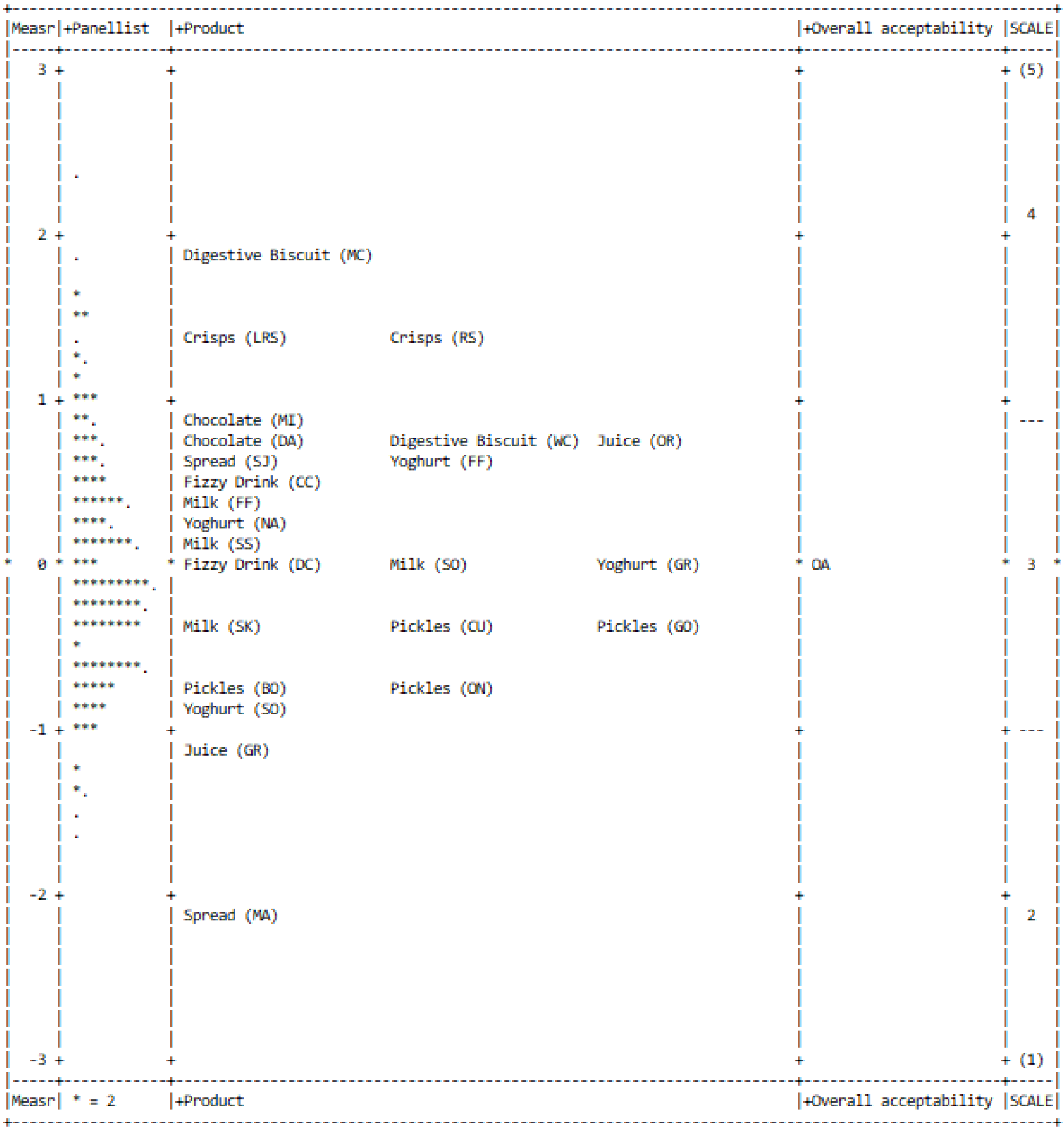
The names of the products and attributes are listed in table 4.1 and 4.2, respectively.



```
+----------------------------------------------------------------------------------------------------------------------------------------------------+
|Measr|+Panellist |+Products                                                                                        |+Overall acceptability  |SCALE|
|-----+-----------+-------------------------------------------------------------------------------------------------+------------------------+-----|
|   1 + .         +                                                                                                 +                        +(11) |
|     |           | Digestive Biscuit (MC)                                                                          |                        |     |
|     | .         |                                                                                                 |                        |  9  |
|     | .         |                                                                                                 |                        |     |
|     | .         | Crisps (LRS)            Crisps (RS)                                                             |                        |     |
|     | *.        |                                                                                                 |                        | --- |
|     | *.        | Chocolate (MI)          Digestive Biscuit (WC)  Juice (OR)              Spread (SJ)             |                        |     |
|     | ****.     | Chocolate (DA)          Fizzy Drink (CC)        Yoghurt (FF)                                    |                        |  8  |
|     | ***.      | Milk (FF)               Yoghurt (NA)                                                            |                        | --- |
|     | ******    | Fizzy Drink (DC)        Milk (SO)               Milk (SS)               Yoghurt (GR)            |                        |     |
*   0 * ******.   *                                                                                                 * OA                     *  7  *
|     | ******.   | Milk (SK)               Pickles (CU)            Pickles (GO)                                    |                        | --- |
|     | ********* | Pickles (BO)            Pickles (ON)            Yoghurt (SO)                                    |                        |  6  |
|     | ***.      | Juice (GR)                                                                                      |                        |  5  |
|     | *.        |                                                                                                 |                        | --- |
|     | *         |                                                                                                 |                        |  4  |
|     | .         |                                                                                                 |                        | --- |
|     | .         | Spread (MA)                                                                                     |                        |     |
|     |           |                                                                                                 |                        |  3  |
|     |           |                                                                                                 |                        |     |
|  -1 +           +                                                                                                 +                        + (1) |
|-----+-----------+-------------------------------------------------------------------------------------------------+------------------------+-----|
|Measr| * = 4     |+Products                                                                                        |+Overall acceptability  |SCALE|
+----------------------------------------------------------------------------------------------------------------------------------------------------+
```

Figure 4.7 Wright maps for Rasch OA with original rating scale

OA = overall acceptability; the names of the products can be seen in table 4.1.



```
+--------------------------------------------------------------------------------------------------------------------------+
|Measr|+Panellist  |+Product                                                                 |+Overall acceptability |SCALE|
|-----+------------+-------------------------------------------------------------------------+-----------------------+-----|
|   3 +            +                                                                         +                       + (5) |
|     |            |                                                                         |                       |     |
|     |            |                                                                         |                       |     |
|     |            |                                                                         |                       |     |
|     |            |                                                                         |                       |     |
|     | .          |                                                                         |                       |     |
|     |            |                                                                         |                       |     |
|     |            |                                                                         |                       |  4  |
|   2 +            +                                                                         +                       +     |
|     | .          | Digestive Biscuit (MC)                                                  |                       |     |
|     |            |                                                                         |                       |     |
|     | *          |                                                                         |                       |     |
|     | **         |                                                                         |                       |     |
|     | .          | Crisps (LRS)            Crisps (RS)                                     |                       |     |
|     | *.         |                                                                         |                       |     |
|     | *          |                                                                         |                       |     |
|   1 + ***        +                                                                         +                       +     |
|     | **.        | Chocolate (MI)                                                          |                       | --- |
|     | ***.       | Chocolate (DA)          Digestive Biscuit (WC)  Juice (OR)              |                       |     |
|     | ***.       | Spread (SJ)             Yoghurt (FF)                                    |                       |     |
|     | ****       | Fizzy Drink (CC)                                                        |                       |     |
|     | ******.    | Milk (FF)                                                               |                       |     |
|     | ****.      | Yoghurt (NA)                                                            |                       |     |
|     | *******.   | Milk (SS)                                                               |                       |     |
*   0 * ***        * Fizzy Drink (DC)        Milk (SO)               Yoghurt (GR)            * OA                    *  3  *
|     | **********.|                                                                         |                       |     |
|     | ********.  |                                                                         |                       |     |
|     | ********   | Milk (SK)               Pickles (CU)            Pickles (GO)            |                       |     |
|     | *          |                                                                         |                       |     |
|     | ********.  |                                                                         |                       |     |
|     | *****      | Pickles (BO)            Pickles (ON)                                    |                       |     |
|     | ****       | Yoghurt (SO)                                                            |                       |     |
|  -1 + ***        +                                                                         +                       + --- |
|     |            | Juice (GR)                                                              |                       |     |
|     | *          |                                                                         |                       |     |
|     | *.         |                                                                         |                       |     |
|     | .          |                                                                         |                       |     |
|     | .          |                                                                         |                       |     |
|     |            |                                                                         |                       |     |
|     |            |                                                                         |                       |     |
|  -2 +            +                                                                         +                       +     |
|     |            | Spread (MA)                                                             |                       |  2  |
|     |            |                                                                         |                       |     |
|     |            |                                                                         |                       |     |
|     |            |                                                                         |                       |     |
|     |            |                                                                         |                       |     |
|     |            |                                                                         |                       |     |
|     |            |                                                                         |                       |     |
|  -3 +            +                                                                         +                       + (1) |
|-----+------------+-------------------------------------------------------------------------+-----------------------+-----|
|Measr| * = 2      |+Product                                                                 |+Overall acceptability |SCALE|
+--------------------------------------------------------------------------------------------------------------------------+
```

Figure 4.8 Wright maps for Rasch OA with collapsed rating scale

OA = overall acceptability; the names of the products can be seen in table 4.1.

4.5.6 Rasch separation and reliability statistics

Table 4.13 reports the Rasch reliability statistics.

The estimates of panellist separation and reliability in the two Rasch ATTRIBUTES models were acceptable, suggesting that the panellists were spread out for 2~3 statistically distinct levels along the scale. However, the values of reliability statistics for the Panellist facet modelled using the single overall acceptability item was below the recommended value of 1.5 for separation and 0.69 for separation reliability (Tennant and Conaghan, 2007) using both the original and collapsed rating scales, suggesting the panellists cannot be differentiated by this measure.

The separation and reliability of Product facet and Attribute facet were satisfied, suggesting several statistical distinct levels can be identified among the products or attributes.

Table 4.13 The estimates of Rasch reliability indices for all models

Facet	Statistic	Rasch ATTRIBUTES		Rasch OA	
		Original rating scale	Collapsed rating scale	Original rating scale	Collapsed rating scale
Panellist	Separation	2.90	2.76	0.83	0.74
	Strata	4.20	4.01	1.44	1.32
	Reliability	0.89	0.88	0.41	0.36
Product	Separation	11.12	10.16	4.98	4.66
	Strata	15.15	13.88	6.98	6.55
	Reliability	0.99	0.99	0.96	0.96
Attribute	Separation	7.05	5.96	-	-
	Strata	9.73	8.28	-	-
	Reliability	0.98	0.97	-	-

4.5.7 Chi-square statistics on fixed effect and random effect

The chi-square statistics on fixed effect and random effect were computed.

Table 4.14 depicts the results for the fixed effect. The chi-square statistics were significant for all facets, suggesting that the elements were not all the same in each facet. Thus the multiple comparisons could be applied to differentiate them.

Table 4.15 shows the chi-square statistics that tested the hypothesis of whether the data can be thought as random samples from a normally distributed population. The non-significance in the statistics for all facets implied that the data sampling was adequate.

Table 4.14 Chi-square statistics on fixed effect

	Rasch ATTRIBUTES		**Rasch OA**	
	Original rating scale	Collapsed rating scale	Original rating scale	Collapsed rating scale
Panellist (d.f. 191)				
Chi-square	1937.1	1872.1	353.8	346.3
P value	0.00	0.00	0.00	0.00
Product (d.f. 23)				
Chi-square	3080.0	2458.3	596.5	512.0
P value	0.00	0.00	0.00	0.00
Attribute (d.f. 7)			-	-
Chi-square	375.6	271.1	-	-
P value	0.00	0.00	-	-

Table 4.15 Chi-square statistics on random effect

	Rasch ATTRIBUTES		**Rasch OA**	
	Original rating scale	Collapsed rating scale	Original rating scale	Collapsed rating scale
Panellist (d.f. 190)				
Chi-square	172.4	173.7	121.3	134.1
P value	0.82	0.80	1.00	1.00
Product (d.f. 22)				
Chi-square	22.8	22.8	22.0	22.0
P value	0.41	0.41	0.46	0.46
Attribute (d.f. 6)			-	-
Chi-square	6.9	6.8	-	-
P value	0.33	0.34	-	-

4.5.8 Multiple comparisons on the overall liking of the products

The replicate measures of the product were estimated after refitting data to the modified model. They were used for multiple comparisons. Table 4.16 depicts the results of ANOVA, Kruskal-Wallis test and the residual analysis for ANOVA assumption.

The key findings are:

(1) Both ANOVA and Kruskal-Wallis test suggested that there were significant difference between the products on their overall liking for all models. The results were consistent with the chi-square test on fixed effects (section 4.5.7).

(2) None of the ANOVA assumption could hold with any data set. Therefore the non-parametric Dunn's test with Benjamini-Hochberg correction was applied to compare the overall liking of the products estimated in different methods. The results of the multiple comparisons can be seen in table 4.17.

(3) The multiple comparisons could discern the products into more statistically different groups with the measures modelled using overall acceptability ratings (Rasch OA). However, more overlaps between the groups were observed with Rasch OA.

(4) Slightly more statistical groups were obtained based on the Rasch measures estimated with the collapsed rating scale than with original scale.

Table 4.16 Results of ANOVA, Kruskal-Wallis test, and the residual analysis for ANOVA assumption based on the estimates of refitted model

	Rasch ATTRIBUTES		**Rasch OA**	
Test	**Original scale**	**Collapsed scale**	**Original scale**	**Collapsed scale**
ANOVA	<0.001	<0.001	<0.001	0.008
Kruskal-Wallis test	<0.001	<0.001	<0.001	<0.001
Residual analysis				
Shapiro-Wilk test	<0.001	<0.001	<0.001	<0.001
Brown-Forsythe test	<0.001	<0.001	<0.001	<0.001
Bonferroni outlier test	10 extreme outlier	3 extreme outlier	No extreme outlier	No extreme outlier

Table 4.17 Multiple comparisons on the overall liking of the products

Product[1]	**Rasch ATTRIBUTES**		**Rasch OA**	
	Original	Collapsed	Original	Collapsed
Digestive Biscuit (MC)	a	a	a	a
Chocolate (MI)	ab	ab	bcdef	ab
Crisps (RS)	abc	ab	abc	ab
Crisps (LRS)	ab	ab	ab	abc
Digestive Biscuit (WC)	bc	bc	bcdef	bcd
Spread (SJ)	bc	bc	abc	abc
Juice (OR)	bc	bcd	bcd	defgh
Fizzy Drink (CC)	bcde	bcd	cdefg	ab
Chocolate (DA)	bcde	bcde	cdefgh	cdef
Yoghurt (FF)	bc	bcde	bcde	bcde
Milk (FF)	cde	cdef	defgh	cdefg
Yoghurt (NA)	bcde	cdef	cdefg	hijk
Milk (SO)	def	defg	fghi	ghijk
Fizzy Drink (DC)	cde	efg	cdefg	defgh
Yoghurt (GR)	def	efgh	efghi	jkl
Milk (SS)	efg	fgh	fghi	ijk
Pickles (CU)	fgh	ghi	ghij	efgh
Pickles (GO)	fgh	ghi	ghij	fghi
Milk (SK)	fgh	ghi	ghij	hijk
Pickles (BO)	fgh	hi	hij	fghi
Yoghurt (SO)	gh	i	ij	fghij
Pickles (ON)	h	i	ij	l
Juice (GR)	hi	i	jk	kl
Spread (MA)	i	j	k	m

[1] The full name of the products can be seen in table 4.1.

a~m represent statistically different groups

4.6 Discussion

4.6.1 Comparison between the composite measurement using attribute ratings and the individual measurement using the overall acceptability ratings

4.6.1.1 Model fit

The global model fit and individual model fit were both better with Rasch ATTRIBUTES than Rasch OA. Compared to the composite measure modelled using multiple items, the measure modelled by a single abstract item may suffer from more serious impact brought by random measurement error and item bias that may lead to unexpected responses that can deteriorate the model fit.

4.6.1.2 Separation/Reliability

The separation statistics of the Panellist facet estimated using the overall acceptability ratings in the model were less than 1.00 (equivalent to the separation reliability of 0.50) with both collapsed and original rating scale, which implied that the panellists cannot be divided into statistical distinct groups on the scale. I contrast, the separation statistics increased to nearly 3.00 when modelling the attribute ratings, suggesting that the panellists were spread out over almost three statistical levels on the scale. The attribute ratings provided more information that can be used for evaluating person's response pattern than the single overall acceptability ratings, thereby providing greater power for discriminating people. In this study, the poor reliability estimated using overall acceptability rating was not a serious issue because the aim of the study was to compare the preference of products according to panellists' overall liking on the them. In fact, for the sensory study involving the trained panels, low person separation/reliability is desired because the trained panels are required to give interchangeable ratings. But for other research that focuses on segmenting persons, it is critical to have relatively high separation/reliability. Using multiple attribute items instead of the single overall acceptability item can help with that purpose.

The same trend could be observed when comparing the reliability statistics of Product facet estimated by the model using attribute ratings and single overall acceptability ratings. The estimates of product separation modelled using attribute ratings were more than twice those modelled using the single overall acceptability ratings, which meant that the difference of the overall liking between the products can be better distinguished by the composite measurement. This again exhibited the necessity of modelling panellists' overall liking over the products using the attribute ratings.

4.6.1.3 Determination of the difference in overall liking of products

Although more statistically different groups can be discerned using the estimates of Rasch OA than Rasch ATTRIBUTES, it did not mean that the former had greater power on discriminating products. On the contrary, the less overlap between the groups with Rasch ATTRIBUTES indicated that difference between products can be better distinguished when modelling the attribute ratings. Furthermore, the product separation statistics were more than doubled when using attribute ratings instead of overall acceptability ratings, implying that the products were dispersed in a broader range on the scale in the unit of standard error with attribute ratings in the model, which also suggested that the estimating the overall liking of product using attribute ratings has greater power to discriminate the products than using single overall acceptability ratings.

4.6.2 Effect of collapsing rating scale

The global model fit and individual fit were better with the collapsed rating scale, indicating that one can construct more meaningful measures after improving the rating scale category effectiveness.

However, the reliability statistics were slightly reduced after collapsing categories, although the decrease in reliability was small enough to be ignored. This finding is consistent with the first case study of this research and the research reported by Ho (2019).

4.6.3 Conclusion

In conclusion, a composite measure of consumers' overall liking on food product can be modelled using multiple sensory attribute ratings by Rasch analysis, which has greater power to differentiate between the overall liking of the products than using the holistic measure modelled using a single overall acceptability item. Moreover, the composite instrument fit the model better than the single item instrument, suggesting that the composite measurement can produce more precise results than the individual measurement.

Therefore, the composite measurement should be applied to sensory liking study. Since the overall liking measure can be modelled using the ratings of specific attributes, there is no need to employ the single overall acceptability item. In addition, because Rasch analysis is an individual-based measurement, the information related to each attribute can be examined and compared, which can help the researchers to better understanding the products.

Chapter 5 Case study III: Application of Rasch analysis in instrument development and validation for consumer insights research

5.1 Introduction

5.1.1 Aim

The case study I (chapter 3) had illustrated an application of using Rasch analysis for evaluating the psychometric properties of an existing consumer research instrument. The results indicated that Rasch analysis could overcome some limitations of CTT such as using discrete ordinal raw scores, excessive number of factors being extracted, and the unrealistic assumptions which cannot hold (*e.g.* one standard error of measurement apply to all scores). Beyond that, this study further explores the application of Rasch analysis in consumer research, focusing on the measurement development under the guidance of Rasch analysis.

5.1.2 Rasch analysis and measurement development

Rasch analysis not only provides a statistical solution that can break through the limits of CTT in data analysis, but also offers a framework to guide the measurement development. The benefits of using Rasch analysis for measurement development in consumer research practice include but not restricted to the following points:

5.1.2.1 Rasch analysis obtains a fundamental basis for construct conceptualisation.

Before measuring anything, one should have a clear idea about what to measure. Therefore, the development of measurement would start from defining the construct, which was considered as the first building block of measurement (Wilson, 2004). Rasch analysis could guide this process.

Firstly, Rasch analysis has been proved to be a special case of additive conjoint measurement in mathematics (Perline *et al.*, 1979). The persons and items are calibrated on the same continuum of the underlying construct, so that they can be measured conjointly. Thus, when defining a construct, one should take all facets of the measurement and the relationship between them into consideration.

Secondly, Rasch analysis requires the persons and items and any other elements of additional facet following a hierarchical pattern on the construct. This should also be reflected in the definition of the construct.

Thirdly, Rasch analysis utilises a tool, which is the Wright map, to assist the conceptualisation of the construct. When defining a construct, one can conceptualise it using a Wright map, which accommodates all measurement elements in predicted patterns.

5.1.2.2 Rasch analysis produces test-free person measures and sample-free item measures at individual levels

The parameters of Rasch model are estimated separately so that their influence on each other could be eliminated. On one hand, the estimates of persons do not rely on a particular item. On the other hand, the item measures are independent of the sampled population. Therefore invariant measures could be obtained. Moreover, unlike CTT where the raw scores of items were merged together to obtain the summated scores for measuring people, Rasch analysis attempts to explain the meanings of different levels of the construct using the estimates of individual persons and items. The errors of measurement are estimated at individual levels. If data fit to Rasch model, then the true locations of persons and items on the continuum of conceptual construct could be estimated. The person measures can be used for classifying consumers into different segments by performing a cluster analysis, while the hierarchy order of item measures can be indicative for the ideation task in new food product development practice.

5.1.3 Research object – Ready meal

Ready meal was selected as the research object in this study.

5.1.3.1 The definition of ready meal

A few definitions of ready meal had been proposed in the literature (Ahlgren *et al.*, 2004; Ahlgren *et al.*, 2005; Costa *et al.*, 2001; Van der Horst *et al.*, 2011).

These definitions were adapted together to form a new definition with the consideration of individual stages of food consumption[1]:

A **ready meal** can be defined as a pre-packaged meal sold in the supermarkets, convenience stores, or coffee shops, consisting of two or more components. It can be stored in chilled, frozen or ambient form. It has been prepared by a manufacturer or in-store so that no further pre-cooking preparation is needed. It should normally be heated[2] by either a microwave oven, or a conventional oven, or any other appropriate method prior to serving. It requires a minimum degree of disposal and/or clean-up.

5.1.3.2 Previous research on ready meal consumption

A number of previous studies (Ahlgren *et al.*, 2004; Ahlgren *et al.*, 2005; Geeroms *et al.*, 2008; Mahon *et al.*, 2006; Prim *et al.*, 2007; Reed *et al.*, 2000; Reed *et al.*, 2001; Reed *et al.*, 2003; Van der Horst *et al.*, 2011; Verlegh and Candel, 1999) have determined the influence of a variety of factors on consumers' consumption of ready meal products. A summary of these studies can be seen in table 5.1. These factors can be classified into eight aspects, which were integrated into the qualitative research part of this study:

(1) Price;
(2) Convenience;
(3) Sensory appeal;
(4) Healthiness;
(5) Novelty & familiarity;
(6) Cooking skills;
(7) Food safety and hygiene; and
(8) Pleasure.

5.1.4 Overview of the study

The development of an instrument for measuring consumer insights in this study followed Wilson's construct modelling approach (Wilson, 2004), in which the process of measurement is decomposed into four building blocks[3]. The study started with defining three initial latent constructs which accounted for three importance aspects regarding the ready meal consumption. Then information

[1] Food consumption can be decomposed into several stages: acquisition, delivery, storage, preparation, cooking, eating, and disposal & clean-up.

[2] The ready meal can be ready to heat or ready to end-cook (Costa *et al.* 2001). Ready to end-cook means it had been partially pre-cooked.

[3] The four building blocks are construct, item responses, outcome space and measurement model. The concept of them can be seen in section 1.2.

related to the defined constructs was collected via six focus group sessions using frequent consumers, occasional consumers and infrequent or non-consumers of ready meals, respectively (two sessions for each type of consumer). Based on the information, the questionnaire was composed with considerations on the item responses and the outcome space. The survey was administrated online. After that, the data were analysed using the Rasch analysis approach. According to the results, the initial constructs and instruments would be refined. After checking the model fit and reliability, the respondents were clustered into segments based on their measures (see section 5.2 for more details). The relationship between the consumer segments and their consumption frequency of ready meals, takeaway meals and restaurant meals were evaluated thereafter. In addition, information was extracted from the estimated item hierarchies of each construct, which can be used as a reference for developing new ready meal products.

Table 5.1 Previous research on ready meal consumption

Previous research	**Research interest**	**Format of the survey**
The consumption of convenience foods: reference groups and eating situations (Verlegh and Candel, 1999)	Belief towards ready meal consumption	Eleven items, 11-point scale from “Very unlikely” to “Very likely”
	Situation influence	
	Attribute (belief) evaluations	9-point scale from “Very negative” to “Very positive”
	Normative beliefs	9-point scale from “Very unlikely” to “Very likely”
	Motivation to comply	6-point scale from “Not agree at all” to “Totally agree”
	Behavioural intentions	9-point scale from “Very unlikely” to “Very likely”)
The retailing environment in Ireland and its effect on the chilled ready meal market (Reed *et al.*, 2000) & Factors affecting consumer acceptance of chilled ready meals on the island of Ireland (Reed *et al.*, 2003)	Consumers’ perceptions, consumption patterns and attitudes to chilled ready meals	The details of the questionnaire were not available
The chilled ready meal market in Northern Ireland (Reed *et al.*, 2001)	Consumers' purchasing habits and the factors influencing consumer chilled food choice	The details of the questionnaire were not available
Attitudes and beliefs directed towards ready-meal consumption (Ahlgren *et al.*, 2004)	Drivers of eating ready meal	Nineteen items 4-point Likert scale
	Attitudes towards ready-meal consumption	Eighteen items 4-point Likert scale
The impact of the meal situation on the consumption of ready meals (Ahlgren *et al.*, 2005)	Reasons for purchasing	Twenty-eight items, 4-point scale from “Not important” to “Very important”
	Frequent ready meal situations	Nineteen items, 4-point scale from “Not decisive” to “Very decisive”
The role of attitudes, subjective norm, perceived control and habit in the consumption of ready meals and takeaways in Great Britain (Mahon *et al.*, 2006)	Attitudes	Three items, 7-point Likert scale
	Subjective norm	One item, 7-point Likert scale
	Perceived control	One item, 7-point scale from “Very unlikely” to “Very likely”
	Behaviour intention	One item, 7-point scale from “Very unlikely” to “Very likely”
The appropriateness of ready meals for dinner (Prim *et al.*, 2007)	Focus group study	Qualitative study
Consumers’ health-related motive orientations and ready meal consumption (Geeroms *et al.*, 2008)	Belief about ready meal	Eleven items, 7-point scale from “Completely disagree” to “Completely agree”
	Perceived importance of different criteria when buying ready meal	Eleven items, 5-point scale from “Not at all important” to “Very important”
Ready-meal consumption: associations with weight status and cooking skills (Van der Horst *et al.*, 2011)	Beliefs about the nutritional value of ready meals	Eight items, 6-point scale from “Does not apply at all” to “Applies very much”
	Beliefs about the taste of ready meals	One item, 6-point scale from “Does not apply at all” to “Applies very much”

5.2 Defining the building blocks of measurement

5.2.1 Defining the initial construct

Three initial constructs were defined at the beginning of the study, in relation to consumers' satisfaction attitudes, product criteria and consumption situations. Figure 5.1 illustrates the proposed scaling orientations for the three constructs by respondent and item.

5.2.1.1 Satisfaction attitudes

Consumers' attitudes towards products or services are often studied in consumer research because attitudes could directly affect behaviour. Myers and Reynolds (1967) stated that "*purchase decisions are based almost solely upon attitudes existing at the time of purchase*". Attitude research is often used for predicting the purchasing intent, or segmenting consumers into subgroups so that different marketing strategies could be used.

In this study, a construct related to consumers' satisfaction attitudes towards the properties of ready meals was defined. Satisfaction towards ready meals refer to a collection of consumers' feelings towards various properties of ready meal products. It is a specific type of attitude. A hypothesis related to this construct is the more satisfied the consumers are, the more frequently they purchase ready meals compared to the less satisfied ones. In addition, the least satisfied properties of ready meals would be identified by investigating the positions of the items on the scale, which can be considered as the properties to be improved in new ready meal product.

5.2.1.2 Product criteria

This construct concerns consumers' decision making patterns and the relative importance of product criteria of ready meals for product selection. Whether the consumer's decision making pattern can affect their purchasing behaviour on ready meals was unknown. This would be explored in this study. Besides, the relative importance of each product criterion was evaluated according to the hierarchy of the items, which could be used as reference in the development of new ready meal products and the selection of advertising strategies, where the most important criteria should be emphasised.

5.2.1.3 Consumption situations

To investigate consumers’ willingness to consume ready meals under different contextual situations, this construct was proposed. The consumers who are willing to consume ready meal in most situations would be the main users of ready meals, who might be targeted for sales or recruited for new ready meal product development project as lead user (Von Hippel, 1986). Moreover, the link between consumers’ willingness to consume ready meals and the specific contextual situations could be studied, which might provide information for the ideation task of new product development.

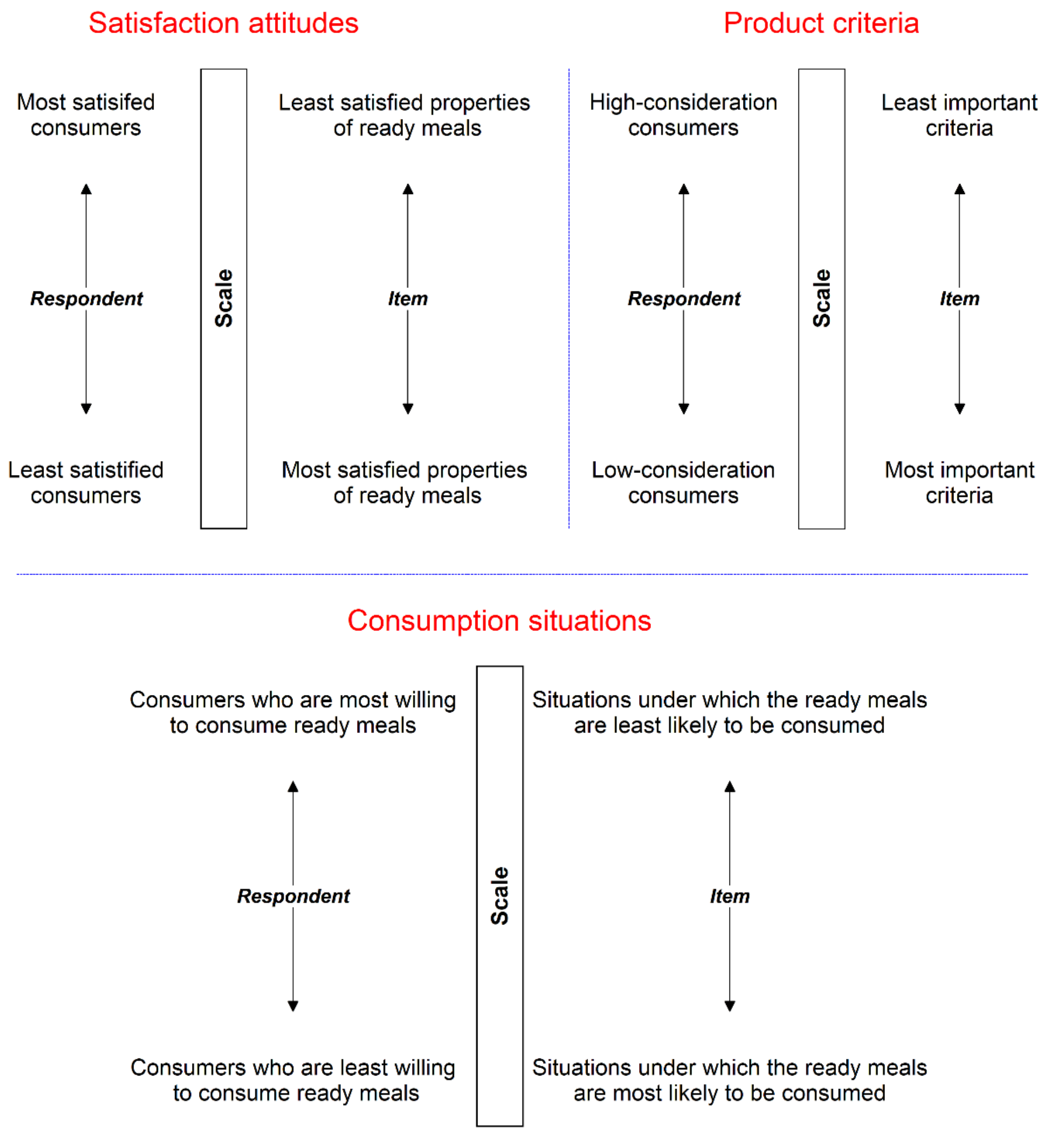

Satisfaction attitudes
Most satisifed consumers
Respondent
Least satistified consumers
Scale
Least satisfied properties of ready meals
Item
Most satisfied properties of ready meals
Product criteria
High-consideration consumers
Respondent
Low-consideration consumers
Scale
Least important criteria
Item
Most important criteria
Consumption situations
Consumers who are most willing to consume ready meals
Respondent
Consumers who are least willing to consume ready meals
Scale
Situations under which the ready meals are least likely to be consumed
Item
Situations under which the ready meals are most likely to be consumed

Figure 5.1 Three initial constructs defined at the beginning of the study

5.2.2 Designing the format of the items with consideration to item responses and outcome space

To measure the constructs, three instruments would be developed. The items were designed to collect consumers' responses in measurable formats with considerations of the item responses and the outcome space, which are the second and third building blocks of measurement.

Firstly, the descriptions of particular properties of ready meals were selected for the instrument measuring the satisfaction attitudes construct. The statements were drafted in the format of Likert items (Likert, 1932). The degree of respondents' agreeability over the items were recorded as the indicators of consumers' satisfaction on the properties.

Secondly, for the measurement of the product criteria construct, individual product criteria were used as items, while an importance rating scale was applied.

Thirdly, to evaluate respondents' willingness to consume ready meals under different contextual situations, the situations were used as the items, while the consumers' willingness to consume ready meals under each situation were rated.

Lastly, arbitrary raw scores were assigned to the categories of the three scales, which would be converted to linear measures via Rasch analysis. Table 5.2 tabulates the designated rating scales used in this study, respectively.

Table 5.2 Designated ratings scales used in this study

	Scale categories		
Raw scores	**Satisfaction attitudes**	**Purchasing criteria**	**Consumption situations**
	7-point Likert scale	**5-point importance scale**	**7-point likelihood scale**
0		Not at all important	
1	Strongly disagree	Slightly important	Unlikely very much
2	Disagree	Important	Unlikely
3	Slightly Disagree	Very important	Somewhat unlikely
4	Neither disagree nor agree	Extremely important	Undecided
5	Slightly agree		Somewhat likely
6	Agree		Likely
7	Strongly agree		Likely very much

5.2.3 Selecting measurement model

To model measures from raw scores, the measurement model, which is the last building block, needs to be selected. In this study, the Rating Scale Rasch model (Andrich, 1978a) was used for initial data analysis.

5.3 Developing the instruments for measuring the construct

5.3.1 Generating the initial item pool via focus group studies

The focus group approach was used for gathering information, from which the initial item pool was generated for the measurement of the three initial constructs proposed in section 5.2. This approach can be defined as a research method that "*collects data through group interaction on a topic determined by the researcher*" (Morgan, 1996). The development of the focus group method can be traced back to 1946, at which time it was called "focussed interview of groups" (Merton and Kendall, 1946).

The focus group study and the following survey research were approved by Faculty Research Ethics Committee (MEEC15-025).

In this study, twenty-eight participants were recruited from students and staff of University of Leeds for the focus group sessions. They were firstly self-classified into three types of ready meal consumers based on their consumption history of ready meals in the past six months prior to recruitment as "frequent consumer", "occasional consumer" and "infrequent/non-consumer". After that, six focus group sessions were conducted, including two sessions for each type of consumers. Each participant was required to attend one sessions with the same type of consumers. Table 5.3 depicts the criteria used for self-classification, and the gender and age distributions of the participants.

Table 5.4 describes the procedures and topics of the focus group studies designed by the researchers according to the three initial constructs and the eight aspects related to ready meal consumption, which was identified from literature (see section 5.1.2.2). Each session lasted for one to one and a half hours accordingly, which were recorded by an audio recorder. The records were transcribed into written scripts. The items were drafted according to the information extracted from the ready meal related quotations associated with the three constructs. Some statements were adapted from the previous research listed in Table 5.1. Eventually eighty-eight items were drafted. Half of the items used to measure the satisfaction attitude construct were positive-worded, such as “ready meals are usually tasty”, while the other half of them were negative-worded, such as “Ready meals don’t taste fresh”.

Table 5.3 Classification of focus group participants and the demographic distribution

Group		Frequent consumer	Occasional consumer	Infrequent/ Non-consumer	Total
Criteria (In past six months, I had purchased ready meal)		Every month	A few times but not in every month	Not at all	
Number of participants		9	12	7	28
Gender	Female	4	8	6	18
	Male	5	4	1	10
Age	18-24	0	4	2	6
	25-34	4	6	4	14
	35-44	2	1	1	4
	45-54	3	1	0	4

5.3.2 Pilot testing and the composition of the survey questionnaire

A pilot testing on the eighty-eight construct-based items was conducted using the Bristol online survey tool with five participants. According to the feedbacks, five items were removed because they were redundant in their meanings, while another five items were rephrased due to unclear wordings. Eventually eighty-three items were retained in the refined item pool, including forty-four, twenty-four and fifteen items about satisfaction attitudes, product criteria and consumption situations, respectively.

After that, the survey was assembled in the Bristol online survey tool using the refined items. Three additional items about the consumption frequency of three types of meals, including take-away meals, restaurant meals and ready meals, were added to the questionnaire before the eighty-three construct-based item, using a 5-point frequency scale from “Never” to “Every day”[1]. Items for gender and age were also included in the questionnaire. More details of the composition of the survey can be seen in table 5.5. The list of item statements will be provided in section 5.6.3.

[1] The scale categories are: 0=“Never”, 1=“Less than once per month”, 2=“One to three times per month”, 3=“A few times per week but not every day”, and 4=“Everyday”.

Table 5.4. The procedures and topics of the focus group studies

<table>
<tr><th></th><th>Frequent consumer</th><th>Occasional consumer</th><th>Infrequent or Non-consumer</th></tr>
<tr><td><b>0. Greetings and warm-up activity (five-ten minutes)</b></td><td colspan="3">The panel members were given a brief introduction about the aims of the research.
They were encouraged to introduce themselves to each other.</td></tr>
<tr><td><b>1. Opening topic (ten-fifteen minutes)</b></td><td colspan="3">All panel members were asked to write down their own definitions of ready meal on paper then pass it to next person clockwise. They were then led to discuss the definition of ready meals with the presence of stimuli (commercial ready meal products).
After that, the definition of each meal used by the research team was given to the panel member to ensure everyone would discuss further topics on the same thing.</td></tr>
<tr><td><b>2. Transition Topic (ten-fifteen minutes)</b></td><td colspan="2">The panel members were asked to describe their latest experience of ready meal consumption. The specific topic is:
Could you recall your last or one of your recent experience of ready meal consumption? Which product did you choose? What criteria of the product made you choose it?</td><td><i>This topic was skipped for rare/non-consumer</i></td></tr>
<tr><td><b>3. Transition Topic (five-ten minutes)</b></td><td>What are the main reasons or drivers that make you buy ready meal frequently?</td><td>What occasions or situation would you like to purchase ready meal? What are reasons or barriers that stopped you being a frequent customer?</td><td>What are the main reasons or barriers that make you not buy ready meal at all?</td></tr>
<tr><td><b>4. Key topic (fifteen-thirty minutes)</b></td><td colspan="3">The panel members were asked to compare four types of meals on the eight aspects summarised from the literature. In addition, they were also encouraged to propose additional aspects that can be compared.
Four type of meals: (1) Ready meals; (2) Takeaway/ready-to-eat meals; (3) Restaurant meal; and (4) Home-made meals.
Eight aspects: (1) Price; (2) Convenience (availability, time and energy) (within food consumption cycle); (3) Sensory appeal; (4) Healthiness; (5) Novelty and familiarity; (6) Cooking skill required; (7) Food Safety and Hygiene; and (8) Pleasure.</td></tr>
<tr><td><b>5. Key topic (five-ten minutes)</b></td><td colspan="3">A discussion about the potential eating situations of ready meal products was raised. The focus was on the context such as “when”, “where”, “with whom”, <i>etc.</i></td></tr>
<tr><td><b>6. Ending topic (five-ten minutes)</b></td><td colspan="3">The suggestion on the improvement/selling point of new ready meal product. The specific question is:
If you are working for the new product development team of food industry, and your task is to develop a new ready meal product, what improvement or selling point would you like to bring into your product?</td></tr>
</table>

Table 5.5 The composition of the survey questionnaire

Sections	Topic	Content or the particular question
Part I	Consent	Participant information and consent form
Part II	Introduction of ready meal product	The definition of ready meal product adopted in this study and some examples in photos
Part III	Consumption frequency of three types of meals[1]	How often do you consume the meals listed below in the past six months? (three items)
Part IV	Satisfaction attitudes construct	To what extent do you agree or disagree with the statements listed below? (forty-four items)
Part V	Product criteria construct	If you are going to buy a ready meal product - to what extent do you think the following criteria are important when you decide which particular ready meal product to buy? (twenty-four items)
Part VI	Consumption situations construct	How likely would you have ready meal when you are under the situations listed below? (fifteen items)
Part VII	Demographic information	Gender and age

[1] Take-away meal, restaurant and ready meal

5.4 Sampling

The survey was opened to UK residents for one year through the Bristol online survey tool. The respondents were recruited via email, poster and personal contact. A total number of 676 visits were recorded, among which only 339 respondents proceeded to the last survey page. Six respondents were further removed because they skipped almost the whole questionnaire. The final data set, assembled from the information, was provided by 333 respondents.

Table 5.6 tabulates the distribution of gender and age among the respondents. The age groups were combined into three because the numbers of respondents in some of the groups were too small for comparison.

Table 5.6 Distributions of gender and age

	Gender		Age		
Group	Female	Male	16-24	25-34	35+
Number of respondents	245	88	145	106	82

*Age group “35-44”, “45-54”, “55-64”, and “65+” were combined to a “35+” group.

5.5 Data analysis procedures

5.5.1 Rasch analysis

The data in relation to the three initial constructs were fitted to the Rating Scale Rasch model individually using WINSTEPS (Linacre, 2014d). The functioning of rating scale categories was determined first, followed by tests for the assumption of unidimensionality and local item independence. The assumption of unidimensionality was tested using PCA on standardised model residuals (PCAR) and the independent t-test protocol. The more detailed description of the methods can be found in section 2.4.2.2. If the assumption of unidimensionality could not hold, then the instrument would be split into subsets, until all subsets were found unidimensional. In addition, the assumption of local item independence was evaluated by accessing the residual correlations between items. Data under the revised instruments[1] would be used for further analysis. The test of fit, Wright maps and reliability statistics would be reported.

5.5.2 Segmenting respondents into groups using cluster analysis

The DIvisive ANAlysis Clustering (DIANA clustering) was conducted for segmenting respondents into groups by R package "cluster" (Maechler *et al.*, 2018), using their measures [2] estimated by Rasch analysis on the satisfaction attitudes construct and product criteria construct as variables, respectively,. In addition, the DIANA clustering was also performed on respondents' raw mean scores on same instruments for comparison purpose. The DIANA clustering (Macnaughton-Smith *et al.*, 1964) is a type of hierarchical clustering approach. It starts with a single cluster containing all elements, and at each step, the cluster is split into smaller groups, until all clusters contain only one element in each (Kaufman and Rousseeuw, 2005).

After the respondents were clustered into segments, the mean measures and the raw mean scores of each segment were compared on the construct basis, using ANOVA and Kruskal-Wallis test (Kruskal and Wallis, 1952) using R program (R Core Team, 2018). If the results were significant, then the parametric Tukey HSD test and non-parametric Dunn's test (Dunn, 1964) with Hochberg correction

[1] The statements of the items can be seen in tables 5.10 ~ 5.16

[2] The person's measures on product criteria construct reflected their decision pattern. The higher the measure, the more consideration the respondent would have.

(Hochberg, 1988) were applied for multiple comparisons, using R package PMCMRplus (Pohlert, 2018).

5.5.3 Exploring the relationship between respondent segments and consumption frequency

The Kruskal-Wallis test was employed to compare the consumption frequency of three types of meals by respondent segments and by age groups. The Wilcoxon test (Wilcoxon, 1945) was performed for the comparisons by gender groups using R program. In addition, if the results of Kruskal-Wallis test were significant, Dunn's test with Hochberg correction was conducted for multiple comparisons.

5.6 Results

5.6.1 Rating scale category effectiveness

The rating scale category effectiveness of the three instruments were evaluated. The results indicated that the rating scales used in the instruments measuring satisfaction attitudes construct and consumption situations construct needed to be optimised, whereas the performance of the rating scale used for measuring product criteria construct was acceptable.

Firstly, all “essential” diagnostic criteria (refer to section 2.3.1.2) of scale categories effectiveness suggested by Linacre (2002a) were fulfilled within all instruments.

Secondly, as shown in table 5.7, although the Rasch-Andrich thresholds advanced monotonically in the satisfaction attitudes instrument, the minimal advancing distance was only 0.10, which was less than the recommended value of 0.58 for a 7-point scale (Linacre, 2006). This implied that the intervals covered by some of the scale categories were too narrow on the construct. There was no issue regarding the Rasch-Andrich thresholds of the 5-point importance scale used in the product criteria instrument. But disordered Rasch-Andrich thresholds were observed with the 7-point likelihood scale of consumption situations instrument, implying that some of the scale categories were never modal. The potential reason might be the meanings of some scale categories were too close, resulting in that the respondents could not clearly distinguish them in use.

Attempts at collapsing categories were made for the satisfaction attitudes instrument and consumption situations instrument. Table 5.7 compares the Rasch-Andrich thresholds of all above instruments before or after the rating scales were revised. It was found that the Rasch-Andrich thresholds could advance monotonically by distances greater than the minimum advancing distance suggested by Linacre (2006), when the 7-point Likert scale were collapsed to a 5-point scale for satisfaction attitudes instrument, and the 7-point likelihood scale were collapsed to a 4-point scale for consumption situations instrument. Therefore, the data sets of the satisfaction attitudes and consumption situations instruments were revised in these ways before further analysis was conducted.

Table 5.7 Rasch-Andrich thresholds and the effect of collapsing categories

Instrument	**Rating scale[1]**	**Number of categories**	**Rasch-Andrich threshold**	**Minimum advancing recommended[2]**	**Actual minimum advancing**
Satisfaction attitudes	Original 7-point scale	7	Ordered	0.58	0.10
	Collapse category 3-4-5	5	Ordered	0.81	0.47
	Collapse category 2-3 and 4-5	5	Ordered	0.81	1.07
Product criteria	Original 5-point scale	5	Ordered	0.81	0.84
Consumption situations	**Original 7-point scale**	7	Disordered	0.58	
	Collapse category 3-4-5	5	Disordered	0.81	
	Collapse category 2-3 and 4-5	5	Ordered	0.81	0.50
	Collapse category 2-3 and 4-5-6	4	Ordered	1.10	0.30
	Collapse category 2-3-4 and 5-6	4	Ordered	1.10	1.35

[1] The labels of the categories are listed in table 5.2. The final solutions were labelled in **bold**.

[2] The minimum advancing distance between Rasch-Andrich thresholds were estimated using equation 2.21(section 2.3.1.2).

5.6.2 Tests of unidimensionality and local item independence

Four and two individual dimensions were identified under the satisfaction attitudes instrument (figure 5.2) and the product criteria instrument (figure 5.3), respectively, while the consumption situations instrument was considered as unidimensional (figure 5.4).

5.6.2.1 Unidimensionality and local item independence of the satisfaction attitudes instrument

The eigenvalue of the unexplained variance in first extracted PCAR contrast in satisfaction attitudes data was 4.70, implying that there might be an additional dimension, as it had a strength of greater than three items. The items were divided into three clusters by WINSTEPS according to their residual loadings on the first PCAR contrast. Cluster 1 was made up of items which had the highest loadings, while cluster 3 consisted of items with lowest loadings. The other items were assigned to cluster 2. The largest disattenuated correlation was observed between the first and second item clusters, which had a value of 0.84. This value was higher than 0.82, above which the persons' measures on the two item clusters would be twice dependent than independent (Linacre, 2014c). The disattenuated correlations between the other pairs of item clusters were all less than this value. This means the connection between items in cluster 1 and 2 were closer than with items in cluster 3. Therefore, the instrument was split into two subsets, according to the cluster membership and the actual meanings of the item statement.

The first subset contained 11 items from cluster 3. The data were refitted to the model, followed by the PCAR test. The eigenvalue of the first contrast was 2.01, implying that the subset might be unidimensional. However, the lower bound of 95% binomial CI of the independent t-test was 10.91%, which was higher than the recommended rule of less than 5%. Further inspection of local item independence found that two item pairs were potentially dependent, as their residual correlations were equal to or greater than 0.30. The statements of the potential LID items were reviewed (table 5.8). A decision was made to combine the two item pairs (A18/A19, and A11/A16) to form two super-items (A1819 and A1116), as their meanings were similar. The summated score of the item pair was used as the score of the super-item. After the creation of the super-items, the data were fitted to a partial-credit model (Masters, 1982), in which all single items were parameterised using the same rating scale structure, and the super-items scored by summating the raw scores of two original items were defined using another scale parameter. The PCAR test and t-test protocol were applied again to the revised data. The eigenvalue of the first PCAR contrast decreased to 1.69, while higher disattenuated correlations between item clusters could observed compared to the original subset. Although the proportion of significant t-tests was still higher than 5% (6.61%), the estimate of the lower bound of the 95% binomial CI was only 4.26%, below the threshold of 5%. This strongly indicated that the revised subset was unidimensional. All items under this subset could measure consumers' satisfaction towards the convenience and cost properties of ready meal. It was named as "satisfaction attitudes towards product purchase properties (of ready meals)[3]".

Another subset split from the whole satisfaction attitudes instrument was also fitted to the model. After repeating the PCAR procedure, it was found that the eigenvalue of the first PCAR contrast was still larger than 3 (3.68). This subset corresponds to the subset labelled as "undefined 1" in figure 5.2. It was split into two more subsets for further analysis. One of the newly defined subsets (eight items) was considered as a unidimensional instrument according to the results of the PCAR test followed by the independent t-tests. The items under it all refer to consumers' satisfaction towards "health benefit" of ready meals. The other new subset, on the contrary, could not be proved as unidimensional instrument, even after two pairs of items were combined to super-items due to LID. It was further split into two more subsets, which were verified as two unidimensional instruments. One of them concerns consumers' satisfaction towards the "product consumption properties" of ready meals, which mainly covers the sensory perception and the enjoyment, while another one is related to the "product characteristics" of ready meals such as the availability and packaging design.

5.6.2.2 Unidimensionality and local item independence of the product criteria instrument

The product criteria instrument was found to be multidimensional, as the eigenvalue of the unexplained variance in the first extracted PCAR contrast was 5.01 (figure 5.4). It was much higher than 3.00, above which would suggest multidimensional (Linacre, 2014c). Again, three item clusters were reported by software WINSTEPS according to the residual loadings on the first contrast. The disattenuated correlation between item clusters 1 and 2 was 0.89. It was above the value of 0.87, implying that the two item sets definitely measure the same thing (Linacre, 2014c). Therefore all items belonging to item clusters 1 and 2 were used to assemble a new data set, and the other items were kept together as

[3] In this study, all instruments were specific to ready meals, even if it was not labelled due to limits of the words.

another data set. The two subsets were fitted to the rating scale Rasch model separately. Both a PCAR test and the independent t-tests suggested that they were unidimensional (figure 5.3), as:

(1) Eigenvalue of the first contrast was below 3.00;
(2) The disattenuated correlation between clusters were quite high;
(3) The estimates of lower bound of 95% binomial CI were less than 5% in either subsets.

In addition, no potential dependent item pairs were found, as the residual correlation between items within the two subsets were all less than 0.3.

After reviewing the meanings of the item statements (table 5.8), the two subsets of product criteria were renamed as intrinsic criteria and extrinsic criteria.

5.6.2.3 Unidimensionality and local item independence of the consumption situations instrument

The eigenvalue of unexplained variance in first contrast extracted by PCAR on the consumption situations instrument was very close to 3.00 (figure 5.4). The disattenuated correlations between item clusters 1 and 2 (1.00), and item clusters 2 and 3 (0.77) both implied unidimensionality. However the estimate of lower bound of 95% binomial CI failed to prove the assumption of unidimensionality.

Potential violations of the assumption of local item independence in three item pairs were indicated as their residual correlations were equal to or greater than 0.30. The statements of these items were reviewed. Eventually the item S8 and S12 were combine to a super-item, whereas the other two item pairs were retained in single item format because they were relatively different in the meanings. The revised data were refitted to a partial credit model. The super-item was modelled using its own rating scale structure.

The PCAR and independent t-test were redone with the revised model. The eigenvalue of the first contrast was less than 3.00. Although the independent t-test had a lower bound of 95% binomial Ci at the value of 13.86%, which was still greater than the threshold of 5%, the consumption situations instrument was determined as a unidimensional instrument because the disattenuated correlation between item clusters 1 and 2, and item clusters 2 and 3 were both higher than 0.82, which would suggest that they probably measure the same thing (Linacre, 2014c).

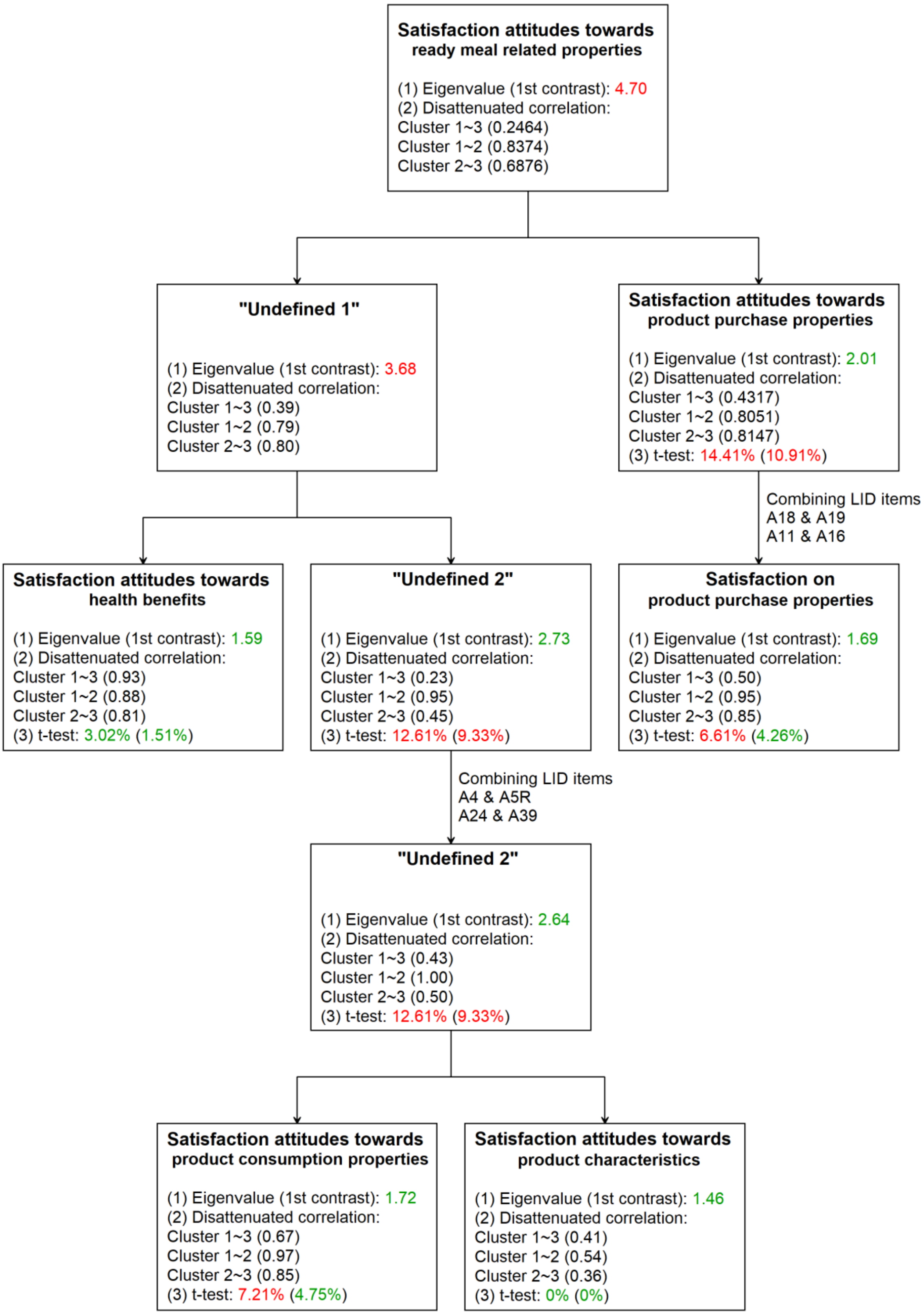

Satisfaction attitudes towards ready meal related properties
(1) Eigenvalue (1st contrast): 4.70
(2) Disattenuated correlation:
Cluster 1~3 (0.2464)
Cluster 1~2 (0.8374)
Cluster 2~3 (0.6876)
"Undefined 1"
(1) Eigenvalue (1st contrast): 3.68
(2) Disattenuated correlation:
Cluster 1~3 (0.39)
Cluster 1~2 (0.79)
Cluster 2~3 (0.80)
Satisfaction attitudes towards product purchase properties
(1) Eigenvalue (1st contrast): 2.01
(2) Disattenuated correlation:
Cluster 1~3 (0.4317)
Cluster 1~2 (0.8051)
Cluster 2~3 (0.8147)
(3) t-test: 14.41% (10.91%)
Combining LID items
A18 & A19
A11 & A16
Satisfaction attitudes towards health benefits
(1) Eigenvalue (1st contrast): 1.59
(2) Disattenuated correlation:
Cluster 1~3 (0.93)
Cluster 1~2 (0.88)
Cluster 2~3 (0.81)
(3) t-test: 3.02% (1.51%)
"Undefined 2"
(1) Eigenvalue (1st contrast): 2.73
(2) Disattenuated correlation:
Cluster 1~3 (0.23)
Cluster 1~2 (0.95)
Cluster 2~3 (0.45)
(3) t-test: 12.61% (9.33%)
Satisfaction on product purchase properties
(1) Eigenvalue (1st contrast): 1.69
(2) Disattenuated correlation:
Cluster 1~3 (0.50)
Cluster 1~2 (0.95)
Cluster 2~3 (0.85)
(3) t-test: 6.61% (4.26%)
Combining LID items
A4 & A5R
A24 & A39
"Undefined 2"
(1) Eigenvalue (1st contrast): 2.64
(2) Disattenuated correlation:
Cluster 1~3 (0.43)
Cluster 1~2 (1.00)
Cluster 2~3 (0.50)
(3) t-test: 12.61% (9.33%)
Satisfaction attitudes towards product consumption properties
(1) Eigenvalue (1st contrast): 1.72
(2) Disattenuated correlation:
Cluster 1~3 (0.67)
Cluster 1~2 (0.97)
Cluster 2~3 (0.85)
(3) t-test: 7.21% (4.75%)
Satisfaction attitudes towards product characteristics
(1) Eigenvalue (1st contrast): 1.46
(2) Disattenuated correlation:
Cluster 1~3 (0.41)
Cluster 1~2 (0.54)
Cluster 2~3 (0.36)
(3) t-test: 0% (0%)

Figure 5.2 Tests of unidimensionality of satisfaction attitudes instrument

The values in brackets for t-tests are the estimated lower bound of 95% binomial CI.

The satisfied statistics are marked in green, while the unsatisfied statistics are marked in red.

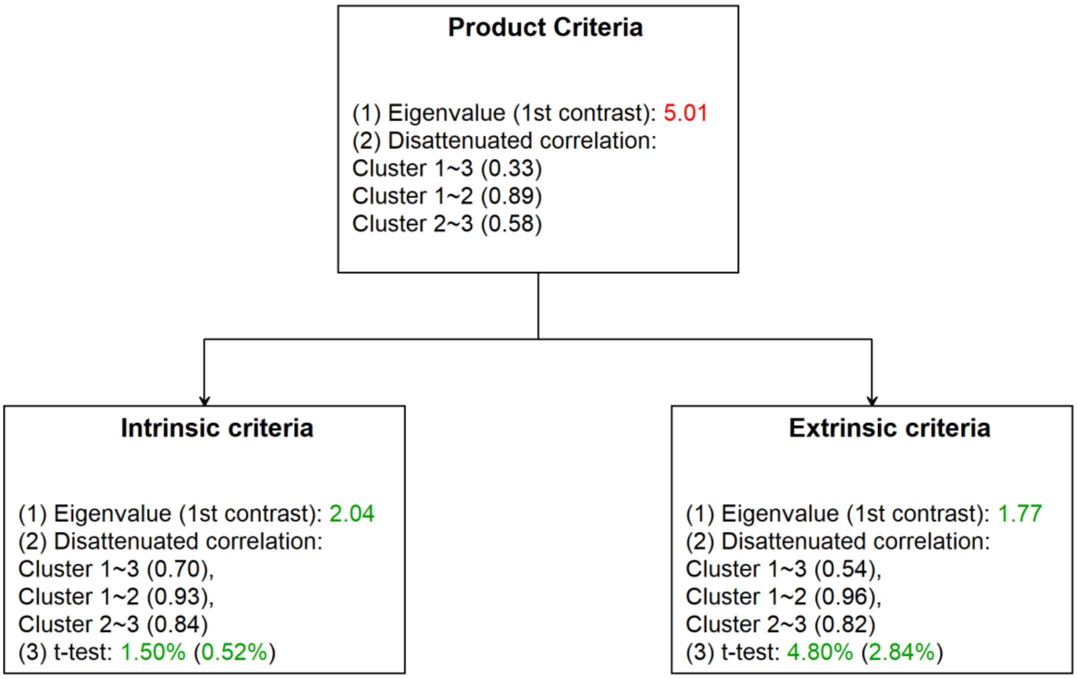

Product Criteria
(1) Eigenvalue (1st contrast): 5.01
(2) Disattenuated correlation:
Cluster 1~3 (0.33)
Cluster 1~2 (0.89)
Cluster 2~3 (0.58)
Intrinsic criteria
(1) Eigenvalue (1st contrast): 2.04
(2) Disattenuated correlation:
Cluster 1~3 (0.70),
Cluster 1~2 (0.93),
Cluster 2~3 (0.84)
(3) t-test: 1.50% (0.52%)
Extrinsic criteria
(1) Eigenvalue (1st contrast): 1.77
(2) Disattenuated correlation:
Cluster 1~3 (0.54),
Cluster 1~2 (0.96),
Cluster 2~3 (0.82)
(3) t-test: 4.80% (2.84%)

Figure 5.3 Tests of unidimensionality of product criteria instrument

The values in brackets for t-tests are the estimated lower bound of 95% binomial CI.

The satisfied statistics are marked in green, while the unsatisfied statistics are marked in red.

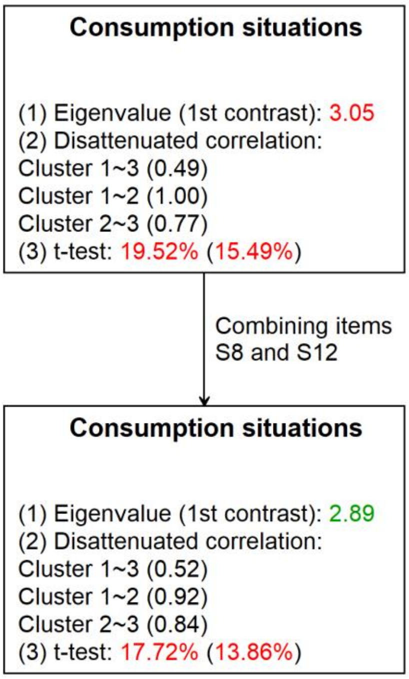

Consumption situations
(1) Eigenvalue (1st contrast): 3.05
(2) Disattenuated correlation:
Cluster 1~3 (0.49)
Cluster 1~2 (1.00)
Cluster 2~3 (0.77)
(3) t-test: 19.52% (15.49%)
Combining items S8 and S12
Consumption situations
(1) Eigenvalue (1st contrast): 2.89
(2) Disattenuated correlation:
Cluster 1~3 (0.52)
Cluster 1~2 (0.92)
Cluster 2~3 (0.84)
(3) t-test: 17.72% (13.86%)

Figure 5.4 Tests of unidimensionality of consumption situations instrument

The values in brackets for t-tests are the estimated lower bound of 95% binomial CI.

The satisfied statistics are marked in green, while the unsatisfied statistics are marked in red.

Table 5.8 Item pairs that had a residual correlation equal to or greater than 0.30

Instrument	Item 1	Item2	Residual correlation	Decision
Satisfaction attitudes				
Product purchase	**A18:** It is convenient to me that the ready meal products normally can be stored for a long time.	**A19:** Ready meals are good back-ups to have at home.	0.32	Combine to item **A1819**
	A11: The price of a ready meal product is usually low.	**A16:** Ready meal is of good value of money for the food you can get.	0.30	Combine to item **A1116**
Product consumption	**A4:** Ready meals are usually tasty	**A5R:** Ready meals usually taste bland (lack of flavours)	0.39	Combine to item **A45R**
Product characteristics	**A24:** Generally speaking, the design of packaging of ready meal product is attractive.	**A39:** Generally speaking, the appearance of ready meal product is attractive.	0.31	Combine to item **A2439**
Consumption situations	**S8:** I'm curious about a new recipe. I found the ready meal version which can be tried before I cook it myself.	**S12:** I spot a new dish (ready meal) that I haven't tried before.	0.43	Combine to item **S812**
	S9: I want to spend time on things other than cooking in my free time.	**S10:** I'm not in the mood to cook.	0.39	Keep separately
	S1: I don't have time to cook my own food.	**S6:** I'm too tired to cook.	0.30	Keep separately

R in item label means the item was negative-worded.

5.6.3 Test of individual fit

To evaluate the individual fit, the outfit MNSQ statistics of the respondents and items in refined instruments were estimated.

The proportions of misfitting respondents within each instrument were calculated, as shown in table 5.9. A number of respondents showed overfit (*i.e.* outfit MNSQ <0.50) or slightly underfit (*i.e.* outfit MNSQ between 1.50-2.00). Although not productive, they would not degrade the measurement (Wright and Linacre, 1994). Small percentages of respondents were found to be strongly underfit, as their outfit MNSQ values were greater than 2.00. However, the impact of misfitting persons on the measurement would be arguably less than that of misfitting items (Wright and Linacre, 1994). They were unlikely to have a great effect on the estimates.

Item A32 and C16 exhibited slightly underfit[1], as their outfit MNSQ statistics were just above the suggested range of 0.50~1.50 (Wright and Linacre, 1994). But they were not likely to degrade the measurement, because the degradation would be indicated by an outfit MNSQ value greater than 2.00 (Wright and Linacre, 1994). Therefore they were retained. To improve the fit, the extreme unexpected responses[2] to A32 and C16 were removed. The revised data sets were refitted to the model. Further inspection on the outfit MNSQ statistics suggested that all items in the revised data sets fit well to the model, as their MNSQ were all in suggested range of 0.50~1.50 (table 5.10-5.16).

Table 5.9 The proportion of misfitting respondents

Instrument	**Overfit**	**Underfit**	
	<0.50[1]	1.50~2.00[1]	>2.00[1]
Satisfaction attitudes towards			
Product purchase	20.42%	9.61%	6.31%
Health benefit	24.62%	8.41%	9.61%
Product consumption	17.72%	9.31%	6.31%
Product characteristics	21.92%	9.31%	6.91%
Product criteria			
Intrinsic criteria	18.62%	10.21%	8.11%
Extrinsic criteria	15.32%	8.71%	6.31%
Consumption situations	18.92%	7.21%	9.91%

[1] Outfit MNSQ statistics

[1] The outfit MNSQ values of item A32 and C16 were 1.52, and 1.68, respectively, before the extreme unexpected responses were removed.

[2] Extreme unexpected response was defined as the response which had an absolute standardised residual equal to or greater than 2.

Table 5.10 The measure, standard error (SE) and outfit MNSQ statistics of items within the instrument about consumers’ satisfaction towards product purchase properties of ready meals

Item	**Statement**[1]	**Measure±SE**	**Outfit MNSQ**
A1116[2]	Ready meal products are usually cheap and cost effective.	1.19±0.05	1.11
A30	Ready meal is of good value of money for the time/energy saving on cooking.	0.76±0.07	0.78
A22	Storing ready meal can save space in fridge and/or freezer than storing raw ingredients because ready meal is an all-in-one meal pre-packed in a regular shape.	0.41±0.07	1.06
A8	It is convenient to me that I can buy multiple ready meal products during one shopping trip.	0.33±0.07	0.90
A14	The price of ready meal product reflects its quality like the taste and ingredient usage (*i.e.* expensive=better quality, and cheap=poor quality.	0.10±0.07	1.38
A1819[3]	It is convenient to me that the ready meal products can normally be stored for a long time as backup.	-0.16±0.05	1.00
A6	It is convenient to me that ready meals can be bought everywhere.	-0.26±0.07	0.80
A23	Having ready meal can save a lot of efforts spent on buying ingredients, and preparing meal & clean-up at home.	-0.56±0.07	0.84
A21	Ready meals are easy to prepare, even for someone who does not have particular knowledge of cooking.	-1.81±0.09	0.99

R means the item was negative-worded in the instrument.

[1] Question: To what extent do you agree or disagree with the statements listed below?

[2] Item A1116 was combined from A11 and A16.

[3] A1819 was combine from A18 and A19.

Table 5.11 The measure, standard error (SE) and outfit MNSQ statistics of items within instrument about consumers' satisfaction towards the health benefit properties of ready meals

Item	Statement[1]	Measure±SE	Outfit MNSQ
A27	Ready meals are normally quite healthy.	0.47±0.11	0.90
A38R	Ready meals contain too much salt that has adverse effects on health.	0.42±0.11	0.85
A44R	I think ready meals are not healthy because they are highly processed.	0.20±0.10	1.27
A12R	Ready meals contain more than desirable amount of additives that can impair health.	0.11±0.10	0.88
A31R	Ready meals have a high-fat content, which is not good for health.	-0.03±0.10	0.80
A42R	Ready meals are high in calories.	-0.14±0.10	0.96
A28R	Ready meals contain excessive sugar, which is bad for health.	-0.28±0.10	0.86
A35R	Ready meals are not nutritionally balanced meals because they are lack of vegetables inside.	-0.75±0.09	1.46

R means the item was negative-worded in the instrument.

[1] Question: To what extent do you agree or disagree with the statements listed below?

Table 5.12 The measure, standard error (SE) and outfit MNSQ statistics of items within instrument about consumers' satisfaction towards product consumption properties of ready meals

Item	Statement[1]	Measure±SE	Outfit MNSQ
A37R	The taste of ready meals are often not as good as the appearance indicated.	1.30±0.09	0.93
A40R	Ready meals don't taste fresh.	0.85±0.08	0.85
A9R	There is no pleasure in cooking ready meal.	0.76±0.08	1.41
A10R	Ready meals are extremely bad to the environment due to the disposal of packaging materials.	0.58±0.08	0.95
A34R	Ready meals often taste quite pasty (I cannot identify individual ingredient from the texture).	0.31±0.08	0.76
A2R	I think the ingredients used in ready meals have bad quality.	0.23±0.08	0.80
A15R	Ready meals are usually too salty.	0.17±0.08	1.28
A29	Having ready meals can reduce the amount of food waste.	-0.09±0.07	1.02
A45R[2]	Ready meals are usually tasty.	-0.18±0.05	1.03
A20R	I don't like the taste of the ready meals of foreign cuisine because they are not authentic at all.	-0.23±0.07	0.89
A33R	I think ready meals are not healthy because I don't know what's inside, even there is a list of ingredients on the label.	-0.25±0.07	1.24
A43R	Ready meals are usually too sweet.	-0.49±0.07	0.87
A26R	There is no pleasure in eating ready meal.	-0.66±0.07	0.72
A1R	I don't enjoy the process of choosing the particular ready meal product from a wide range of products on shelves.	-0.93±0.07	1.39
A3	I think the ready meals are safe to eat because the production are governed by legislation and industry standards.	-1.34±0.07	1.05

R means the item was negative-worded in the instrument.

[1] Question: To what extent do you agree or disagree with the statements listed below?

[2] Item A45R was combined from A4 and A5R.

Table 5.13 The measure, standard error (SE) and outfit MNSQ statistics of items within the instrument about consumers' satisfaction towards product characteristics of ready meals

Item	Statement	Measure±SE	Outfit MNSQ
A25	The serving suggestion (*e.g.* "Serves") on the labels of ready meals are always correct.	1.01±0.08	0.89
A32	More variety of ready meal products will cause difficulty in selecting the product.	0.63±0.08	[3]0.96
A41	The proportions of individual components (*e.g.* solid components and sauce) in ready meals are always proper.	0.58±0.08	0.81
A7R	The range of ready meal in family pack is too small.	0.17±0.08	0.97
A2439[2]	Generally speaking, the appearance of ready meal product is attractive.	0.16±0.05	1.01
A13	Ready meals are sold in sufficient portion to me.	-0.56±0.07	1.08
A17	There is sufficient nutrition information on the labels of ready meal products.	-0.81±0.07	1.19
A36	There are sufficient variety of ready meal products on the market.	-1.18±0.07	1.03

R means the item was negative-worded in the instrument.

[1] Question: To what extent do you agree or disagree with the statements listed below?

[2] Item A2439 was combined from A23 and A39.

[3] After removing the extreme unexpected responses

Table 5.14 The measure, standard error (SE) and outfit MNSQ statistics of items belong to intrinsic criteria instrument

Item	Statement[1]	Measure±SE	Outfit MNSQ
C5	Amount of protein per meal	1.22±0.07	1.21
C16	Origin of the ingredients (*e.g.* Angus beef, British vegetable, *etc.*)	0.52±0.07	[1]1.05
C3	“Five-a-Day” equivalence per meal	0.35±0.07	1.47
C15	Additional nutritional information (fibre, vitamins, *etc.*)	0.19±0.07	0.81
C8	Nutritional claim (*e.g.* high in fibre, high in Omega-3 fatty acids)	0.13±0.07	0.73
C7	Amount of sugar per meal	0.05±0.07	0.81
C19	Amount of salt per meal	-0.15±0.07	0.82
C9	Total energy (calories) per meal	-0.27±0.07	1.03
C17	Amount of fat per meal	-0.33±0.07	0.74
C6	Freshness	-0.64±0.07	1.41
C18	Ingredients inside the product	-1.08±0.07	1.10

[1] Question: If you are going to buy a ready meal product - to what extent do you think the following criteria are important when you decide which particular ready meal product to buy?

[2] After removing the extreme outliers.

Table 5.15 The measure, standard error (SE) and outfit MNSQ statistics of items within extrinsic criteria instrument.

Item	Statement[1]	Measure±SE	Outfit MNSQ
C4	It is something new to me.	0.87±0.06	1.19
C20	Claims of the taste (*e.g.* “restaurant quality” and “taste like home-made”)	0.66±0.06	1.18
C1	Packaging material (*e.g.* plastic tray or foil tray)	0.41±0.06	0.92
C23	Brand	0.40±0.06	0.95
C24	Design of outer packaging	0.40±0.06	0.80
C21	Promotion	0.24±0.06	1.06
C11	Shelf-life	-0.10±0.06	1.06
C22	Is it suitable for domestic freezer (if it is displayed as a chilled product)	-0.14±0.06	1.11
C10	Cooking method (*e.g.* microwaveable, oven cook only, *etc.*)	-0.27±0.06	0.97
C2	I’m familiar with the dish.	-0.28±0.06	1.05
C13	Portion size	-0.44±0.06	0.68
C14	Whether I can see the real appearance of product inside the outer packaging	-0.66±0.06	1.05
C12	Price	-1.07±0.06	0.96

[1] Question: If you are going to buy a ready meal product - to what extent do you think the following criteria are important when you decide which particular ready meal product to buy?

Table 5.16 The measure, standard error (SE) and outfit MNSQ statistics of items within consumption situations instrument

Item	Statement	Measure±SE	Outfit MNSQ
S3	I invited my friends for lunch, or dinner or having a party.	1.96±0.09	1.38
S7	I want to enjoy my evening with my family.	1.48±0.09	0.94
S15	I'm in a mood to eat something delicious.	1.08±0.08	0.86
S812[1]	I spot something (ready meal) new to me.	0.64±0.05	1.40
S11	I'm alone.	0.24±0.08	1.06
S9	I want to spend time on things other than cooking in my free time.	0.08±0.08	0.81
S2	I'm stressed.	-0.15±0.08	0.81
S14	I haven't prepared any lunch to take to work/school.	-0.20±0.08	1.28
S5	I found a ready meal product which is on promotion/special offer.	-0.45±0.08	1.00
S10	I'm not in the mood to cook.	-0.50±0.08	0.70
S13	I'm on a trip. The accommodation (*e.g.* hostel, hotel, Airbnb, *etc.*) provides basic cooking facilities (*e.g.* kitchen with microwave).	-0.72±0.08	1.39
S6	I'm too tired to cook.	-1.06±0.08	0.87
S4	I'm hungry and I wanted to eat something quickly.	-1.09±0.08	0.84
S1	I don't have time to cook my own food.	-1.32±0.09	1.01

[1] Item S812 was combined from S8 and S2.

5.6.4 Wright maps

The estimates of respondents and items on the refined constructs were visualised for initial inspection using the Wright maps.

5.6.4.1 Wright maps of satisfaction attitudes related constructs

Figure 5.5 illustrates the Wright maps of the four refined satisfaction attitudes constructs. Similar to that defined in the initial construct (figure 5.1), the respondents located at the bottom of the scale were the least satisfied consumers, whereas the others at the top of the scale were the most satisfied consumers. The items were mapped in the opposite orientation, where the least satisfied properties of ready meals were at the top and the most satisfied properties of ready meals were at the bottom. Apart from the hierarchy order of the items, which implies the ranking order of the degrees of consumer satisfaction towards individual properties of ready meals, a few things could be observed from the Wright maps:

(1) The ceiling or floor effects of the items were found in all constructs, as long tails of respondents could be observed on the Wright maps (figure 5.5). The items spread out in a much narrow range on the scale compared to the respondents. The respondents located on the tails presented a tendency to rate the items using extreme scale categories.

(2) All items from the instrument that measures consumers' satisfaction towards health benefits of ready meals are located on top of the scale, indicating that they were too difficult to be endorsed during the survey. In other words, the consumers disagree that the ready meals can obtain as many health benefits as they want.

(3) The statement[1] of item A21 was highly agreed on by all respondents, as its location on the Wright map was extremely low. This was expected, because this item measures consumers' satisfaction towards the convenience of ready meals associated with minimum requirement of preparation and cooking.

5.6.4.2 Wright maps of product criteria constructs

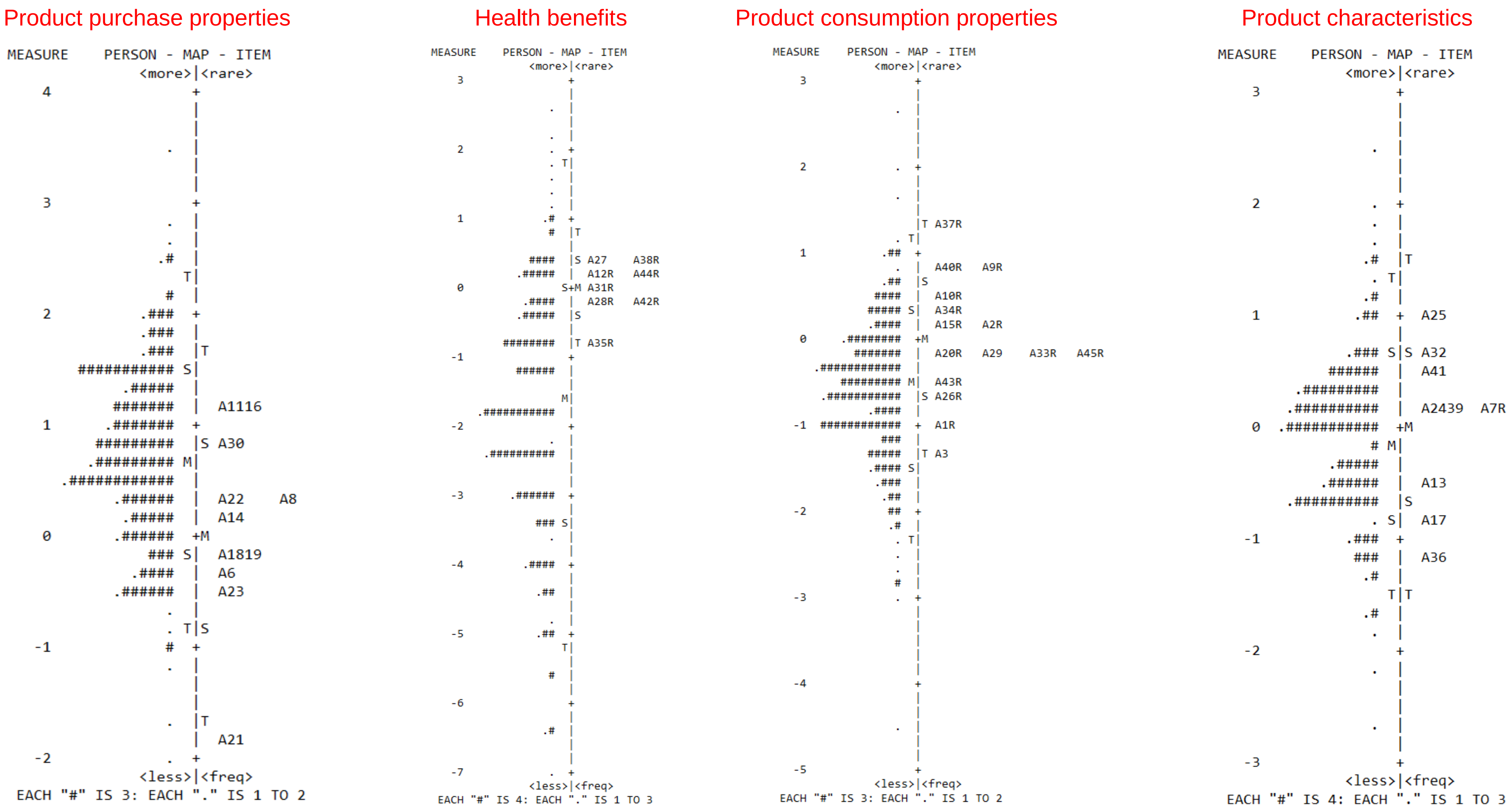
Similarly, the distribution of the respondents were heavily-tailed on the two product criteria constructs as shown in figure 5.6, while no extreme item was evident on the Wright maps. The tails at the top represent the high-consideration consumers, while the tails at the bottom represent the low-consideration consumers.

5.6.4.3 Wright map of consumption situations

A single side of a long tail towards the bottom of the scale was evident in the distribution of respondents on the Wright map of the consumption situations construct, as marked in figure 5.6. These respondents, according to the construct, were those who were not willing to consume ready meal under any situations.

According to the item hierarchy on the scale, the respondents were most willing to consume ready meals under situation S1, when they "don't have time to cook" their own food. The willingness decreased in the order of situations corresponding to their locations on Wright map from bottom to top, until S3, when the respondents invited friends "for lunch, or dinner or having a party".

[1] A21: Ready meals are easy to prepare, even for someone who does not have particular knowledge of cooking.



Product purchase properties

```
MEASURE    PERSON - MAP - ITEM
               <more>|<rare>
    4                +
                     |
                     |
                  .  |
                     |
                     |
    3                +
                  .  |
                  .  |
                 .#  |
                    T|
                  #  |
    2          .###  +
               .###  |
               .###  |T
        ########### S|
             .#####  |
            #######  |  A1116
    1      .#######  +
          #########  |S A30
         .######### M|
      .############  |
            .######  |  A22    A8
             .#####  |  A14
    0       .######  +M
                ### S|  A1819
              .####  |  A6
            .######  |  A23
                  .  |
                  . T|S
   -1             #  +
                  .  |
                     |
                     |
                  .  |T
                     |  A21
   -2             .  +
               <less>|<freq>
  EACH "#" IS 3: EACH "." IS 1 TO 2
```

Health benefits

```
MEASURE    PERSON - MAP - ITEM
               <more>|<rare>
    3                +
                     |
                  .  |
                     |
                  .  |
    2             .  +
                  . T|
                  .  |
                  .  |
                  .  |
    1            .#  +
                  #  |T
                     |
               ####  |S A27    A38R
             .#####  |  A12R   A44R
    0               S+M A31R
              .####  |  A28R   A42R
             .#####  |S
                     |
           ########  |T A35R
   -1                +
             ######  |
                     |
                    M|
       .###########  |
   -2                +
                  .  |
        .##########  |
                     |
                     |
   -3       .######  +
                     |
                ### S|
                  .  |
                     |
   -4         .####  +
                     |
                .##  |
                     |
                  .  |
   -5           .##  +
                    T|
                     |
                  #  |
                     |
   -6                +
                     |
                 .#  |
                     |
                     |
   -7             .  +
               <less>|<freq>
  EACH "#" IS 4: EACH "." IS 1 TO 3
```

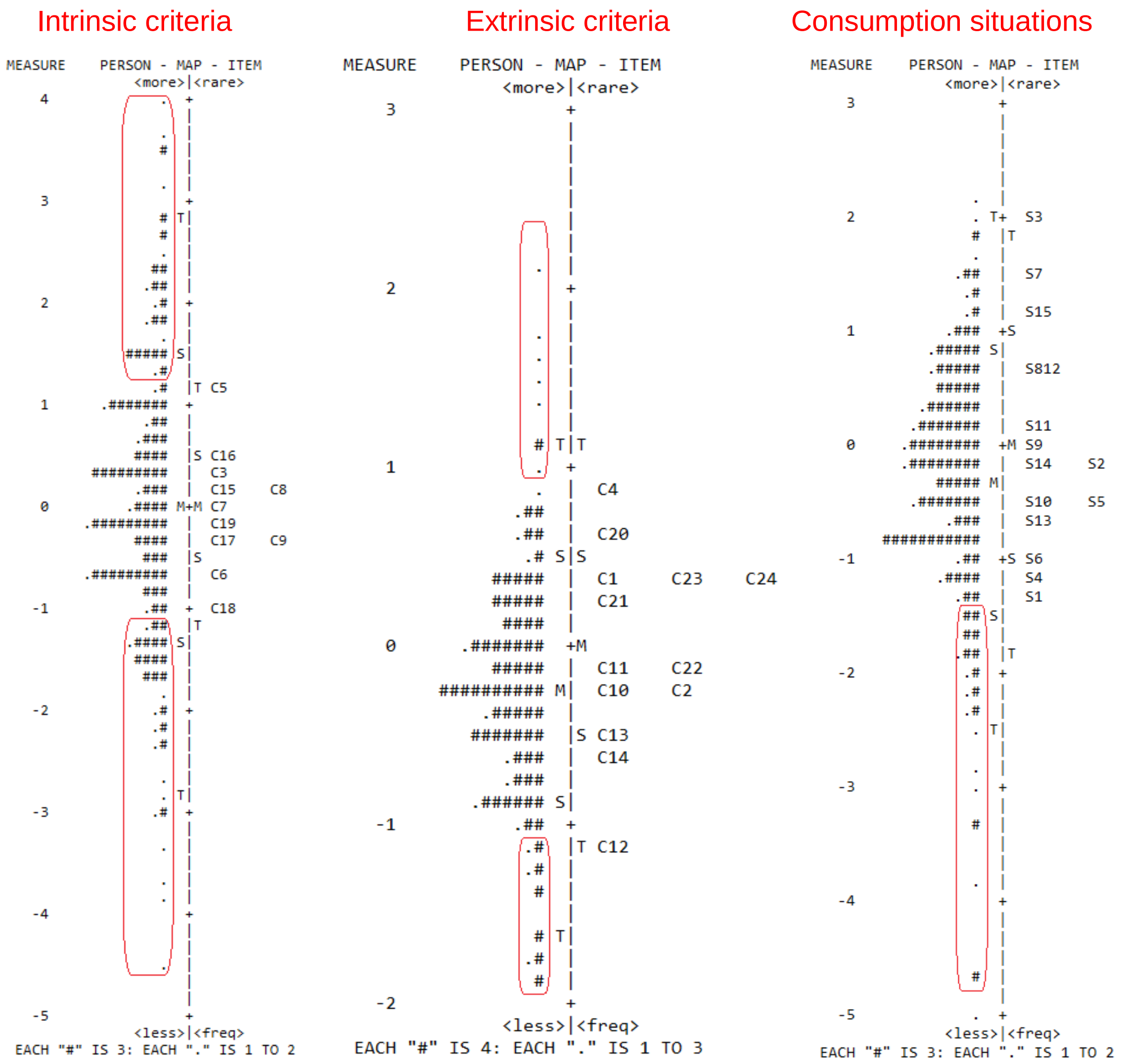
Product consumption properties

```
MEASURE    PERSON - MAP - ITEM
               <more>|<rare>
    3                +
                     |
                  .  |
                     |
                     |
                     |
    2             .  +
                     |
                  .  |
                     |
                     |T A37R
                  . T|
    1           .##  +
                  .  |  A40R   A9R
                .##  |S
               ####  |  A10R
              ##### S|  A34R
              .####  |  A15R   A2R
    0     .########  +M
            #######  |  A20R   A29    A33R   A45R
      .############  |
          ######### M|  A43R
       .###########  |S A26R
              .####  |
   -1  ############  +  A1R
                ###  |
              #####  |T A3
              .#### S|
               .###  |
                .##  |
   -2            ##  +
                 .#  |
                  . T|
                  .  |
                  .  |
                  #  |
   -3             .  +
                     |
                     |
                     |
                     |
                     |
   -4                +
                     |
                     |
                  .  |
                     |
                     |
   -5                +
               <less>|<freq>
  EACH "#" IS 3: EACH "." IS 1 TO 2
```

Product characteristics

```
MEASURE    PERSON - MAP - ITEM
               <more>|<rare>
    3                +
                     |
                     |
                  .  |
                     |
                     |
    2             .  +
                  .  |
                  .  |
                 .#  |T
                  . T|
                 .#  |
    1           .##  +  A25
                     |
               .### S|S A32
             ######  |  A41
         .#########  |
        .##########  |  A2439  A7R
    0  .###########  +M
                  # M|
             .#####  |
            .######  |  A13
        .##########  |S
                  . S|  A17
   -1          .###  +
                ###  |  A36
                 .#  |
                    T|T
                 .#  |
                  .  |
   -2                +
                  .  |
                     |
                     |
                  .  |
                     |
   -3                +
               <less>|<freq>
  EACH "#" IS 4: EACH "." IS 1 TO 3
```

Figure 5.5 Wright maps for the four satisfaction-related constructs



Intrinsic criteria

```
MEASURE    PERSON - MAP - ITEM
                <more>|<rare>
    4                .  +
                        |
                     .  |
                     #  |
                        |
                     .  |
    3                   +
                     # T|
                     #  |
                     .  |
                    ##  |
                   .##  |
    2               .#  +
                   .##  |
                     .  |
                 ##### S|
                    .#  |
                    .#  |T C5
    1         .#######  +
                   .##  |
                  .###  |
                  ####  |S C16
             #########  |  C3
                  .###  |  C15    C8
    0            .#### M+M C7
            .#########  |  C19
                  ####  |  C17    C9
                   ###  |S
            .#########  |  C6
                   ###  |
   -1              .##  +  C18
                   .##  |T
                 .#### S|
                  ####  |
                   ###  |
                     .  |
   -2               .#  +
                    .#  |
                    .#  |
                        |
                     .  |
                     . T|
   -3               .#  +
                        |
                     .  |
                        |
                     .  |
                     .  |
   -4                   +
                        |
                        |
                     .  |
                        |
                        |
   -5                   +
                <less>|<freq>
EACH "#" IS 3: EACH "." IS 1 TO 2
```

Extrinsic criteria

```
MEASURE    PERSON - MAP - ITEM
                 <more>|<rare>
    3                  +
                       |
                       |
                       |
                       |
                       |
                       |
                   .   |
    2                  +
                       |
                   .   |
                   .   |
                   .   |
                   .   |
                       |
                   # T|T
    1              .   +
                   .   |  C4
                 .##   |
                 .##   |  C20
                  .# S|S
               #####   |  C1      C23     C24
               #####   |  C21
                ####   |
    0        .#######  +M
               #####   |  C11     C22
         ########### M|  C10     C2
              .#####   |
             #######   |S C13
                .###   |  C14
                .###   |
             .###### S|
   -1            .##   +
                  .#   |T C12
                  .#   |
                   #   |
                       |
                   # T|
                  .#   |
                   #   |
   -2                  +
                 <less>|<freq>
EACH "#" IS 4: EACH "." IS 1 TO 3
```

Consumption situations

```
MEASURE    PERSON - MAP - ITEM
                 <more>|<rare>
    3                  +
                       |
                       |
                       |
                       |
                   .   |
    2              . T+  S3
                   #   |T
                   .   |
                 .##   |  S7
                  .#   |
                  .#   |  S15
    1           .###   +S
             .##### S|
             .#####    |  S812
              #####    |
            .######    |
          .#######     |  S11
    0    .########     +M S9
         .########     |  S14     S2
              ##### M|
          .#######     |  S10     S5
               .###    |  S13
        ###########    |
   -1           .##    +S S6
             .####     |  S4
                .##    |  S1
                 ## S|
                 ##    |
                .##    |T
   -2            .#    +
                 .#    |
                 .#    |
                  . T|
                       |
                  .    |
   -3             .    +
                       |
                  #    |
                       |
                       |
                  .    |
   -4                  +
                       |
                       |
                       |
                  #    |
                       |
   -5             .    +
                 <less>|<freq>
EACH "#" IS 3: EACH "." IS 1 TO 2
```

Figure 5.6 Wright maps for product criteria constructs and consumption situations construct

The long tails of respondents in the distributions are marked in red.

5.6.5 Reliability

The “real” Rasch separation, strata and reliability indices for respondents and items were estimated within the refined instruments, as presented in table 5.17.

In this study, the strata is more suitable for describing how spread the respondents were located on the scale because the distribution of the person measures were heavily-tailed, which can be observed from the Wright maps (figure 5.5 and 5.6). According to the values of the strata, the respondents might be discerned into two to three levels based on individual instruments about their satisfaction attitudes, three to four levels according to their measures within product criteria instruments, and three levels using the consumption situations instruments. The smallest value of strata was found within the measures of consumers' satisfaction towards product characteristics of ready meals, which was only 1.63. This implied that the respondents could not be well distinguished according to their measures using this instrument, as they were only across one to two statistical levels on the scale.

The indices for item reliability were all quite high, which indicates that the sample size of the study was adequately large to spread out the items along the scale.

Table 5.17 Rasch reliability statistics of the instruments

Instrument	**Respondents**			**Items**		
	Separation	Strata	Reliability	Separation	Strata	Reliability
Satisfaction attitudes						
Product purchase properties	1.56	2.41	0.71	11.46	15.61	0.99
Health benefits	2.05	3.07	0.81	3.37	4.83	0.92
Product consumption properties	1.97	2.96	0.80	8.75	12.00	0.99
Product characteristics	**0.97**	**1.63**	**0.49**	9.38	12.84	0.99
Product criteria						
Intrinsic criteria	2.81	4.08	0.89	7.80	10.73	0.98
Extrinsic criteria	1.76	2.68	0.76	8.50	11.67	0.99
Consumption situations	2.36	3.48	0.85	11.28	15.37	0.99

The lowest observation of reliability statistics are labelled in red.

5.6.6 Consumer segmentation

5.6.6.1 Consumer segmentation based on satisfaction attitudes

(1) Segmentation by Rasch measures

The dendrogram obtained from the DIANA clustering on respondents' measures of the four satisfaction attitudes instruments was displayed in figure 5.7. It is a tree-diagram, which displays the information about the hierarchy relationship between the sub-clusters. The respondents were clustered into four segments according to the height on the dendrogram. However there were only eighteen respondents in the smallest cluster, which was not suitable for comparison. Therefore this cluster was merged into the closest one. Eventually three segments were identified. Table 5.18 depicts the number of respondents in each segment, which were all reasonable large for comparison. The mean measures by segments on each construct were also reported in same table. Both parametric and non-parametric statistical procedures suggested that the respondents in cluster 1 and 2 represented relatively high satisfaction levels towards product purchase properties and product characteristics of ready meals, whereas three relative satisfaction levels (high, medium and low) could be used to describe the mean measures of respondents in the three clusters estimated from their responses to health benefits and product consumption properties of ready meals. These results were consistent with the indications of the strata statistics (table 5.17).

(2) Segmentation by raw mean scores

The respondents were clustered into four segments based on their the raw mean scores of the four satisfaction attitudes instruments, which were more evenly spread across all construct levels compared to the initial four clusters identified based on the Rasch measures. Figure 5.8 illustrates the dendrogram, while table 5.19 depicts the descriptive information of each segment and the results of multiple comparisons by segments.

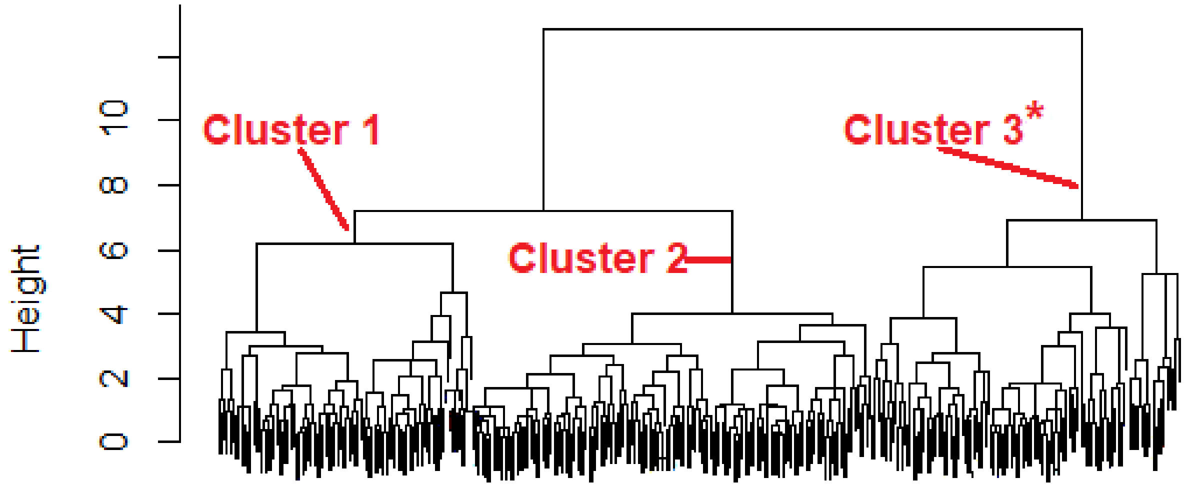


Figure 5.7 Dendrogram of DIANA clustering on respondents’ measures of the four satisfaction attitudes instruments

*Cluster 3 was merged from two initial clusters.

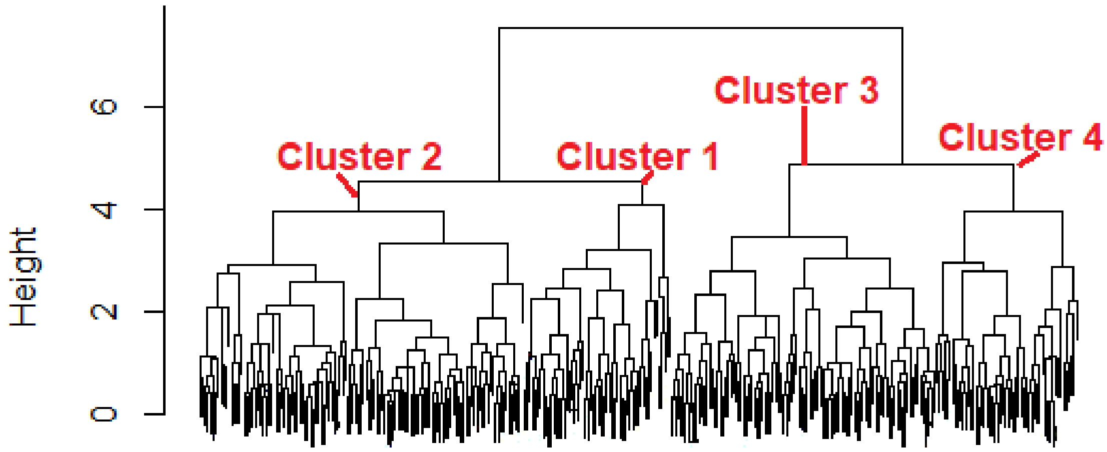


Figure 5.8 Dendrogram of DIANA clustering on respondents’ raw mean scores of the four satisfaction attitudes instruments

Table 5.18 Segmenmtation based on respondents' Rasch measures of the satisfaction attitudes instruments

Segment (cluster)	Number of respondents	Product purchase properties			Health benefits			Product consumption properties			Product characteristics		
		Mean measure	Statistical group[1]	Level	Mean measure	Statistical group[1]	Level	Mean measure	Statistical group[1]	Level	Mean measure	Statistical group[1]	Level
1	88	0.94	a	High	0.37	a	High	0.14	a	High	-0.02	a	High
2	138	0.72	ab	High/Low	-1.49	b	Medium	-0.48	b	Medium	0.03	a	High
3[2]	107	0.55	b	Low	-3.81	c	Low	-1.25	c	Low	-0.31	b	Low

[1] Same results were obtained from Tukey HSD test and Dunn's test with Hochberg correction

[2] The Segment 3 was merged from two initial clusters, including eighteen respondents belonging to the initial cluster 4, who exhibited the lowest satisfaction level towards the ready meal

Table 5.19 Consumer segmenmtation based on resondents' raw mean scores of the satisfaction attitudes instruments

Segment (cluster)	Number of respondents	Product purchase properties			Health benefits			Product consumption properties			Product characteristics		
		Mean score	Statistical group[1]	Level	Mean score	Statistical group[1]	Level	Mean score	Statistical group[1]	Level	Mean score	Statistical group[1]	Level
1	56	5.55	a	High	4.08	a	High	4.79	a	High	4.49	a	High
2	123	4.98	b	Medium	3.35	b	Medium	3.9	b	Medium	4.26	(ab)b[2]	High/Medium
3	100	5.12	b	Medium	2.08	c	Low	3.3	c	Medium/Low	4.06	b	Medium
4	54	3.76	c	Low	2.23	c	Low	2.67	d	Low	3.71	c	Low

[1] The same results were obtained from the Tukey HSD test and the Dunn's test with Hochberg correction

[2] Slightly different results were drawn from the Tukey HSD test and the Dunn's test with Hochberg correction (in bracket)

5.6.6.2 Consumer segmentation based on consumers' degree of consideration on product criteria.

(1) Segmentation by Rasch measures

Same as the segmentaion based on satisfaction attitudes, four initial clusters were discerned by DIANA clustering using respondents' measures from the two product criteria instruments. Figure 5.9 illustrates the dengrogram. Again, a small cluster (thirty eight respondents) was spot. It was combined with the nearest cluster to form a larger one. Three final segments were identified. The same trends on the distribution of respondents' measures on the two product criteria constructs were observed, where the three segments were associated with the high, medium and low levels by clusters on the degree of consideration of the consumers (table 5.20).

(2) Segmentation by raw mean scores

The dendrogram obtained from the DIANA clustering can be seen in figure 5.10. The respondents were divided into 4 segments. Table 5.21 describes the results of multiple comparisons by segments.

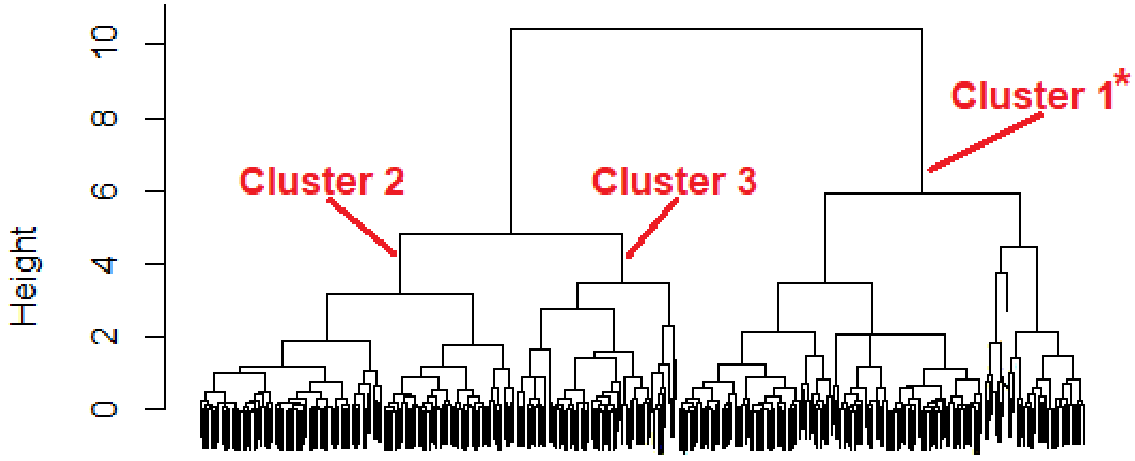


Figure 5.9 Dendrogram of DIANA clustering on respondents' Rasch measures of the product criteria instruments

*Cluster 1 was merged from two initial clusters.

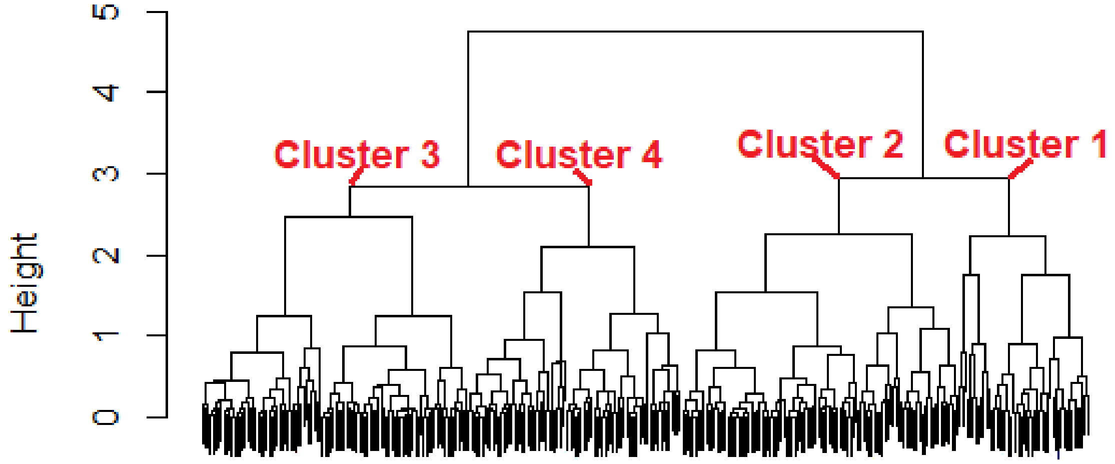


Figure 5.10 Dendrogram of DIANA clustering on respondents' raw mean scores of the product criteria instruments

Table 5.20 Consumer segmenmtation based on respondents' Rasch measures of the product criteria instruments

Segment (cluster)	Number of respondents	Intrinsic criteria			Extrinsic criteria		
		Mean measure	Statistical groups[1]	Level	Mean measure	Statistical groups[1]	Level
1[2]	153	1.24	a	High	0.01	a	High
2	61	-0.45	b	Medium	-0.39	b	Medium
3	119	-1.97	c	Low	-0.52	c	Low

[1] Same results were obtained from Tukey HSD test and Dunn's test with Hochberg correction

[2] The Segment 1 was combined from two clusters, including thirty-eight respondents belonging to initial cluster 4, who exhibited the highest degree of consideration when purchasing ready meals

Table 5.21 Consumer segmenmtation based on respondents' raw mean scores of the product criteria instruments

Segment (cluster)	Number of respondents	Intrinsic criteria			Extrinsic criteria		
		Mean score	Statistical Groups[1]	Level	Mean score	Statistical Groups[1]	Level
1	49	3.17	a	High	2.55	a	High
2	104	2.6	b	Medium	1.78	b	Medium
3	100	1.73	c	Medium	1.8	b	Medium
4	80	0.92	d	Low	1.48	c	Low

[1] Same results were obtained from Tukey HSD test and Dunn's test with Hochberg correction

5.6.6.3 Distribution of gender and age in the segments

In addition, the distribution of gender and age in all segments was demonstrated in tables 5.22 and 5.23. No special pattern was found.

Table 5.22 Gender and Age distribution in consumer segments based on Rasch measures

Construct	Segment	Gender		Age		
		Female	Male	16-24	25-34	35+
Satisfaction attitudes	1	62	28	39	28	21
	2	103	35	65	42	31
	3	82	25	41	36	30
Product criteria[1]	1	120	33	67	52	34
	2	41	20	27	16	18
	3	84	35	51	38	30

[1] The levels are about the degree of consideration

Table 5.23 Gender and Age distribution in consumer segments based on raw mean scores

Construct	Segment	Gender		Age		
		Female	Male	16-24	25-34	35+
Satisfaction attitudes	1	82	41	62	35	26
	2	44	12	24	20	12
	3	78	22	41	28	31
	4	41	13	18	23	13
Product criteria[1]	1	69	31	42	36	22
	2	79	25	41	36	27
	3	41	8	27	15	7
	4	56	24	35	19	26

[1] The levels are about the degree of consideration

5.6.7 Relationship between consumer segments and the consumption frequency of three type of meals

The results of Kruskal-Wallis test followed by Dunn’s test with Hochberg correction were illustrated in figures 5.11 ~ 5.13. The main findings are:

(1) The consumption frequency of ready meals were significantly different between the segments clustered on satisfaction attitudes measures or raw mean scores. The highest consumption frequency was associated with the highest level of the satisfaction attitudes, while the respondents who had lowest Rasch measures or raw mean scores on satisfaction attitudes constructs were likely to consume ready meals least often. This verified the hypothesis proposed when defining the construct (section 5.2.1).

(2) How considerable the respondents are did not affect their consumption frequency of ready meals.

(3) There was no significant difference between respondents’ consumption frequency of restaurant meals or takeaway meals by segments based on Rasch measures. However, an unexpected negative correlation was observed between the consumption frequency of takeaway meals and the segments based on raw mean scores associated with respondents’ degree of consideration when consuming ready meals. This may be a false positive result.

(4) People aged 25-34 are more likely to consume restaurant meals.

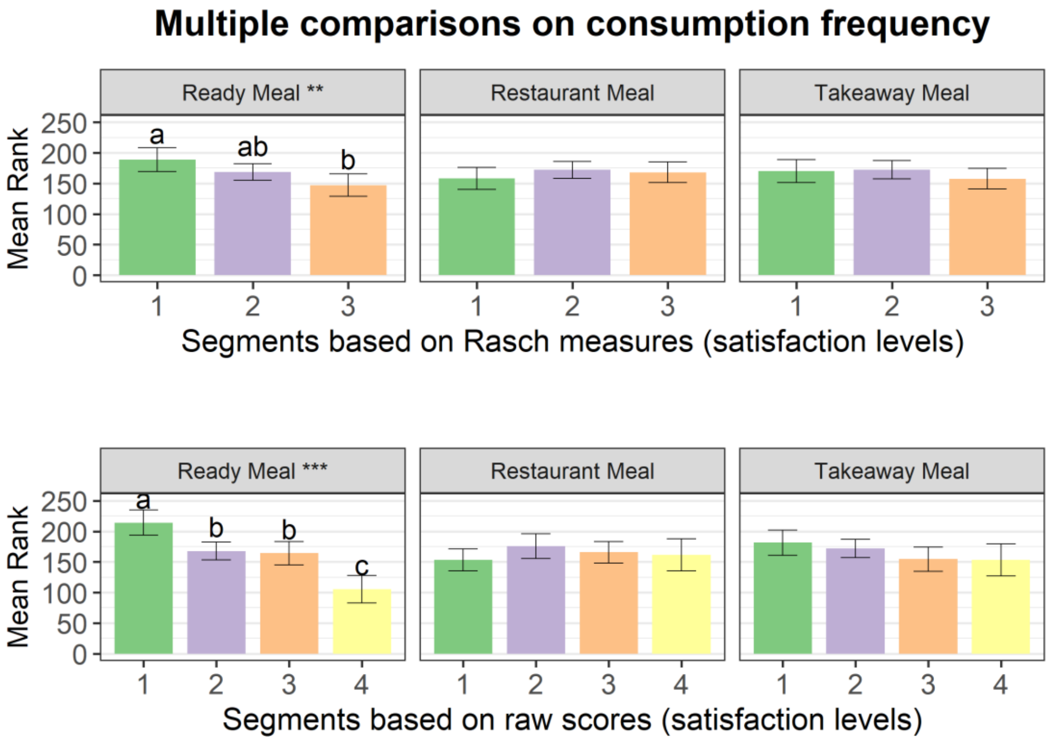


Figure 5.11 Multiple comparisons on consumption frequency of three types of meals by segments based on respondents satisfaction levels

(**=<0.01 and ***=<0.01 for Kruskal-Wallis test, error bars=95% CI)

The segments refer to tables 5.18 and 5.19

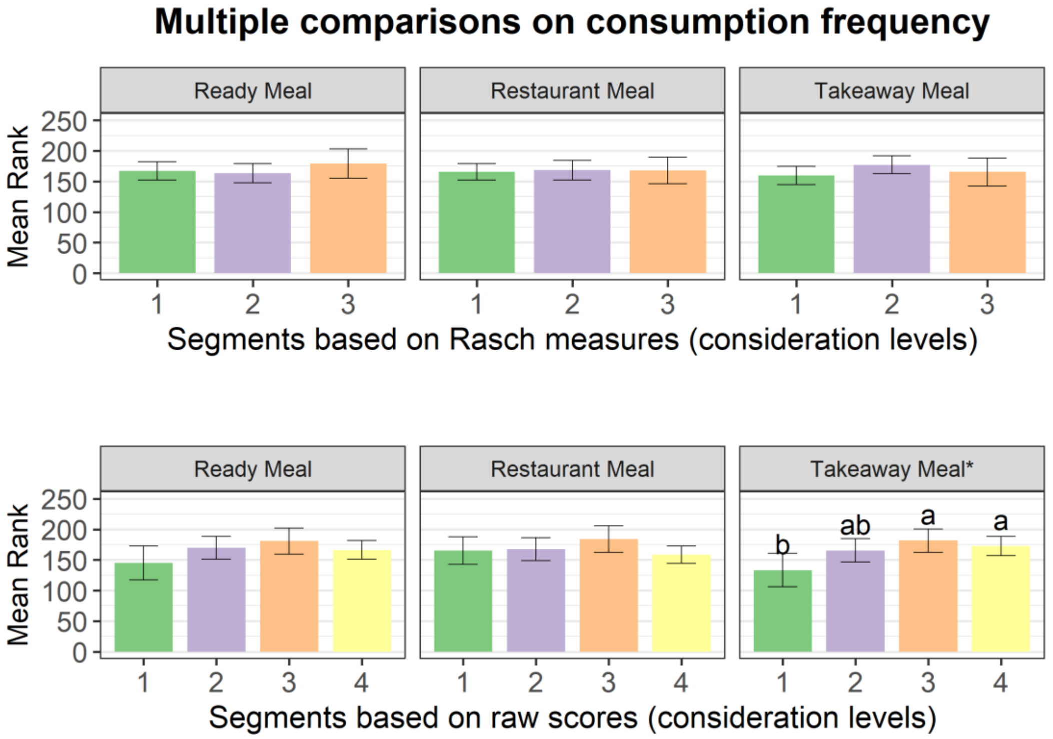


Figure 5.12 Multiple comparisons on consumption frequency of three types of meals by segments based on respondents' consideration levels

(*=<0.05 for Kruskal-Wallis test, error bars=95% CI)

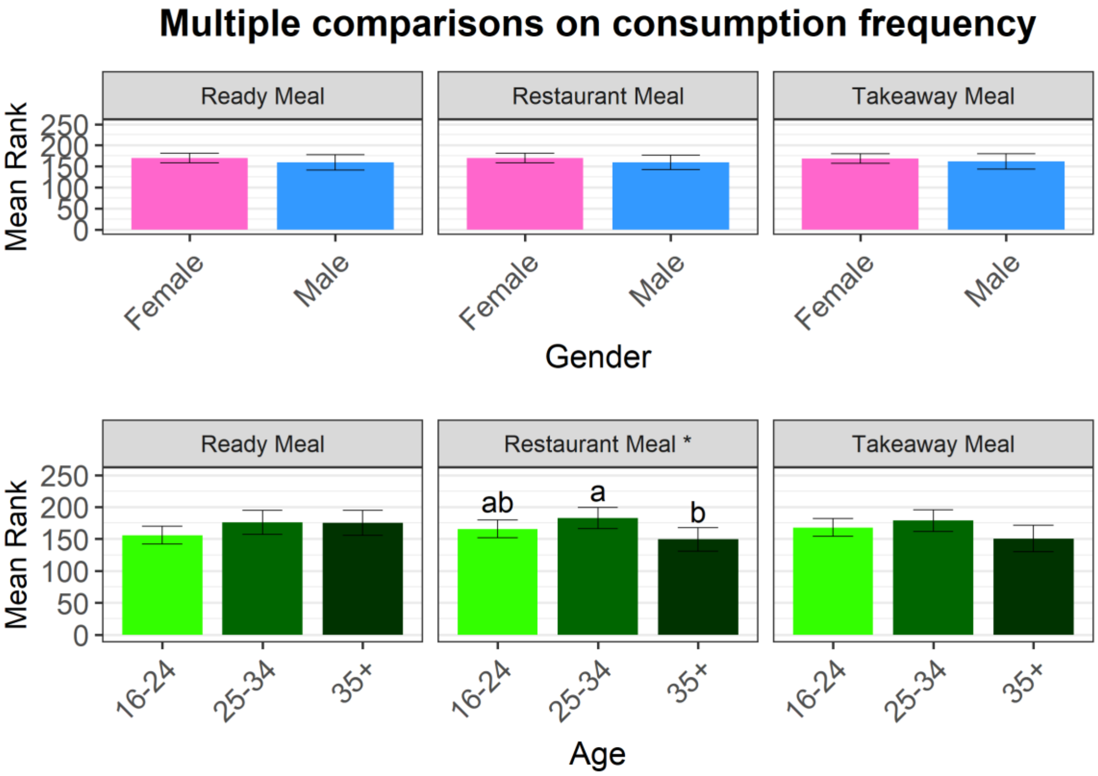


Figure 5.13 Multiple comparisons on consumption frequency of three types of meals by gender or age

(*=<0.05 for Kruskal-Wallis test, error bars=95% CI)

5.7 Discussion

5.7.1 Consumer insights detected from this study

5.7.1.1 Consumers' satisfaction attitudes towards ready meals.

With the help of Rasch analysis, four underlying aspects associated with consumers' satisfaction attitudes towards ready meals were identified. Cluster analysis divided the respondents into three consumer segments, corresponding to different satisfaction levels. The results of multiple comparisons implied that the segmentation by satisfaction levels could be used to predict the consumers purchasing behaviours of ready meals.

In practice, one can use a single instrument to measure consumers' satisfaction towards a particular aspect, or use all four instruments together for segmenting consumers. Also, the hierarchy rank of the item measures can be used as reference in the development or improvement of new ready meal products. For instance, the least satisfied product consumption properties of ready meals identified in this research were:

(1) The taste of ready meals are worse than the expectation made according to the appearance.
(2) Ready meals don't taste fresh.
(3) There is no pleasure in cooking ready meals.

The above descriptions are corresponded to item A37R, A40R and A9R, respectively, which are located on top of the Wright map (figure 5.5). In real development projects, one can treat these deficiencies of current ready meals as development opportunities, generating breakthrough new product ideas to overcome these issues.

5.7.1.2 Consumers' decision making patterns and relative importance of product criteria of ready meals

The results indicated that the consumers' decision making pattern did not affect their consumption frequency on ready meals and the other meals. Therefore, it cannot be used for predicting consumers' behaviour in relation to ready meals.

On the other hand, the results revealed the relative importance of individual product criteria. For example, when selecting ready meals, the ingredients inside the product (item C18), the freshness (item C6), the amount of fat per meal (item C17) were considered as most important intrinsic criteria, while the price (item C12), whether the real appearance of the product can be see through the outer packaging (item C14), and the portion size (item C13) were flagged as the most important extrinsic criteria. In real development projects, these criteria should be considered first. This does not mean the least important criteria such as the amount of protein per meal (item C5) could be ignored completely, but in a time-tight project, the research team should spent most of time on the more important criteria.

In addition, it is of interest to point out that, the familiarity of the ready meal (item C2) was rated as the fourth most important extrinsic criteria, whereas "it is new to me" (item C4) was considered as the least important extrinsic criteria. This implies that when choosing ready meals, consumers might prefer familiar dishes over the completely unknown meal.

5.7.1.3 The willingness of consuming ready meals at particular contextual situations

The results indicated that the consumers would be more willing to consume ready meals when they are short of time or energy. The three particular situations under which the ready meals would be most likely to be consumed are “I don’t have time to cook my own food” (item S1), “I’m hungry and I wanted to eat something quickly” (item S4), and “I’m too tired to cook” (item S6). However, the ready meals were least likely to be consumed if the consumers wanted to enjoy a moment by themselves or with others, as the items refer to the enjoyment had the highest measures on the scale[1], such as “I invited my friends for lunch, or dinner or having a party” (item S3), “I want to enjoy my evening with my family” (item S7), and “I’m in a mood to eat something delicious” (item S15).

Likewise, the super-item S812 “I spot something (ready meal) new to me” could not stimulate consumers’ interest to consume ready meals, as it had relatively high measures compared to other items (on top of the Wright map, see figure 5.6). This result was consistent with the findings within the product criteria instrument that the consumers may tend to choose a ready meal that was familiar to them.

5.7.2 Consumer segmentation – Rasch vs. CTT

The DIANA clustering discerned the respondents into segments based on their Rasch measures or raw mean scores in the same manner across the instruments used for segmentation. However, the memberships of the segmentation were disagreed by Rasch and CTT.

It appears like the segmentation based on Rasch measures is more sensitive to the extreme scores, as a small-size cluster made up of respondents at the tails of sample distribution was identified by both sets of instruments, however further investigation is needed.

[1] The items had highest measures on the scale were conceptualised as the situations under which the ready meals were least likely to be consumed. See figure 5.1.

5.7.3 Rasch analysis provides a number of additional quality control procedures that can govern the development and validation of the instrument.

A major benefit of using Rasch analysis for developing instrument is a number of additional quality control procedures are routinely performed in data analysis (Boone, 2016).

In this study, several item pairs with local item dependence (LID) were identified. Some LID item pairs could be explained. For instance, item A24 asked respondents about their opinions on the attractiveness of the packaging design, which is covered by item A39 about the attractiveness of appearance. The other candidates of LID items were also connected to each in certain extent. The LID might affect the quality of measurement, which had been described in section 2.3.2.1. Minimising LID can increase the accuracy of measurement. Therefore, after reviewing the item statements, some of the LID items were combined to super-items. In addition, LID is associated with the unidimensionality of the instrument, because the LID items can be considered as an additional dimension under the main dimension. As shown in figure 5.2, the assumption of unidimensionality in the item subset concerns consumers’ satisfaction towards product purchase properties could only hold after resolving LID.

Another example is that Rasch analysis obtains fit statistics to examine the performance of individual elements. Fit statistics reflect the degree of divergence between the observation score and expected response. In this study, A32 and C16 were flagged as misfitting items because their outfit MNSQ statistics were greater than recommended range. After removing the extreme unexpected responses, their outfit MNSQ values were brought back to acceptable levels, therefore the quality of the measurement could be improved.

5.7.4 Summary

This case study illustrated an example of applying Rasch analysis in consumer research. The whole study, from the conceptualisation of initial constructs to the validation of final instruments, was guided by Rasch analysis. The benefits of using Rasch analysis in developing measurement support the proposal that it should be used in consumer research in relation to new food product development.

Chapter 6 Case study IV: Application of the Rasch analysis in instrument development and validation for sensory product benchmarking test on beef lasagne ready meal products

6.1 Introduction

The case study II (chapter 4) tested the hypothesis that a composite measure of overall liking modelled with a Many-Facet Rasch Model using attribute ratings can provide better discrimination power on product preference than the single holistic measure modelled using a single overall acceptability item. The products used in study II were selected from eight food and beverage categories, thus only the common sensory modalities and individual attributes were used in the instrument. However, in a consumer benchmarking test, similar type of products and particular attributes related to test objects would be used. Therefore it is necessary to conduct another study to explore the application of Rasch analysis for a benchmarking test on similar product using product-specific attributes as items.

This study used beef lasagne ready meals as test objects. The development of the instrument followed the construct modelling approach (Wilson, 2004). Rasch analysis was applied for governing the development and validation of the instrument. In addition, the results obtained from data analysis could be potentially used in real product development project for beef lasagne ready meals by industry.

6.2 Developing the instrument

6.2.1 Overview of the instrument development

Wilson’s construct modelling approach (Wilson, 2004) was used to define the construct and the basic formats of the instrument. After that, the initial attribute pool was created based on literature and researchers’ consumption experience. Thereafter, a series of one-to-one consumer interviews were conducted. The information collected from these interviews was used to refine the attribute pool. Eventually, the final questionnaire was constructed with twenty attributes selected from the refined attribute pool.

6.2.2 Defining the construct

For the benchmarking test, the construct was defined as panellists' overall liking of the products. It was conceptualised using a construct map (figure 6.1). Three facets (*i.e.* products, panellists and attributes) were involved in the measurement of the construct. Their orientations on the scale were all defined in a positive way as shown in figure 6.1. The higher the measures, the greater degree of overall liking could be reflected.

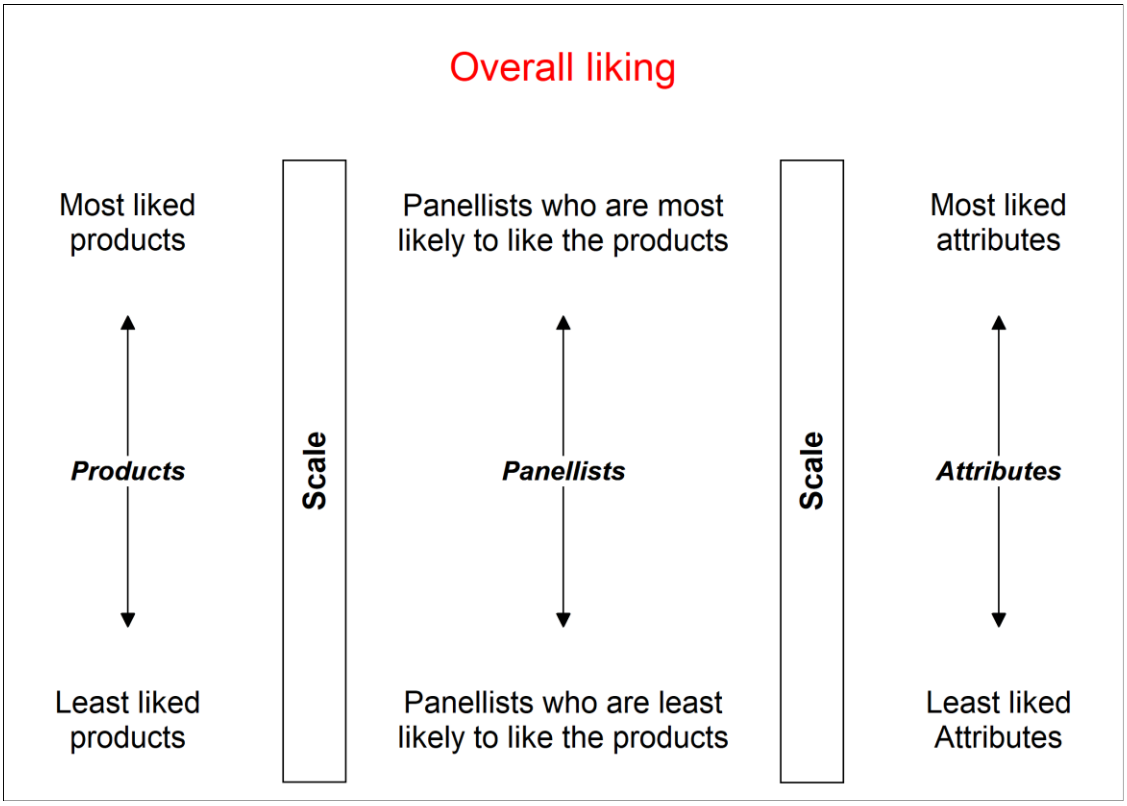


Figure 6.1 Construct and orientation of the measurement elements defined in this study

6.2.3 Defining the item responses and outcome space

In this study, consumers' affective responses to the construct were collected using the industry standard 9-point hedonic scale. The arbitrary raw scores 1-9 were assigned to the categories along the scale from "Dislike extremely" to "Like extremely".

6.2.4 Defining the measurement model

To measure the three facet, a Many-Facet Rasch Rating scale model (MFR-RS model) would be applied. All model parameters were parametrised to positive direction so that the elements on all facets can follow the defined positive orientation.

6.2.5 Selecting distinct products for sensory evaluation

Six commercial beef lasagne ready meal products, which represent three marketing-orientations, were selected as samples. Table 6.1 provides their basic information. They were expected to be quite different on their overall liking. According to the product description on the packaging material, the two premium products were made using high quality ingredients. Therefore they were expected to be the most liked products. The two standard products were expected to hold the middle positions on the scale, while the healthy-eating products were predicted to be least liked products because there were less flavour-rich and aroma-rich ingredients such as meat and cheese inside them for the purpose of decreasing the energy density. Moreover, the same prediction might also be made using the market prices as the second indicators[1]. (Table 6.1).

Table 6.1 6 Beef lasagne ready meal product used in this research

Brand	**Marketing orientation**	**Purchase price per single pack**	**Predicted rank of overall liking**
Tesco	Premium Product[1]	£3.7	Most liked product
Iceland	Premium Product[2]	£2.67	↑
Morrisons	Standard Product[1]	£2.23	
Tesco	Standard Product[2]	£1.5	
Morrisons	Healthy-eating Product[2]	£1	Least liked product
WeightWatchers	Healthy-eating Product[2]	£1	Least liked product

[1] chilled. [2] frozen

[1] The higher the price, the better quality of the product. It should be noted that this is not always true.

6.2.6 Generating the initial attribute pool

The aim of this step was to select a list of sensory attributes of beef lasagne as the initial attribute pool.

The research started from reviewing the literature. Only one previous research (Farley and Reed, 2005) used beef lasagne as samples for sensory evaluation. The searching range of literature was then extended to the main components of beef lasagne such as lasagne sheet, meat sauce, béchamel sauce and cheese topping. Several papers and books concerned with the sensory properties of these components were found (AL‐OBAIDY *et al.*, 1984; Arocas *et al.*, 2010; Larmond and Voisey, 1973; Haraldsson, 2010; Landy *et al.*, 2002; Olivera and Salvadori, 2006).

In addition, several attributes based on researchers' own experience, which were not included by previous research, were added. Eventually, the initial attribute pool (see Appendix D) was made up of seventy-seven attributes, including sixty-three attributes derived from the literature and fourteen additional attributes proposed by the research team. The attributes were classified into four aspects, including aroma (five attributes), appearance (fifteen attributes), taste-flavour (thirty-one attributes) and texture-mouthfeel (26 attributes). The attributes related to the visual perception of texture were categorised into the appearance aspect.

6.2.7 Refining the attribute pool via consumer interviews

The construction of initial attribute pool was purely relied on literature and researchers' own experience. However, it may not be able to cover all attributes of beef lasagne which can be perceived by consumers. Therefore, it should be refined. A series of one-to-one consumer interviews were carried out for this purpose.

The consumer interviews and following benchmarking test were approved by Faculty Research Ethics Committee (MEEC16-020).

Prior to the consumer interviews, a set of attribute cards which had the names of the attributes printed on them were produced based on the initial attribute pool. Figure 6.2 exhibits some examples of the cards. These cards were used in the interviews in order to translate consumers' sensory vocabulary into the standard terms.

The interviews were conducted in a controlled environment simulated to a normal eating context. Each participant attended the one-to-one interview only once either during the lunch time between 12pm-2pm or the dinner time after 5pm. The samples were baked strictly following the cooking instruction printed on the product’s outer packaging. The cooking time was well calculated so that all participants were served the samples right after they were cooked.

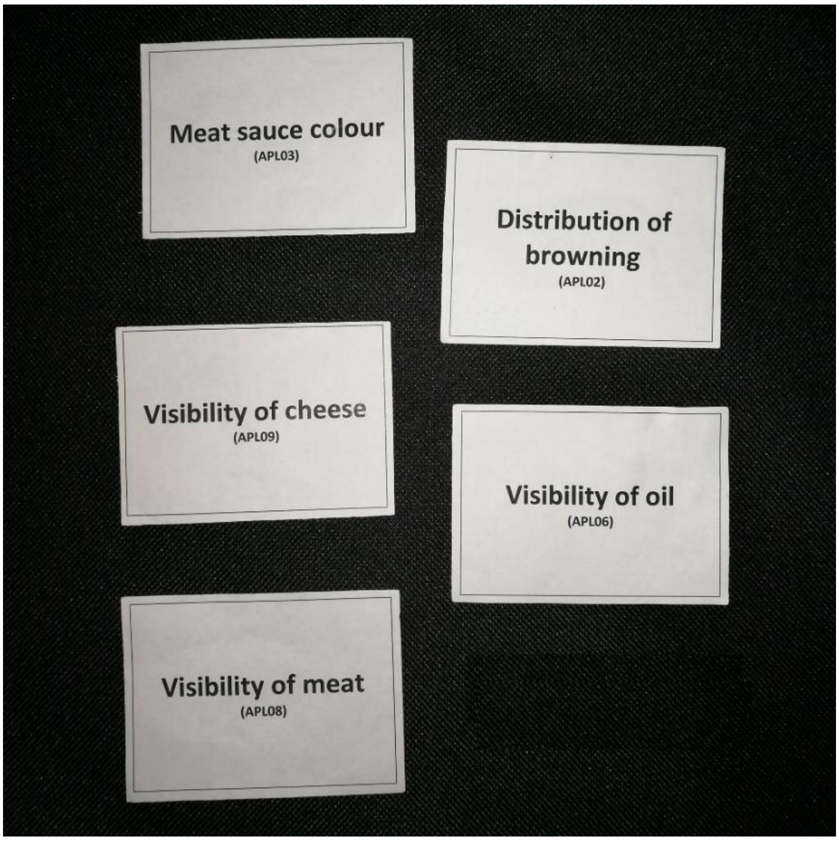


Figure 6.2 Examples of the attribute cards

In the interview, each participant was provided two half-portion of lasagne ready meal for comparison. The examples of lasagne samples can be seen in figure 6.3. The participants were encouraged to describe the similarity or difference that could be perceived between the two samples using their own vocabulary. Once the participants identified a characteristic, the researcher presented one or a few attribute cards to the interviewees according to their description to confirm the standard sensory term that could define the attribute. Extra explanation was provided if the term was not commonly used by the participant who was not familiar with sensory science. By doing this, the consensus of attributes was made. In addition, if the description did not fit to any attribute card from the initial attribute pool, the researcher would record a new term after discussing with the interviewee.

The interviews were semi-structured. The participants were required to compare the samples from the appearance aspect first until they could not elicit any more attribute about appearance. Then the interviews would proceed to the other aspects, in the order of aroma, taste-flavour and texture-mouthfeel.

Forty-five panellists were recruited for the one-to-one interview. They were split evenly into three groups because there were fifteen possible combinations of sample pairs. To minimise the potential bias introduced by unknown differences between production batches, the panellists in each group were presented the samples produced in the same batch. A total number of one hundred sensory attributes were elicited by the panellists, including twenty-four, eighteen, twenty-nine and twenty-nine attributes related to appearance, aroma, taste/flavour and texture/mouthfeel, respectively. They comprised the refined attribute pool. The full list of them can be viewed in the Appendix E.

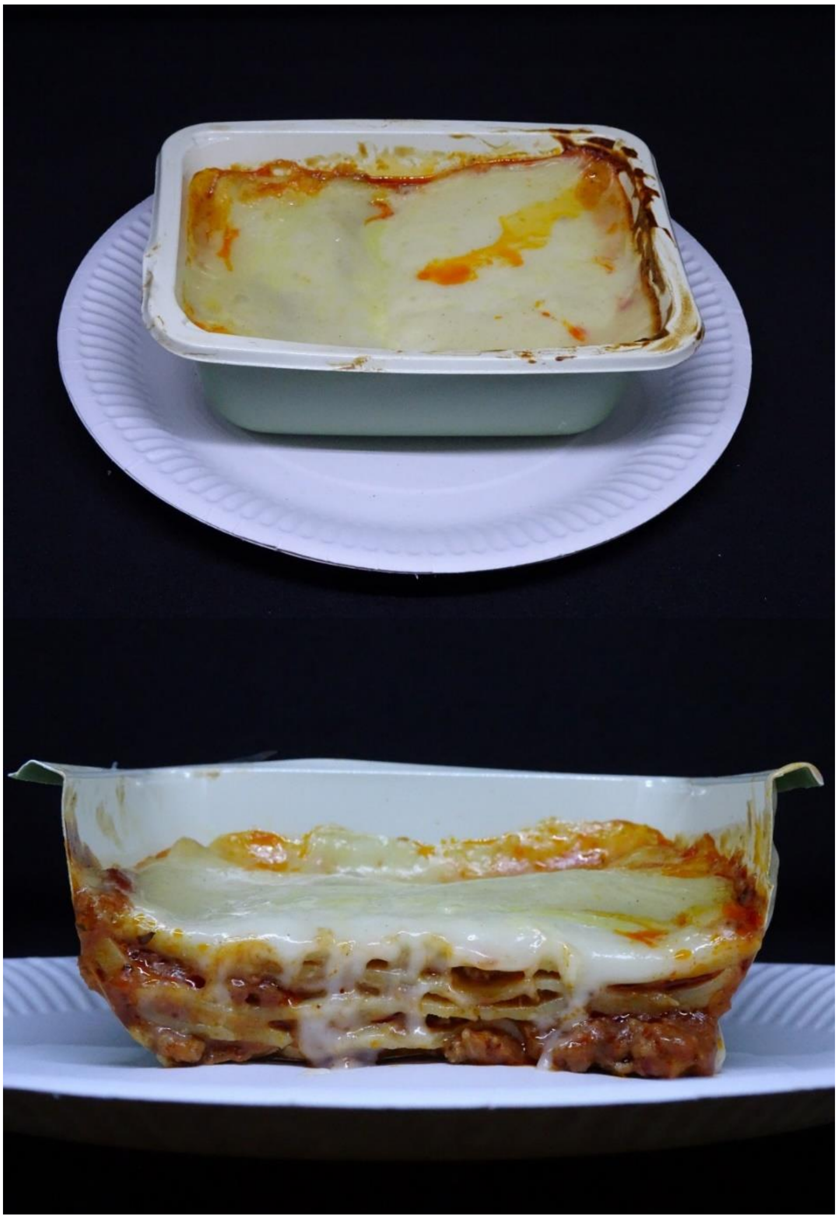


Figure 6.3 Examples of beef lasagne ready meal product tested in this study

6.2.8 Composing of the final questionnaire

The refined attribute pool was reviewed by the research team. Twenty attributes were selected for the instrument according the frequency of them being elicited by the interviewees and their representativeness. Table 6.2 depicts their statements and labels. The final questionnaire started from aroma related attributes because the intensity of aroma has the tendency to decrease quickly. The remaining items were arranged in the order of appearance related attributes, taste-flavour related attributes and texture-mouthfeel related attributes.

Table 6.2 Final questionnaire of the benchmarking test

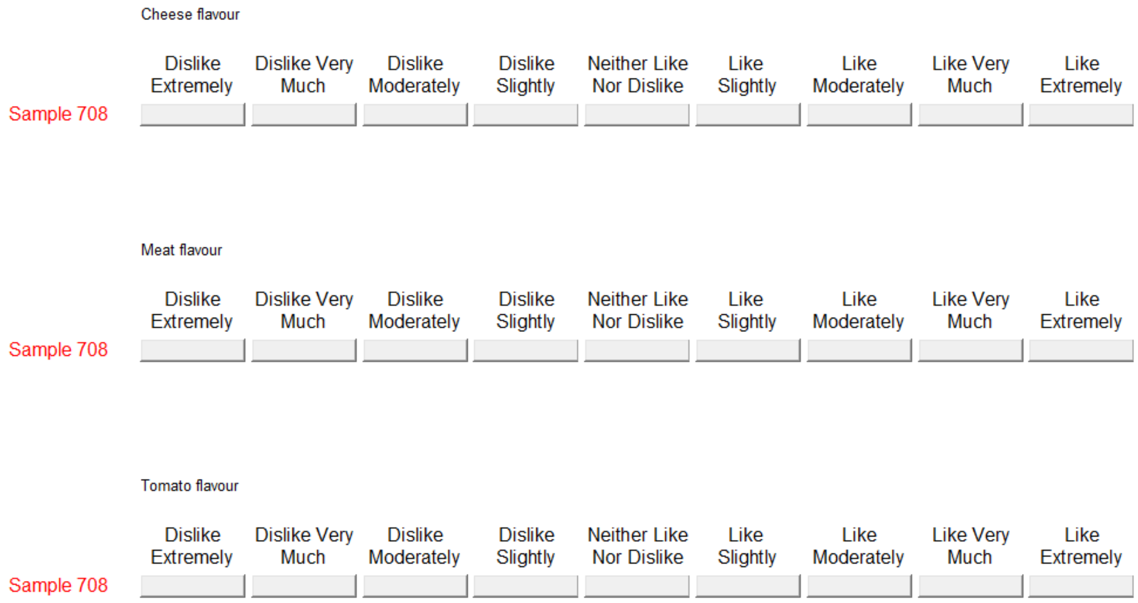
Aspect	Label	Attributes
Aroma	AR1	Cheese aroma
	AR2	Meat aroma
	AR3	Tomato aroma
	AR4	Herb aroma
Appearance	AP1	Amount of oil that you can see on the top surface
	AP2	Proportion of the browning part that you can see on the surface
	AP3	Overall firmness (visual perception)
	AP4	Amount of cheese you that can see
	AP5	Amount of herb you that can see
	AP6	Amount of vegetable chunks you that can see
Taste-flavour	TA1	Cheese flavour
	TA2	Meat flavour
	TA3	Tomato flavour
	TA4	Herb flavour
	TA5	Cream flavour
	TA6	Saltiness
Texture-mouthfeel	TE1	Firmness of pasta (lasagne sheet only)
	TE2	Chewiness of meat (meat only)
	TE3	Thickness of the mixed sauces
	TE4	Creaminess of the mixed sauces

6.3 Procedures of instrument validation using benchmarking test

6.3.1 Experimental work

The benchmarking test was conducted in the same centrally located sensory lab which was used previously for case study II. The data were collected using computer software Compusense 5 (Compusense Inc., 2013). Figure 6.4 shows an example of the computer interface, on which the panellists were required to rate how much they liked or disliked the individual attributes of samples.

In the benchmarking test, the experimental plan was made according to a combination of a Williams Design and incomplete block design (Patterson, 1951; Wakeling and MacFie, 1995; Williams, 1949). It was produced by R package crossdes version 1.1-1 (Sailer, 2013). Every panellist was provided with four samples for evaluation.



Cheese flavour

	Dislike Extremely	Dislike Very Much	Dislike Moderately	Dislike Slightly	Neither Like Nor Dislike	Like Slightly	Like Moderately	Like Very Much	Like Extremely
Sample 708									

Meat flavour

	Dislike Extremely	Dislike Very Much	Dislike Moderately	Dislike Slightly	Neither Like Nor Dislike	Like Slightly	Like Moderately	Like Very Much	Like Extremely
Sample 708									

Tomato flavour

	Dislike Extremely	Dislike Very Much	Dislike Moderately	Dislike Slightly	Neither Like Nor Dislike	Like Slightly	Like Moderately	Like Very Much	Like Extremely
Sample 708									

Figure 6.4 An example of the computer interface of the benchmarking test

6.3.2 Rasch analysis

The data were fitted in the MFR-RS model using Facets (Linacre, 2014a). After that, it was examined using the same procedure used in study II. The analysis started from evaluating the rating scale category effectiveness. The criteria proposed by Linacre (2002a; 2006) was used, which can be seen in section 2.3.1.2. If the rating scale category effectiveness was not satisfied, then the scale would be revised. If the proper functioning of rating scale could be ensured, then the assumption of unidimensionality and local item independence of the tested attributes would be inspected. Following that, the global model fit and individual fit were evaluated. The measures of elements of the three facets, were visualised on the Wright map for initial comparison. They were all parameterised towards positive direction.

6.3.3 Statistical analysis

The reliability statistics of each facet were obtained by Facet. Chi-square tests for fixed effect and random effect were also conducted using Facet.

To compare the overall liking of the six products, the estimates of each replicate of product overall liking were modelled using the method proposed by Ho (2019). The procedure had been described in section 4.4.4. Both parametric ANOVA and non-parametric Kruskal-Wallis test were employed to examine the difference between the mean values of the estimates by products.

Residual analysis was conducted for checking ANOVA assumption. Shapiro-Wilk test (Shapiro and Wilk, 1965) and Kolmogorov-Smirnov test with Lilliefors correction (Kolmogorov, 1933; Lilliefors, 1967; Smirnov, 1948) were conducted for the evaluation of normality of standardised residuals. For the homogeneity of variance, both parametric Levene’s test (Levene, 1960) and non-parametric Brown-Forsythe test (Brown and Forsythe, 1974) were employed. Furthermore, the Bonferroni outlier test was conducted on the studentised residuals.

If the ANOVA assumptions could hold, then the Tukey-HSD test (Tukey, 1949) would be used for post-hoc analysis. If they were slightly violated, then dropping a few extreme outliers might solve the problem. In addition, the non-parametric Dunn’s test with Hochberg correction (Hochberg, 1988) was also applied for multiple comparisons.

6.4 Results

Ninety-six panellists were recruited for the benchmarking test. Each sample was rated sixty-four times in total.

6.4.1 Rating scale category effectiveness

Most of criteria[2] proposed by Linacre (2002a; 2006) such as "*at least 10 observations of each category*", and "*average measures advance monotonically with category*" were satisfied. The outfit MNSQ statistics of the categories were acceptable as none of them located greater than 2.0. However, the results of the t-test between the Rasch-Andrich thresholds suggested that there was no significant difference between some of the adjacent ones (table 6.4), indicating that the intervals on the scale covered by those categories were too narrow. To improve the category effectiveness, several trials of combining scale categories had been conducted. The best solution was collapsing the 9-point scale into a 4-point scale in the way described in table 6.4. The optimised scale structure could meet all criteria suggested by Linacre (2002a; 2006). The probability curves of the collapsed categories on the scale are illustrated in figure 6.5, which shows clear boundaries between the categories.

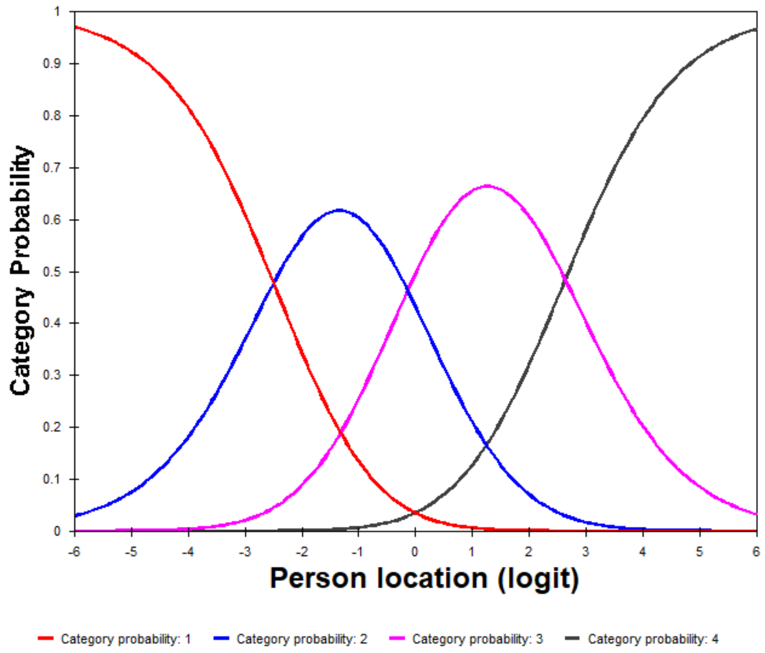


Figure 6.5 Category probability plot of the collapsed 4-point scale

[2] See section 2.3.1.2 for more details

Table 6.3 Rasch-Andrich thresholds of the original 9-point scale

Scale category	**Raw Score**	**Obs**[1]	**Exp**[2]	**Rasch-Andrich threshold ± SE**	**Outfit MNSQ**
Original 9-point scale					
Dislike extremely	1	-0.46	-0.36	NA	0.8
Dislike very much	2	-0.24	-0.28	-1.33±0.09	1.2
Dislike moderately	3	-0.17	-0.19	-0.66±0.05*	1.1
Dislike slightly	4	-0.05	-0.09	-0.67±0.04*	1.2
Neither like nor dislike	5	-0.01	0.01	-0.30±0.03**	1.0
Like slightly	6	0.08	0.11	-0.27±0.03**	1.0
Like moderately	7	0.21	0.22	0.32±0.03	1.0
Like very much	8	0.36	0.33	0.59±0.03	0.9
Like extremely	9	0.45	0.43	2.32±0.08	1.0
Collapsed 4-point scale					
Dislike extremely	1	-0.31	-0.11	NA	0.9
Dislike somewhat *(Dislike very much + Dislike moderately + Dislike slightly)*	2	0.47	0.42	-2.50±0.09	1.1
Like somewhat *(Neither like nor dislike + Like slightly + Like moderately)*	3	1.01	1.04	-0.14±0.03	1.0
Like very much *(Like very much + Like extremely)*	4	1.71	1.67	2.63±0.03	1.0

[1] Modelled average measure in logits.

[2] Expected average measure if data fitted the model.

* and **: There was no significant difference between the two thresholds labelled with * and between the two thresholds labelled with **.

6.4.2 Tests of unidimensionality and local item independence

The Panellist and Product facets were combined to create a new Panellist-by-Product facet, so that the PCA on standardised residual (PCAR) can be applied to test the assumption of unidimensionality of Attribute facet. The result of PCAR with the revised two-facet model[3] showed that the eigenvalue of the raw unexplained variance of the first extracted contrast (2.4883) had a strength less than 3 items, which was not large enough to form an additional dimension (table 6.4). Therefore the Attribute facet might be unidimensional.

The results of evaluation of item fit also indicated that the measurement might be unidimensional. Only one attribute item (AP1) in this two-facet model showed misfit. Its outfit MNSQ statistic (1.51) was just above than the recommended range of 0.50~1.50, suggesting underfit. Although it might not be productive, it would not degrade the measurement if the value of outfit MNSQ was less than 2.0 (Wright and Linacre, 1994).

However, the independent t-tests on the estimates individually modelled using the attribute items with high positive loading (≥0.30) and negative loading (≤-0.30) on the first residual component did not support the assumption of unidimensionality. 14.58% of the t-tests were significant at 0.05 level, while the estimated lower bound of binomial CI was 11.29%. Both of them were greater than the rule of thumb “<5%” (Smith, 2002; Tennant and Conaghan, 2007).

After that, the disattenuated correlation coefficients between the three item clusters on the first extracted contrast were reviewed. The results supported the unidimensionality of Attribute facet. The disattenuated correlation between the cluster 1 that consisted of items with highest loadings on the contrast and the cluster 3 composed by the items with lowest loadings on the same contrast was 0.78. It was greater than 0.71, which was the lower bound suggested by Linacre (2014d) that the two item clusters measures were more likely to be dependent. Moreover, the disattenuated correlations between cluster 1 and 2, and between cluster 2 and 3 were both higher than 0.87, which was the cut-off point used by Linacre (2014d) as an evidence of the two item clusters definitely measuring the same thing.

Further inspection was done on the local item independence. The residual correlations of a few item pairs were slightly higher than 0.30, implying that they might be dependent on each other. This might be the reason that contradictory suggestions about the unidimensionality of attribute items were given by different indicators. Table 6.5 depicts the problematic attribute item pairs. These attributes

[3] *i.e.* Panellist-by-Product facet and Attribute facet

were different between each other of their meanings, although they might be slightly related in pairs. For instance, panellists' visual perception on the overall firmness may be partly implied by the proportion of the browning part. But the proportion of the browning part itself mainly concerned the degree of pre-baking operation in production. Since they all represented unique sensory characteristics of beef lasagne ready meals, they should neither be combined to form a "super-item" nor be removed for the purpose of eliminating local item dependence issue.

Taking everything together into account, although the independent t-tests showed a slight violation of the assumption of unidimensionality, the instrument was still considered to be unidimensional according to the other indicators.

Table 6.4 Tests of unidimensionality

PCA on standardised model residuals (Attributes)	
Raw unexplained variance (total) - 1st contrast	Eigenvalue: 2.4883
Pair of clusters	**Disattenuated Correlation**
1~3	0.7768
1~2	0.9555
2~3	0.9158
Item fit (outfit MNSQ)[1]	0.66-1.51
Misfitting item (and its outfit MNSQ value)	AP1 (1.51)
Independent t-test	
Proportion of significant t statistics	14.58%
Lower bound of binomial CI	11.29%

[1]The outfit MNSQ statistics used in this table were estimated with the two-facet model for the purpose of testing assumption of dimensionality, which were different with those estimated with the three-facet model.

Table 6.5 The item pairs that exhibited potential LID

Item 1	**Item 2**	**Residual correlation**
Tomato aroma	Herb aroma	0.35
Proportion of the browning part that you can see on the surface	Overall firmness (visual perception)	0.32
Meat aroma	Tomato aroma	0.31
Thickness of the sauce	Creaminess of the sauce	0.30

6.4.3 Global model fit

The global model fit was inspected by counting the proportion of extreme standardised residuals of the original three-facet model. Table 6.6 depicts the details of the results. The global model fit was satisfied according to the criteria suggested by (Linacre, 2014b), as less than 5% of the absolute standardised residuals were equal to or greater than 2 and less than 1% of that were equal to or greater than 3.

Table 6.6 Counts and proportion of extreme residuals

Total Responses	**Absolute value of standardised residuals ≥2**		**Absolute value of standardised residuals ≥3**	
	Count	Proportion	Count	Proportion
7680	348	4.53%	14	0.18%

6.4.4 Test of individual fit

Individual fit was examined for the three-facet model. The number of misfitting panellists suggested by checking the outfit MNSQ statistics were summarised in table 6.7. The strong level of underfit (outfit MNSQ>2) was identified with only one panellist, which would not be likely to have a great impact to the measurement.

Table 6.7 Misfitting panellists

Outfit MNSQ	**Count[1]**	**Proportion**
>2	1	1.04%
1.5~2.0	9	9.38%
<0.5	6	6.25%

[1] The total number of panellists is 96.

All products and attributes fitted well in the model, as their outfit MNSQ statistics were all between the recommended range of 0.5~1.5 (Wright and Linacre, 1994). Tables 6.8 and 6.9 depict the outfit MNSQ statistics for Product and Attribute facet, respectively. In addition, the measures of each elements were also exhibited in two tables, in an order from highest measure to lowest measure.

Table 6.8 Measures and outfit MNSQ statistics of products

Product	**Measure±SE**	**Outfit MNSQ**
Tesco chilled premium	1.62±0.05	1.04
Iceland frozen premium	1.50±0.05	1.12
Morrisons chilled standard	1.26±0.05	0.93
Tesco frozen standard	0.94±0.05	0.87
Morrisons frozen healthy-eating	0.37±0.05	1.00
WeightWatchers frozen healthy-eating	0.12±0.05	1.06

Table 6.9 Measure and outfit MNSQ statistics of attributes

Attribute[1]	**Measure±SE**	**Outfit MNSQ**
AR1	0.56±0.09	1.20
TA2	0.20±0.09	0.89
AR3	0.19±0.09	1.28
TE2	0.16±0.09	0.80
AP4	0.13±0.09	1.15
TA3	0.10±0.09	1.02
TE3	0.10±0.09	1.02
TA6	0.06±0.09	1.08
TE1	0.05±0.09	1.13
AR4	0.04±0.09	0.99
AP5	0.04±0.09	0.82
AR2	0.02±0.09	0.97
AP3	-0.04±0.09	0.94
TA4	-0.04±0.09	1.02
TE4	-0.06±0.08	1.32
AP1	-0.09±0.08	0.79
TA1	-0.17±0.08	0.89
AP2	-0.31±0.08	0.98
TA5	-0.34±0.08	1.29
AP6	-0.60±0.08	0.98

[1] The full names of the attributes can be found in table 6.2.

6.4.5 Wright maps

The distribution of the elements were visualised on the Wright map (figure 6.6). A clear hierarchical order could be observed in all facets.

(1) Product facet

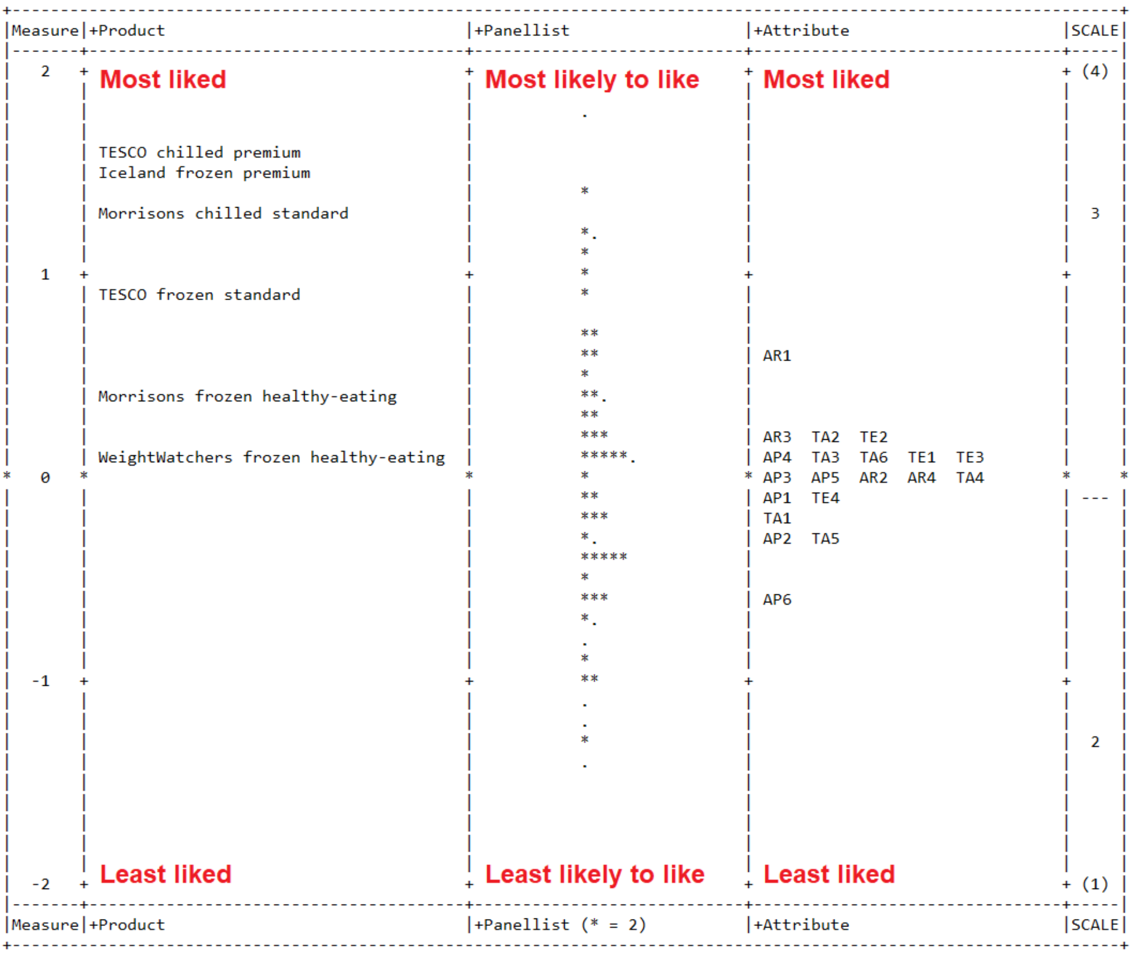
The locations of the two premium products on the scale were quite close, implying that they might be liked by the panellists at similar degree. They were the most liked product, as they were on the top of the map. The two standard products were less liked than the premium ones, while the two healthy-eating products were the least liked samples.

(2) Panellist facet

The panellists were widely spread on the scale, implying that distinctive differences could be identified between panellists on their overall liking to the products. The distribution of panellists were heavily-tailed, therefore when analysing the reliability statistics, the strata statistic would be more appropriate for describing the dispersion of the panellists on the scale than separation statistic, because the extreme levels of measures would be taken into consideration in the computation of strata.

(3) Attribute facet

The estimates of two attributes AR1 and AP6 were clearly separated from the other attributes, where AR1 was the most liked attributes and AP6 was the least liked attributes. Therefore, when describing the dispersion of the attributes on the scale, strata would be inspected, too.



```
+-------------------------------------------------------------------------------------------------------------------+
|Measure|+Product                               |+Panellist                  |+Attribute                      |SCALE|
|-------+---------------------------------------+----------------------------+--------------------------------+-----|
|   2   + Most liked                            + Most likely to like        + Most liked                     + (4) |
|       |                                       |                            |                                |     |
|       |                                       |           .                |                                |     |
|       |                                       |                            |                                |     |
|       | TESCO chilled premium                 |                            |                                |     |
|       | Iceland frozen premium                |                            |                                |     |
|       |                                       |           *                |                                |     |
|       | Morrisons chilled standard            |                            |                                |  3  |
|       |                                       |           *.               |                                |     |
|       |                                       |           *                |                                |     |
|   1   +                                       +           *                +                                +     |
|       | TESCO frozen standard                 |           *                |                                |     |
|       |                                       |                            |                                |     |
|       |                                       |           **               |                                |     |
|       |                                       |           **               | AR1                            |     |
|       |                                       |           *                |                                |     |
|       | Morrisons frozen healthy-eating       |           **.              |                                |     |
|       |                                       |           **               |                                |     |
|       |                                       |           ***              | AR3  TA2  TE2                  |     |
|       | WeightWatchers frozen healthy-eating  |           *****.           | AP4  TA3  TA6  TE1  TE3        |     |
*   0   *                                       *           *                * AP3  AP5  AR2  AR4  TA4        *     *
|       |                                       |           **               | AP1  TE4                       | --- |
|       |                                       |           ***              | TA1                            |     |
|       |                                       |           *.               | AP2  TA5                       |     |
|       |                                       |           *****            |                                |     |
|       |                                       |           *                |                                |     |
|       |                                       |           ***              | AP6                            |     |
|       |                                       |           *.               |                                |     |
|       |                                       |           .                |                                |     |
|       |                                       |           *                |                                |     |
|  -1   +                                       +           **               +                                +     |
|       |                                       |           .                |                                |     |
|       |                                       |           .                |                                |     |
|       |                                       |           *                |                                |  2  |
|       |                                       |           .                |                                |     |
|       |                                       |                            |                                |     |
|       |                                       |                            |                                |     |
|       |                                       |                            |                                |     |
|       |                                       |                            |                                |     |
|       |                                       |                            |                                |     |
|  -2   + Least liked                           + Least likely to like       + Least liked                    + (1) |
|-------+---------------------------------------+----------------------------+--------------------------------+-----|
|Measure|+Product                               |+Panellist (* = 2)          |+Attribute                      |SCALE|
+-------------------------------------------------------------------------------------------------------------------+
```

Figure 6.6 Wright map of the benchmarking test

6.4.6 Reliability

The Rasch reliability statistics of the three facets are tabulated in table 6.10.

Firstly, the Product facet had the greatest reliability. This was expected because distinct products were selected for test objects as part of research design.

Secondly, according to the strata, four statistical distinguishable levels can be identified in the Panellist facet.

Thirdly, although the strata of the Attribute facet was less than that of Panellist facet and Product facet, it was still acceptable as the value of it (3.68) was above the recommended threshold 2.33 given by Tennant and Conaghan (2007), implying that the attributes were spread out on the scale for at least three statistical levels. It also implies that the sample size of panellists and the selection of products were adequacy.

Table 6.10 Rasch reliability statistics of the three facets

Facet	Separation	Strata	Reliability
Panellist	3.17	4.57	0.91
Product	12.89	17.52	0.99
Attribute	2.51	3.68	0.86

6.4.7 Fixed effect and random effect

The results of chi-square tests for evaluating the fixed effect and random effect were reported in table 6.11. The significant p-values of chi-square test for fixed effect further confirmed the heterogeneity of the estimates in all three facets, while all elements can be considered as randomly sampled from a normal distributed population because the p-values for Random effect were great than 0.05.

Table 6.11 Chi-square tests for fixed effect and random effect

Facet	Fixed effect			Random effect		
	Chi-square	d.f.	P value	Chi-square	d.f.	P value
Panellist	1213.5	95	0.00	88.3	94	0.65
Product	875.6	5	0.00	5.0	4	0.29
Attribute	148.8	19	0.00	16.9	18	0.53

6.4.8 Multiple comparisons on product preference

Multiple comparisons were conducted on the sixty-four replicated overall liking estimates of each product using the refitted MFR-RS model. The results of ANOVA and Kruskal-Wallis test were tabulated in table 6.12, as well as the results of residual analysis for ANOVA assumptions. Both ANOVA and Kruskal-Wallis test suggested that there was significant difference between the products on their overall liking, which was consistent with the results of chi-square test for fixed effect.

6.4.8.1 Residual analysis

Firstly, the p-value of Shapiro-Wilk test on the standardised residuals was smaller than 0.001, implying that the assumption of normal distribution of the residuals could not hold with the full data set. However, the opposite suggestion could be drawn from the Kolmogorov-Smirnov test with Lilliefors correction. So whether the residuals were normally distributed was questionable. Secondly, both parametric Levene's test and non-parametric Brown-Forsythe test confirmed the homogeneity of the variances. Finally, the Bonferroni outlier test on the studentised residuals indicated that two observations were extreme outliers.

After dropping the two extreme outliers, the p-value of Shapiro-Wilk test was increased from less than 0.001 to 0.031. It was still significant, although the value was close to 0.05. On the other hand, the p-value of the Kolmogorov-Smirnov test with Lilliefors correction was still insignificant. Therefore no conclusion about the assumption of normality of residuals had been drawn. Secondly, the statistics of tests for homogeneity of variance and the consistency of variance on the data set without the two extreme outliers were similar with the full data set. Lastly, no extreme outliers could be found using Bonferroni outlier test after dropping the two outliers.

6.4.8.2 Multiple comparisons using both parametric and non-parametric procedure

Since the assumption of normality of residuals cannot be confirmed by statistical tests. Both parametric Tukey HSD test and Non-parametric Dunn's test with Hochberg correction were applied. They obtained the same results (table 6.13), thus there is no need to worry about the residual assumptions. The Tesco chilled premium product, Iceland frozen premium product and Morrisons chilled standard product were the most liked products, followed by Tesco frozen standard product. The two healthy-eating products were the least liked products.

Table 6.12 Results for ANOVA, Kruskal-Wallis test and residual analysis

	Test	Criteria	Full data	After dropping Obs 183	After dropping Obs 183 and Obs 116
Residual analysis					
Normality	Shaprio-Wilk test	p-value	<0.001	0.027	0.031
	Kolmogorov-Smirnov test with Lilliefors correction	p-value	0.089	0.090	0.075
Homogeneity of variance	Brown-Forsythe test	p-value	0.262	0.295	0.278
	Levene's test	p-value	0.226	0.250	0.224
Outlier	Bonferroni Outlier test	Extreme outlier	Obs 183 & Obs 116	Obs 116	NA
		Bonferroni adjusted p-value	0.003 (Obs 183) & 0.025 (Obs 116)	0.016	NA
ANOVA		p-value	<0.001	<0.001	<0.001
Kruskal-Wallis test		p-value	<0.001		

“Obs” represents the particular data point

Table 6.13 Multiple comparisons between products

Beef lasagne ready meal product	**Tukey HSD test on full data set**	**Tukey HSD test on revised data[1]**	**Dunn's test with Hochberg correction**
Tesco chilled premium	a	a	a
Iceland frozen premium	a	a	a
Morrisons chilled standard	ab	ab	ab
Tesco frozen standard	b	b	b
Morrisons frozen healthy-eating	c	c	c
WeightWatchers frozen healthy-eating	c	c	c

[1] After dropping the two extreme outlier obs183 and obs116.

6.5 Discussion

An instrument made up of specific sensory attributes in relation to beef lasagne ready meal products had been developed and validated in this study under the guidance of Rasch analysis. The results of benchmarking test were consistent with the expected rank order of product overall liking, suggesting the instrument is appropriate for measuring consumers' overall liking on beef lasagne ready meal products. It has a potential to be used by industry for product testing and consumer segmentation purpose.

Rasch analysis expects all measurement elements (*e.g.* panellists, products and attributes) in the same facet to follow a hierarchical probabilistic pattern (see section 1.4.1). In other words, the panellist would have more chance to like an attribute than any other attributes which were more difficult to be liked. Therefore the degree of the liking over the attributes can be compared using the estimates of attributes. In this study, item AP6 was the least liked attribute, followed by the AP2 (proportion of the browning part that you can see on the surface) and TA5 (cream flavour). These should be improved first if one plan to develop a new beef lasagne ready meals.

In addition to that, the residual correlation implied that three item pairs exhibited LID. This obtained additional information for developing new instrument in future.

The researchers should avoid using the LID item simultaneously. Alternatively, a thorough description that can help the panellists to distinguish the two dependent attributes should be provided, or at least a statement should be offered to remind the panellists try to judge the attribute without connecting the other in mind.

Chapter 7 Summary and further discussion

7.1 Summary of the research

The main aim of this research was to explore the applications of Rasch analysis in food-related consumer research for new food product development. Two consumer survey case studies have been delivered using existing instrument Health and Taste Attitudes Scales (Roininen *et al.*, 1999; Roininen *et al.*, 2000) and three new instruments associated with ready meal consumption developed by the researchers. Two sensory liking tests have been conducted on a broad range of food and beverage products using general sensory attributes shared by most of products, and specific products (*i.e.* beef lasagne ready meal) using product-related attributes elicited from consumer interviews.

Case study I compared the difference in using CTT approach and Rasch analysis for evaluating the underlying structure and other psychometric properties of an existing instrument Health and Taste Attitude Scales. The results shows the superiority of using Rasch analysis over CTT approach. For example, the major deficiencies of factor analysis had been outlined in section 3.6.1.1. In this study, different scale structures under taste-related scale were identified using exploratory factor analysis (EFA) and Rasch analysis individually, none of which were consistent with the original study (Roininen *et al.*, 1999). The results indicated that the factor analysis had reported excessive factors. For instance, subscale “Craving for sweet foods” was reported by EFA as two factors. However according to Rasch analysis, the items under these two factors should be interpreted as two levels on the same construct. In addition, the reliability of subscale “pleasure” reported by EFA was below the recommended threshold 0.7 (Nunnally and Bernstein, 1994), which might be a consequence brought by false factor analysis partition.

Case study II explored the application of the Many-Facet Rasch Rating scale (MFR-RS) model in sensory overall liking test. In this study, a composite measure of overall liking modelled using panellists’ ratings of eight sensory attributes (composite measurement) and a holistic measure modelled using a single overall acceptability item (individual measurement) were compared. The results showed that the product separation estimated from the composite measurement were more than twice the size of the individual measurement, implying greater degree of dispersion of the product can be identified with the composite measurement. Although the multiple comparisons differentiated the products into slightly more statistical groups based on the holistic measure than the composite measure, there were less overlapping in the groups in latter one, which also support the hypothesis that the composite measure modelled using Rasch analysis has better discriminating power to differentiate products. In addition, Rasch analysis obtains a set of procedure to guarantee the quality of the measurement.

Case study III explored the application of Rasch analysis in developing and validating consumer insight instrument. Three aspects associated with ready meal consumption were studied. The initial constructs were conceptualised using the Wright maps, on which the expected response patterns were defined. Based on the construct, three instruments were composed according to the information collected using focus group studies with consumers. Rasch analysis identified the optimised rating scale categories and the underlying structure of the instruments (*i.e.* dimensionality). The estimates of respondents on each subscale were modelled. Consumers were clustered into three segments based on their measures of four satisfaction attitudes related instruments, and three segments based on the estimates of product criteria related instruments[1]. A comparison between the consumption frequency of three types of meals by consumer segments, gender and age groups were conducted. The results verified the hypothesis that the consumers' consumption frequency of ready meals can be predicted by the segmentation based on satisfaction attitudes, which represented different satisfaction levels. The results also indicated that there was no difference between the high-consideration consumers and low-consideration consumers on the consumption frequency of ready meals. In addition, according to the expectation of Rasch analysis, the items can be interpreted as locating along the scale in a hierarchical manner. This hierarchical order obtained information such as the least satisfied properties of ready meals, relative importance of product criteria in relation to ready meal selection, and the most likely eating situation of ready meals. These information could be used for new ready meal development by industry.

Case study IV developed and validated a sensory instrument for benchmarking test of beef lasagne ready meals under the guidance of Rasch analysis. The initial attribute pool was developed according to literature and researchers' own consumption experience. It was refined via 45 one-to-one consumer interviews. 20 attributes from the refined attribute pool were selected for the benchmarking test. 6 distinct products were chosen as samples for instrument validation. A composite measure was modelled using Many-Facet Rasch Rating scale (MFR-RS) model on the ratings of the twenty attributes. The results of multiple

[1] From the respondents perspective, the product criteria instruments differentiated the respondents into high-consideration consumers and low-consideration consumers.

comparisons suggested that this composite measure can differentiate the products in expected rank order. Therefore, the instrument can be used for the benchmarking purpose. In addition, the product development opportunity can be identified according to the hierarchical order of the twenty attributes calibrated on the scale, among which the amount of visible vegetable chunks was the least liked attribute.

7.2 Contribution of knowledge

This research filled the gap of measurement theory in the food-related consumer insights and sensory study by introducing Rasch analysis, which is an advanced measurement paradigm for constructing invariant measures.

It demonstrated the advantages of Rasch analysis over CTT-based approaches through a series of comparisons on different aspects of measurement, from dimensionality test to assessing measurement stability.

It showed the ability of Rasch analysis on modelling a composite measure of sensory overall liking from multiple attributes, which exhibits a greater power of differentiating product than using the single overall acceptability item.

It not only obtained evidence of the benefits of using Rasch analysis in food-related consumer insights research and sensory study, but also provided examples on how to develop and validate an instrument from the beginning under the guidance of Rasch analysis.

A set of instruments that concerned with ready meal consumption were developed during the research, which can be adapted by food companies for developing new ready meal products.

7.3 Additional issues identified from the case studies

7.3.1 Optimal number of categories in rating scales

Problematic Rasch-Andrich thresholds were found in all four case studies, where the 7-, 9- and 11- point rating scales were collapsed to 4- or 5-point rating scales in order to improve the scale category effectiveness. The only exemption was observed in the instrument used for measuring consumer's decision patterns and the relative importance of product criteria in case study III, where a 5-point scale was used without the need for collapsing. This raised a question about how many categories should be used in the measurement.

According to these results, it seemed like 4- to 5- point rating scales may perform better than rating scales with 7-point or more categories. This empirical assertion is consistent with the tentative evidence found by Toland and Usher (2016), which suggest that 4-point scale is needed for a mathematics self-efficacy scale. However, none of the observation made from this research or by Toland and Usher (2016) can be supported by decisive evidence.

In fact, different answers can be found in the literature. For survey, the 7-point scale seems to be preferred by most of the researches who studied the impact of the number of categories on the reliability or validity of instruments, such as Cicchetti *et al.* (1985) and Oaster (1989). This was also concluded by Cox (1980) in his review paper. Some other researchers suggested that less categories should be used, such as 4-point (Chang, 1994), and 5-point (Jenkins and Taber, 1977; Lissitz and Green, 1975). By contrast, Coelho and Esteves (2007) found using a 10-point scale could obtain higher discriminant validity than using a 5-point scale in a consumer satisfaction study, while Cummins and Gullone (2000) observed higher scale sensitivity for a 10-point scale than a 5- or 7-point scale in the subjective quality of life questionnaire.

For sensory evaluation, the number of categories vary depending on the purpose of the measurement. Lawless *et al.* (2010) compared the standard 9-point hedonic scale with an 11-point hedonic scale, and the labelled magnitude scale, which indicated that the three scales have equal ability to differentiate the acceptability of products. No comparison had been done between the 9-point hedonic scale and a revised scale with less number of categories.

In short, there is no agreement about the optimal number of rating scale categories. Therefore, as suggested by Linacre (2002a), one should always examine the functioning of rating scale in data analysis, which can benefit from conducting Rasch analysis.

It is noticeable that, the research which attempted at directly comparing the use of rating scale with different number of categories were all conducted within CTT approach. Perhaps a future study in sensory test area can be done within the framework of Rasch analysis.

7.3.2 Issue associated with local item independence

Another common issue found in three of the four case studies was the violation of assumption of local item independence. Rasch analysis provides an effective frame that can help to identify and tackle this issue. The test of local item independence in Rasch analysis can help the researchers to identify the problem related to item design. For instance, in case study III, the item S8 (I'm curious about a new recipe, I found the ready meal version which can be tried before I cook it myself) and S12 (I spot a new dish (ready meal) that I haven't tried before) were combined to a super-item due to LID (residual correlation=0.43). According to the feedback from some respondents, these items were considered redundant because they were not designed effectively. In a future study, the two items should be rephrased to one item.

7.3.3 Sample size consideration

7.3.3.1 Sample size consideration for constructing stable measures for exploratory research

Generally speaking, larger sample size produces more stable results under the same research framework (Linacre, 1994b). The results of case study I indicated that Rasch analysis can obtain more stable person estimates than CTT at same sample size. This finding is consistent with previous research conducted by Magno (2009). Therefore, to obtain same level of estimation stability, Rasch analysis requires a smaller sample size than CTT. This is an advantage of applying Rasch analysis in the consumer research, especially for the exploratory research, because it is time and cost saving to collect data from a relatively small sample. The rule of thumb (Linacre, 1994b) for the minimal sample size for exploratory purpose is 30 with dichotomous Rasch model and 50 with polytomous Rasch rating scale model (Andrich, 1978a). It should be noted that, if the polytomous Rasch partial credit model (Masters, 1982) is used, a larger sample size is needed because this model specifies the scale structure individually for each item.

7.3.3.2 Sample size consideration for conducting definitive statistical tests

Although Rasch analysis requires smaller sample size than CTT for constructing stable measures, the requirement of sample size are same for conducting definitive statistical tests followed by Rasch analysis or CTT.

For the dimensionality test using factor analysis or PCA, a few suggestions can be found from the literature. Guilford (1954) suggested that a sample of 200 or more should be used for these tests; while Kline (1979) pointed that a sample of 100 should be sufficient to obtain reliable result. Some other scholars also suggested that the ratio of respondents to items were taken into consideration, where the recommended minimum ratio were 3 to 10 (Cattell, 1978; Everitt, 1975; Gorsuch, 1983; Nunnally and Bernstein, 1994). In addition, Arrindell and Van der Ende (1985) argued that neither the absolute sample size nor the ratio of person over item are relevant. Their research indicated that the sample size should be related to ratio of respondents to the number of factors or components extracted from the test with a minimum value of 20.

For comparing the difference between subgroups by t-test or ANOVA, a minimal sample size of 30 per group is recommended by VanVoorhis and Morgan (2007). They also suggested at least 50 participants are needed for estimating the Pearson and Spearman’s correlation coefficients.

Nevertheless, in this research, all minimum sample size requirements were fulfilled in the four case studies.

7.3.4 Missing data

7.3.4.1 The mechanisms of missing data

The mechanisms of missing data can be classified into three categories (Little and Rubin, 1987; Rubin, 1976). For any data with missing responses:

(1) If the missingness is not related to any measurement variables, then the data are “Missing complete at random” (MCAR).

(2) If the probability of missing responses is dependent on the observed data, but not related to missing responses, then the data are “Missing at random” (MAR).

(3) If there is a systematic relationship between the probability of missing responses and their own values, then the data are “Missing not at random” (MNAR).

In this research, the incomplete data collected based on planned missing data design in the two sensory studies (case study II and IV) belong to MCAR. In additional, a few missing responses were observed in the survey data collected in case study III, which are also considered as MCAR because no special pattern of missingness was found in the data.

7.3.4.2 Impact of missing data on Rasch analysis estimation

Rasch analysis is robust to missing data (Scott, 2001; Wright, 1992). This is because the most commonly used estimation methods in Rasch analysis such as JMLE, PAIR, CMLE and MMLE (see Appendix B) are all based on the maximum likelihood method, which does not require complete data. The estimation of model parameters can be done by using all available information based on the likelihood. The marginal raw scores and counts of observed data are sufficient statistics for each parameter (Linacre, 2014c).

7.3.4.3 The technical considerations when applying Rasch analysis with incomplete data

Although Rasch analysis is robust against missing data, the degree of robustness may be affected by the mechanisms of missing data and the estimation methods.

(1) The mechanisms of missing data

The impacts of the three mechanisms of missing data on the quality of measurement are different (Kang, 2013; Mack *et al.*, 2018). Firstly, the MCAR type of missing data would reduce the statistical power due to the reduction of sample size. But, they do not introduce bias to the measurement. Secondly, the MAR type of missing data may or may not produce a biased result. Thirdly, if the missingness of data is MNAR, the results are likely to be biased. A recent simulation study on the impact of the mechanisms of missing data on the estimation of Rasch model parameters (Waterbury, 2019) found that the item estimates were not biased when the missing data were MCAR or MAR, while the negative bias was identified when the missing data are MNAR. Currently there is no formal procedure for minimising the impact of MNAR type of missing data on Rasch parameter estimation. Therefore, one should be more cautious when handling missing data that are considered as MNAR.

(2) The effects of missing data on parameter estimation vary between the estimation methods

A few research have compared the performance of different estimation methods in the presence of missing data within the framework of Rasch analysis. DeMars (2002; 2003) compared the impacts of missing data associated with MCAR and MAR on estimating Rasch model parameters by JMLE and MMLE, suggesting that MMLE is less robust against the missing data than JMLE. Heine and Tarnai (2015) evaluated the estimates and standard errors produced by CMLE, MMLE and PAIR with missing data in MCAR at different rates of missingness. They found that the three estimation methods are equally stable. SOYSAL *et al.* (2016) investigated the impact of MCAR and MAR types of missing data on the estimation of CMLE, JMLE and MMLE, concluding that the performance of JMLE is generally better than the others.

7.3.5 Inconsistency in the person-item response pattern

Rasch analysis requires a hierarchical probabilistic person-item response pattern (Engelhard, 2013), which can be described in two aspects:

(1) All persons are expected to have greater probability to endorse an item than other items which are more difficult to endorse.

(2) Any person who ranks higher on the scale should have more chance to endorse all items than the other persons located lower on the scale.

In the real world, however, certain degree of inconsistence to the expected pattern always exists in consumer research data. Therefore it is of importance to evaluate the degree of inconsistence associated with each measurement element to the required response. In Rasch analysis, this can be done by scanning the individual fit statistics (Wright, 1994a). If the fit statistics are located in the acceptable range (see table 2.6), then the data can be considered as appropriated for modelling using Rasch analysis.

In case study II of this research, the product marmite, which has been branded as a “love it or hate it” product (Reynolds, 2002) was evaluated together with the other products. If the marketing slogan of marmite is true, then it is expected that the panellists’ responses to marmite would exhibit a bimodal distribution. Surprisingly, the MNSQ statistics of Marmite in all models were satisfied[2], indicating that the panellists’ responses to Marmite were consistent with the model expectation.

In future sensory liking study, if a product is identified as misfitting and the source of misfit is linked to a bimodal response pattern, then a few approaches may be used:

(1) Removing the product exhibits bimodal response pattern from the data. The data of this product can be fitted to the model by its own for analysis.

(2) Conceptualising the product as two products for two groups of panellists

Similar to one of the methods used for resolving DIF (see section 2.3.4.3), the product may be conceptualised as two individual products, including one related to the panellists who like it very much and another one bound with the panellists who dislike it. The responses to the original product item are then split into two parts accordingly. Eventually two measures of same product will be estimated, corresponding to the location of the product on the scale rated by the panellists who “love” it or “hate” it, respectively.

7.4 Disadvantages of applying Rasch analysis in consumer research

Despite that Rasch analysis can break through a number of restrictions of CTT, it also has a few disadvantages.

(1) Rasch analysis requires data fit model, which cannot be met perfectly

Unlike CTT which makes weak assumptions (see section 2.1.1), Rasch analysis requires a few strong assumptions such as local independence, equal item

[2] Ranged from 1.13 to 1.28.

discrimination, no guessing. It requires data fit model. However, no data, especially which collected from the consumer research, can fit the model specification perfectly.

Therefore, when applying Rasch analysis, one should always diagnose to what extent the observed data cooperate with the model. A few approaches for improving the model fit exist, such as removing the extreme unexpected responses from the data or dropping the entire item.

(2) To produce the measures at interval level, Rasch analysis converts the raw scores into logits, which are not often used in the food-related consumer research area

Within the framework of Rasch analysis, the raw scores are converted into the interval additive logits, which can be defined as the natural log of an odds ratio (Cox, 1970). A positive difference of one logit on the measurement scale represents an increase of the odds of observing a model-specific event by 2.718[3] (Linacre and Wright, 1989a). However, this unit is not familiar to the consumer research practitioners.

In practice, to enhance the interpretability of Rasch analysis results, one can rescale the Rasch measures in logits back to the same range of the raw scores.

(3) The implementation of Rasch analysis requires additional training and specific software

Applying the Rasch analysis requires additional knowledge and understandings in statistics and the measurement theories. However, the practitioners are usually taught in CTT. They need to spend plenty of time on studying the relatively complicated Rasch analysis theory.

Moreover, the most commonly used statistical software do not fully support the application of Rasch analysis. The practitioners need to purchase additional software that are dedicated to Rasch analysis (see Appendix B), which will incur extra cost. They also have to spend time on learning how to use the Rasch software.

[3] Euler’s number *e*

7.5 Limitations of the research

The demographic information was not recorded in the two sensory studies (case study II and IV). Consequently, the effects of gender and age on the sensory liking could not be assessed.

For the two survey studies (case study I and III), although the gender and age were recorded, their distributions were unbalanced. There were much more females than males among the participants, while most of them were aged between 16-35. The samples may not be representative enough.

7.6 Conclusion

The benefits of using Rasch analysis in consumer research for new food product development related sensory practice had been demonstrated using the four case studies. Rasch analysis can overcome the limits of CTT, improving the quality of the measurement. It should be introduced to more consumer research related to new food product development.

Reference

Ahlgren, M., Gustafsson, I. B. and Hall, G. 2004. Attitudes and beliefs directed towards ready - meal consumption. *Food Service Technology,* **4**(4), pp.159-169.

Ahlgren, M. K., Gustafsson, I. B. and Hall, G. 2005. The impact of the meal situation on the consumption of ready meals. *International Journal of Consumer Studies,* **29**(6), pp.485-492.

Al - Obaidy, H., Khan, M. and Klein, B. 1984. Comparison Between Sensory Quality of Freshly Prepared Spaghetti with Meat Sauce Before and After Hot Hording on a Cafeteria Counter. *Journal of Food Science,* **49**(6), pp.1475-1477.

Alexandrowicz, R. W. 2012. GANZ RASCH. (Version 1.0). [Software].

Allen, I. E. and Seaman, C. A. 2007. Likert scales and data analyses. *Quality progress,* **40**(7), pp.64-65.

Álvarez, P. and Blanco, M. A. 2000. Reliability of the sensory analysis data of a panel of tasters. *Journal of the Science of Food and Agriculture,* **80**(3), pp.409-418.

Amerine, M., Roessler, E. and Filipello, F. 1959. Modern sensory methods of evaluating wine. *Hilgardia,* **28**(18), pp.477-567.

Andersen, E. B. 1973. Conditional inference for multiple - choice questionnaires. *British Journal of Mathematical and Statistical Psychology,* **26**(1), pp.31-44.

Andrich, D. 1978a. A rating formulation for ordered response categories. *Psychometrika,* **43**(4), pp.561-573.

Andrich, D. 1978b. Scaling attitude items constructed and scored in the Likert tradition. *Educational and Psychological Measurement,* **38**(3), pp.665-680.

Andrich, D. 1988. *Rasch models for measurement.* Sage.

Andrich, D. 2013. An expanded derivation of the threshold structure of the polytomous Rasch model that dispels any “threshold disorder controversy”. *Educational and Psychological Measurement,* **73**(1), pp.78-124.

Andrich, D. and Luo, G. 2003. Conditional pairwise estimation in the Rasch model for ordered response categories using principal components. *Journal of applied measurement,* **4**(3), pp.205-221.

Andrich, D., Sheridan, B. and Luo, G. 2010. RUMM2030: Rasch unidimensional models for measurement. (Version 5.1). [Software].

Arocas, A., Sanz, T., Salvador, A., Varela, P. and Fiszman, S. 2010. Sensory properties determined by starch type in white sauces: effects of freeze/thaw and hydrocolloid addition. *Journal of food science,* **75**(2), pp.S132-S140.

Arrindell, W. A. and Van Der Ende, J. 1985. An empirical test of the utility of the observations-to-variables ratio in factor and components analysis. *Applied Psychological Measurement,* **9**(2), pp.165-178.

Baranowski, T., Missaghian, M., Watson, K., Broadfoot, A., Cullen, K., Nicklas, T., Fisher, J. and O’donnell, S. 2008. Home fruit, juice, and vegetable pantry management and availability scales: A validation. *Appetite,* **50**(2-3), pp.266-277.

Bartlett, M. S. 1950. Tests of significance in factor analysis. *British Journal of statistical psychology,* **3**(2), pp.77-85.

Benjamini, Y. and Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pp.289-300.

Bergkvist, L. and Rossiter, J. R. 2007. The predictive validity of multiple-item versus single-item measures of the same constructs. *Journal of marketing research,* **44**(2), pp.175-184.

Bernstein, I. H. and Teng, G. 1989. Factoring items and factoring scales are different: Spurious evidence for multidimensionality due to item categorization. *Psychological Bulletin,* **105**(3), p467.

Bishop, P. A. and Herron, R. L. 2015. Use and misuse of the Likert item responses and other ordinal measures. *International journal of exercise science,* **8**(3), p297.

Bock, R. D. and Aitkin, M. 1981. Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika,* **46**(4), pp.443-459.

Boone, W. J. 2016. Rasch analysis for instrument development: why, when, and how? *CBE - Life Sciences Education,* **15**(4), pp.1-7.

Boone, W. J., Staver, J. R. and Yale, M. S. 2013. *Rasch Analysis in the Human Sciences.* Springer.

Brentani, E. and Golia, S. 2007. Unidimensionality in the Rasch model: how to detect and interpret. *Statistica,* **67**(3), pp.253-261.

Brown, M. B. and Forsythe, A. B. 1974. Robust tests for the equality of variances. *Journal of the American Statistical Association,* **69**(346), pp.364-367.

Brown, W. 1910. Some experimental results in the correlation of mental abilities 1. *British Journal of Psychology, 1904 ‐1920,* **3**(3), pp.296-322.

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.

Campbell, N. R. 1920. *Physics: The Elements.* Cambridge University Press.

Carifio, J. and Perla, R. 2008. Resolving the 50 ‐ year debate around using and misusing Likert scales. *Medical education,* **42**(12), pp.1150-1152.

Cattell, R. B. 1978. The scientific use of factor analysis. *New York*.

Chang, L. 1994. A psychometric evaluation of 4-point and 6-point Likert-type scales in relation to reliability and validity. *Applied psychological measurement,* **18**(3), pp.205-215.

Choppin, B. 1968. Item bank using sample-free calibration. *Nature,* **219**(5156), p870.

Chrisman, N. R. 1998. Rethinking levels of measurement for cartography. *Cartography and Geographic Information Systems,* **25**(4), pp.231-242.

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.

Churchill Jr, G. A. 1979. A paradigm for developing better measures of marketing constructs. *Journal of marketing research*, pp.64-73.

Cicchetti, D. V., Shoinralter, D. and Tyrer, P. J. 1985. The effect of number of rating scale categories on levels of interrater reliability: A Monte Carlo investigation. *Applied Psychological Measurement,* **9**(1), pp.31-36.

Clauser, B. and Linacre, J., M. 1999. Relating Cronbach and Rasch reliabilities. *Rasch Measurement Transactions,* **13**(2), p696.

Clauser, B. E. and Mazor, K. M. 1998. Using statistical procedures to identify differentially functioning test items. *Educational Measurement: issues and practice,* **17**(1), pp.31-44.

Coelho, P. S. and Esteves, S. P. 2007. The choice between a fivepoint and a ten-point scale in the framework of customer satisfaction measurement. *International Journal of Market Research,* **49**(3), pp.313-339.

Cohen, J. 1988. *Statistical power analysis for the behavioral sciences.* 2nd ed. L. Erlbaum Associates.

Cohen, L. 1979. Approximate expressions for parameter estimates in the Rasch model. *British Journal of Mathematical and Statistical Psychology,* **32**(1), pp.113-120.

Compusense Inc. 2013. Compusense 5. (Version 5.6). [Software].

Cooper, R. G. 1988. The new product process: a decision guide for management. *Journal of Marketing Management,* **3**(3), pp.238-255.

Cooper, R. G. and Kleinschmidt, E. J. 1986. An investigation into the new product process: steps, deficiencies, and impact. *Journal of product innovation management,* **3**(2), pp.71-85.

Cooper, R. G. and Sommer, A. F. 2016. Agile-Stage-Gate: New idea-to-launch method for manufactured new products is faster, more responsive. *Industrial Marketing Management,* **59**, pp.167-180.

Cortina, J. M. 1993. What is coefficient alpha? An examination of theory and applications. *Journal of applied psychology,* **78**(1), p98.

Costa, A. I. D. A., Dekker, M., Beumer, R. R., Rombouts, F. M. and Jongen, W. M. 2001. A consumer-oriented classification system for home meal replacements. *Food Quality and Preference,* **12**(4), pp.229-242.

Costello, A. B. and Osborne, J. W. 2005. Best Practices in Exploratory Factor Analysis: Four Recommendations for Getting the Most From Your Analysis. *Practical Assessment Research & Evaluation,* **10**(7).

Cox, D. R. 1970. *The analysis of binary data.* Methuen.

Cox, E. P. 1980. The optimal number of response alternatives for a scale: A review. *Journal of marketing research,* **17**(4), pp.407-422.

Cronbach, L. J. 1951. Coefficient alpha and the internal structure of tests. *psychometrika,* **16**(3), pp.297-334.

Cronbach, L. J., Gleser, G. C., Nanda, H. and Rajaratnam, N. 1972. *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles.* John Wiley & Sons.

Cronbach, L. J. and Meehl, P. E. 1955. Construct validity in psychological tests. *Psychological bulletin,* **52**(4), p281.

Cronbach, L. J., Rajaratnam, N. and Gleser, G. C. 1963. Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology,* **16**(2), pp.137-163.

Cummins, R. A. and Gullone, E. 2000. Why we should not use 5-point Likert scales: The case for subjective quality of life measurement. *In: Proceedings, second international conference on quality of life in cities*, p.93.

Daher, A. M., Ahmad, S. H., Than, W. and Selamat, M. I. 2015. Impact of rating scale categories on reliability and fit statistics of the Malay Spiritual Well-Being Scale using Rasch Analysis. *The Malaysian journal of medical sciences: MJMS,* **22**(3), p48.

De Battisti, F., Nicolini, G. and Salini, S. 2005. The Rasch model to measure the service quality. *The Journal of Services Marketing,* **3**(3), pp.58-80.

De Battisti, F., Nicolini, G. and Salini, S. 2010. The Rasch model in customer satisfaction survey data. *Quality Technology & Quantitative Management,* **7**(1), pp.15-34.

Demars, C. E. 2002. Incomplete data and item parameter estimates under JMLE and MML estimation. *Applied measurement in education*

15(1), pp.15-31.

Demars, C. E. 2003. Missing data and IRT item parameter estimation. *In: Annual meeting of the American Educational Research Association, Chicago*.

Devellis, R. F. 2006. Classical test theory. *Medical care*, pp.S50-S59.

Devellis, R. F. 2011. *Scale Development: Theory and Applications.* SAGE Publications.

Duncan, O. D. 1984. *Notes on social measurement: Historical and critical.* Russell Sage Foundation.

Dunn, O. J. 1961. Multiple comparisons among means. *Journal of the American statistical association,* **56**(293), pp.52-64.

Dunn, O. J. 1964. Multiple comparisons using rank sums. *Technometrics,* **6**(3), pp.241-252.

Eckes, T. 2011. *Introduction to Many-facet Rasch Measurement: Analyzing and Evaluating Rater-mediated Assessments.* Peter Lang.

Embretson, S. E. 1996. The new rules of measurement. *Psychological assessment,* **8**(4), p341.

Embretson, S. E. 1999. Issues in the measurement of cognitive abilities. *In:* S. E. EMBRETSON and S. L. HERSHBERGER, eds. *The new rules of measurement: What every psychologist and educator should know.* Psychology Press.

Embretson, S. E. and Hershberger, S. L. 1999. *The new rules of measurement: What every psychologist and educator should know.* Psychology Press.

Engelhard, G. 2013. *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences.* Routledge.

Everitt, B. 1975. Multivariate analysis: The need for data, and other problems. *The British Journal of Psychiatry,* **126**(3), pp.237-240.

Farley, H. A. and Reed, Z. 2005. An integrated sensory study of selected chilled lasagne ready meals. *Food Service Technology,* **5**(1), pp.35-45.

Feldt, L. S., Steffen, M. and Gupta, N. C. 1985. A comparison of five methods for estimating the standard error of measurement at specific score levels. *Applied psychological measurement*

9(4), pp.351-361.

Ferguson, G. A. 1941. The factorial interpretation of test difficulty. *Psychometrika,* **6**(5), pp.323-329.

Fisher, W. P. 1992. Reliability, Separation, Strata Statistics. *Rasch measurement transactions,* **6**(3), p238.

Fox, J., Weisberg, S., Adler, D., Bates, D., Baud-Bovy, G., Ellison, S., Firth, D., Friendly, M., Gorjanc, G. and Graves, S. 2018. R package 'car'. (Version 3.0-2). [Software].

Fuller, G. W. 2011. *New Food Product Development: From Concept to Marketplace, Third Edition.* Taylor & Francis.

Ganglmair, A. and Lawson, R. 2003. Advantages of Rasch modelling for the development of a scale to measure affective response to consumption. *In:*

D. TURLEY and S. BROWN, eds. *E - European Advances in Consumer Research.* Provo, UT: Association for Consumer Research, pp.162-167.

García, C., Ventanas, J., Antequera, T., Ruiz, J., Cava, R. and Alvarez, P. 1996. Measuring sensorial quality of Iberian ham by Rasch model. *Journal of food quality,* **19**(5), pp.397-412.

Garner, M. and Engelhard Jr, G. 2000. Rasch measurement theory, the method of paired comparisons, and graph theory. *Objective measurement: Theory into practice,* **5**, pp.259-286.

Geeroms, N., Verbeke, W. and Van Kenhove, P. 2008. Consumers' health-related motive orientations and ready meal consumption behaviour. *Appetite,* **51**(3), pp.704-712.

Gorsuch, R. L. 1983. *Factor Analysis.* L. Erlbaum Associates.

Green, B. G., Shaffer, G. S. and Gilmore, M. M. 1993. Derivation and evaluation of a semantic scale of oral sensation magnitude with apparent ratio properties. *Chemical senses,* **18**(6), pp.683-702.

Green, S. B. and Thompson, M. S. 2005. Structural equation modeling in clinical psychology research. *In:* R. M and I. S, eds. *Handbook of Research Methods in Clinical Psychology.* Oxford: Wiley-Blackwell, p.138.

Guilford, J. P. 1954. *Psychometric methods.* McGraw-Hill.

Haberman, S. J. 2007. The interaction model. *In:* M. VON DAVIER and C. H. CARSTENSEN, eds. *Multivariate and mixture distribution Rasch models.* Springer, pp.201-216.

Hagquist, C. and Andrich, D. 2017. Recent advances in analysis of differential item functioning in health research using the Rasch model. *Health quality of life outcomes,* **15**(1), pp.181-188.

Hair, J. F. 1995. *Multivariate Data Analysis: With Readings.* Prentice Hall.

Hambleton, R. K. 2006. Good practices for identifying differential item functioning. *Medical Care,* **44**(11), pp.S182-S188.

Hand, D. J. 1996. Statistics and the theory of measurement. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, pp.445-492.

Haraldsson, J. 2010. *Development of a Method for Measuring Pasta Quality*. MSc thesis, Linnaeus University.

Harvill, L. M. 1991. Standard error of measurement. *Educational Measurement: issues and practice*

10(2), pp.33-41.

Hasson, D. and Arnetz, B. B. 2005. Validation and findings comparing VAS vs. Likert scales for psychosocial measurements. *International Electronic Journal of Health Education,* **8**, pp.178-192.

Hays, W. L. 1963. *Statistics for Psychologists.* New York: Holt, Rinehart and Winston.

Heine, J.-H. 2017. R package 'pairwise'. (Version 0.4.3-2). [Software].

Heine, J.-H. and Tarnai, C. 2015. Pairwise Rasch model item parameter recovery under sparse data conditions. *Psychological Test Assessment Modeling,* **57**(1), pp.3-36.

Ho, P. 2019. A new approach to measuring Overall Liking with the Many-Facet Rasch Model. *Food Quality and Preference,* **74**, pp.100-111.

Hochberg, Y. 1988. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika,* **75**(4), pp.800-802.

Hoeppner, B. B., Kelly, J. F., Urbanoski, K. A. and Slaymaker, V. 2011. Comparative utility of a single-item versus multiple-item measure of self-efficacy in predicting relapse among young adults. *Journal of substance abuse treatment,* **41**(3), pp.305-312.

Holland, P. W. and Thayer, D. T. 1988. Differential item performance and the Mantel-Haenszel procedure. *In:* H. WAINER and H. I. BRAUN, eds. *Test validity.* Hillsdale, NJ: Lawrence Erlbaum Associates, pp.129-145.

Horn, J. L. 1965. A rationale and test for the number of factors in factor analysis. *Psychometrika,* **30**(2), pp.179-185.

Hsu, T.-C. and Feldt, L. S. 1969. The effect of limitations on the number of criterion score values on the significance level of the F-test. *American Educational Research Journal,* **6**(4), pp.515-527.

Jamieson, S. 2004. Likert scales: how to (ab) use them. *Medical education,* **38**(12), pp.1217-1218.

Jenkins, G. D. and Taber, T. D. 1977. A Monte Carlo study of factors affecting three indices of composite scale reliability. *Journal of Applied Psychology,* **62**(4), p392.

Jöreskog, K. G. 1973. A general method for estimating a linear structural equation system. *In:* A. S. GOLDBERGER and O. D. DUNCAN, eds. *Structural Equation Models in the Social Sciences.* New York: Academic Press, pp.83-112.

Jöreskog, K. G. and Van Thiilo, M. 1972. Lisrel A general computer program for estimating a linear structural equation system involving multiple indicators of unmeasured variables. *ETS Research Bulletin Series,* **1972**(2), pp.i-71.

Juster, F. T. 1966. Consumer buying intentions and purchase probability: An experiment in survey design. *Journal of the American Statistical Association,* **61**(315), pp.658-696.

Kaiser, H. F. 1958. The varimax criterion for analytic rotation in factor analysis. *Psychometrika,* **23**(3), pp.187-200.

Kaiser, H. F. 1970. A second generation little jiffy. *Psychometrika,* **35**(4), pp.401-415.

Kaiser, H. F. 1974. An index of factorial simplicity. *Psychometrika,* **39**(1), pp.31-36.

Kaiser, H. F. and Rice, J. 1974. Little jiffy, mark IV. *Educational and psychological measurement,* **34**(1), pp.111-117.

Kang, H. 2013. The prevention and handling of the missing data. *Korean journal of anesthesiology,* **64**(5), pp.402-406.

Kaufman, L. and Rousseeuw, P. J. 2005. *Finding Groups in Data: An Introduction to Cluster Analysis.* Wiley.

Kelderman, H. 1984. Loglinear Rasch model tests. *Psychometrika,* **49**(2), pp.223-245.

Kelderman, H. and Rijkes, C. P. 1994. Loglinear multidimensional IRT models for polytomously scored items. *Psychometrika,* **59**(2), pp.149-176.

Kim, J.-O. and Mueller, C. W. 1978. *Factor analysis: Statistical methods and practical issues.* Sage.

Kirk, R. E. 1996. Practical significance: A concept whose time has come. *Educational psychological measurement,* **56**(5), pp.746-759.

Kline, P. 1979. *Psychometrics and Psychology.* Academic Press.

Kline, P. 1994. *An Easy Guide to Factor Analysis.* Psychology Press.

Kolmogorov, A. 1933. Sulla determinazione empirica di una lgge di distribuzione. *Inst. Ital. Attuari, Giorn.,* **4**, pp.83-91.

Kowalkowska, J., Lonnie, M., Wadolowska, L., Czarnocinska, J., Jezewska-Zychowicz, M. and Babicz-Zielinska, E. 2018. Health-and taste-related attitudes associated with dietary patterns in a representative sample of Polish girls and young women: A cross-sectional study (GEBaHealth Project). *Nutrients,* **10**(2), p254.

Kruskal, W. H. and Wallis, W. A. 1952. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association,* **47**(260), pp.583-621.

Kuder, G. F. and Richardson, M. W. 1937. The theory of the estimation of test reliability. *Psychometrika,* **2**(3), pp.151-160.

Kuzon, W., Urbanchek, M. and Mccabe, S. 1996. The seven deadly sins of statistical analysis. *Annals of plastic surgery,* **37**, pp.265-272.

Landy, P., Boucon, C., Kooyman, G. M., Musters, P. A., Rosing, E. A., De Joode, T., Laan, J. and Haring, P. G. 2002. Sensory and chemical changes in tomato sauces during storage. *Journal of agricultural and food chemistry,* **50**(11), pp.3262-3271.

Larmond, E. and Voisey, P. W. 1973. Evaluation of Spaghetti Quality by a Laboratory Panela. *Canadian Institute of Food Science and Technology Journal,* **6**(4), pp.209-211.

Lawless, H. T. and Heymann, H. 2010. *Sensory Evaluation of Food: Principles and Practices.* Springer New York.

Lawless, H. T., Popper, R. and Kroll, B. J. 2010. A comparison of the labeled magnitude (LAM) scale, an 11-point category scale and the traditional 9-point hedonic scale. *Food Quality and Preference,* **21**(1), pp.4-12.

Lazarsfeld, P. F. 1959. Latent structure analysis. *In:* S. E. KOCH, ed. *Psychology: A study of a science.* New York: McGraw-Hill, pp.476-543.

Lee, Y.-W. 2004. Examining passage-related local item dependence (LID) and measurement construct using Q3 statistics in an EFL reading comprehension test. *Language testing,* **21**(1), pp.74-100.

Levene, H. 1960. Robust Tests for Equality of Variances. *In:* I. OLKIN, ed. *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling.* Stanford University Press, pp.278-292.

Likert, R. 1932. A technique for the measurement of attitudes. *Archives of psychology*.

Lilliefors, H. W. 1967. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American statistical Association,* **62**(318), pp.399-402.

Lim, J. and Fujimaru, T. 2010. Evaluation of the labeled hedonic scale under different experimental conditions. *Food quality and preference,* **21**(5), pp.521-530.

Lim, J., Wood, A. and Green, B. G. 2009. Derivation and evaluation of a labeled hedonic scale. *Chemical senses,* **34**(9), pp.739-751.

Linacre, J. M. 1989. *Many-facet Rasch Measurement.* MESA Press.

Linacre, J. M. 1994a. PROX with missing data, or known item or person measures. *Rasch Meas Trans,* **8**(3), p378.

Linacre, J. M. 1994b. Sample Size and Item Calibration Stability. *Rasch Measurement Transactions,* **7**(4), p328.

Linacre, J. M. 1995a. Prioritizing misfit indicators. *Rasch Measurement Transactions,* **9**(2), pp.422-423.

Linacre, J. M. 1995b. PROX for polytomous data. *Rasch Measurement Transactions,* **8**(4), p400.

Linacre, J. M. 1997. KR-20/Cronbach alpha or Rasch person reliability: which tells the “truth”? . *Rasch Measurement Transactions,* **11**(3), pp. 580-581.

Linacre, J. M. 1998. Detecting multidimensionality: which residual data-type works best? *Journal of outcome measurement,* **2**, pp.266-283.

Linacre, J. M. 1999. Understanding Rasch measurement: estimation methods for Rasch measures. *Journal of outcome measurement,* **3**, pp.382-405.

Linacre, J. M. 2001. Category, step and threshold: definitions & disordering. *Rasch measurement transactions,* **15**(1), p794.

Linacre, J. M. 2002a. Optimizing rating scale category effectiveness. *J Appl Meas,* **3**(1), pp.85-106.

Linacre, J. M. 2002b. What do infit and outfit, mean-square and standardized mean. *Rasch Measurement Transactions,* **16**(2), p878.

Linacre, J. M. 2006. Dichotomous Equivalents to Rating Scales. *Rasch Measurement Transactions,* **20**(1), p1052.

Linacre, J. M. 2014a. Facets Rasch measurement computer program. (Version 3.7.1). [Software].

Linacre, J. M. 2014b. *A user's guide to FACETS: Rasch-model computer program. Version 3.71.* Chicago: Winsteps.com.

Linacre, J. M. 2014c. *A user's guide to WINSTEPS MINISTEP Rasch-model computer programs. Version 3.81.* Chicago IL: Winsteps.com.

Linacre, J. M. 2014d. WINSTEPS Rasch measurement computer program. (Version 3.8.1). [Software].

Linacre, J. M. and Wright, B. D. 1989a. The "Length" of a Logit. *Rasch Measurement Transactions,* **3**(2), pp.54-55.

Linacre, J. M. and Wright, B. D. 1989b. Mantel-Haenszel DIF and PROX are Equivalent! *Rasch Measurement Transactions,* **3**(2), pp.52-53.

Lissitz, R. W. and Green, S. B. 1975. Effect of the number of scale points on reliability: A Monte Carlo approach. *Journal of Applied Psychology,* **60**(1), p10.

Little, R. J. A. and Rubin, D. B. 1987. *Statistical Analysis With Missing Data.* Wiley.

Loo, R. 2002. A caveat on using single-item versus multiple-item scales. *Journal of managerial psychology,* **17**(1), pp.68-75.

Lord, F. M. 1980. *Applications of Item Response Theory to Practical Testing Problems.* Routledge.

Lord, F. M. and Novick, M. R. 1968. *Statistical theories of mental test scores.* Addison-Wesley Pub. Co.

Lüdecke, D. 2019. R package 'sjstats'. (Version 0.17.4). [Software].

Mack, C., Su, Z. and Weistreich, D. 2018. *Managing Missing Data in Patient Registries: Addendum to Registries for Evaluating Patient Outcomes: a User's Guide.* Agency for Healthcare Research and Quality (US).

Macnaughton-Smith, P., Williams, W., Dale, M. and Mockett, L. 1964. Dissimilarity analysis: a new technique of hierarchical sub-division. *Nature,* **202**(4936), p1034.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K., Studer, M., Roudier, P., Gonzalez, J. and Kozlowski, K. 2018. R package 'cluster'. (Version 2.0.7-1). [Software].

Magno, C. 2009. Demonstrating the difference between classical test theory and item response theory using derived test data. *The International Journal of Educational and Psychological Assessment,* **1**(1), pp.1-11.

Mahon, D., Cowan, C. and Mccarthy, M. 2006. The role of attitudes, subjective norm, perceived control and habit in the consumption of ready meals and takeaways in Great Britain. *Food Quality and Preference,* **17**(6), pp.474-481.

Mair, P., Hatzinger, R., Maier, M. J. and Rusch, T. 2018. R package 'eRm'. (Version 0.16-2). [Software].

Mantel, N. 1963. Chi-square tests with one degree of freedom; extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association,* **58**(303), pp.690-700.

Mantel, N. and Haenszel, W. 1959. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the national cancer institute,* **22**(4), pp.719-748.

Marais, I. and Andrich, D. 2008. Formalizing dimension and response violations of local independence in the unidimensional Rasch model. *J Appl Meas,* **9**(3), pp.200-15.

Marcus-Roberts, H. M. and Roberts, F. S. 1987. Meaningless statistics. *Journal of Educational Statistics,* **12**(4), pp.383-394.

Martinez-Martin, P. 2010. Composite rating scales. *Journal of the Neurological Sciences,* **289**(1-2), pp.7-11.

Marton, F. 1981. Phenomenography—describing conceptions of the world around us. *Instructional science,* **10**(2), pp.177-200.

Masters, G. N. 1982. A Rasch model for partial credit scoring. *Psychometrika,* **47**(2), pp.149-174.

Mciver, J. and Carmines, E. G. 1981. *Unidimensional scaling.* Sage.

Meiser, T. 1996. Loglinear Rasch models for the analysis of stability and change. *Psychometrika,* **61**(4), pp.629-645.

Merton, R. K. and Kendall, P. L. 1946. The focused interview. *American journal of Sociology,* **51**(6), pp.541-557.

Michell, J. 2002. Stevens's theory of scales of measurement and its place in modern psychology. *Australian Journal of Psychology,* **54**(2), pp.99-104.

Morgan, D. L. 1996. Focus groups. *Annual review of sociology,* **22**(1), pp.129-152.

Mosteller, F. and Tukey, J. W. 1977. Data analysis and regression: a second course in statistics. *Addison-Wesley Series in Behavioral Science: Quantitative Methods*.

Myers, J. H. and Reynolds, W. H. 1967. *Consumer Behavior and Marketing Management.* Houghton Mifflin.

Nelder, J. 1990. The knowledge needed to computerise the analysis and interpretation of statistical information. *Expert Systems and Artificial Intelligence: the need for information about data*, pp.23-27.

Norman, G. 2010. Likert scales, levels of measurement and the “laws” of statistics. *Advances in health sciences education,* **15**(5), pp.625-632.

Novick, M. R. and Lewis, C. 1967. Coefficient alpha and the reliability of composite measurements. *Psychometrika,* **32**(1), pp.1-13.

Nunnally, J. C. and Bernstein, I. H. 1994. *Psychometric theory.* McGraw-Hill.

Oaster, T. 1989. Number of alternatives per choice point and stability of Likert-type scales. *Perceptual and Motor Skills,* **68**(2), pp.549-550.

Olivera, D. F. and Salvadori, V. O. 2006. Textural characterisation of lasagna made from organic whole wheat. *International journal of food science & technology,* **41**, pp.63-69.

Olsson, U. 1979. On the robustness of factor analysis against crude classification of the observations. *Multivariate Behavioral Research,* **14**(4), pp.485-500.

Ough, C. and Winton, W. 1976. An evaluation of the Davis wine-score card and individual expert panel members. *American Journal of Enology and Viticulture,* **27**(3), pp.136-144.

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.

Pantouvakis, A. and Renzi, M. F. 2016. Exploring different nationality perceptions of airport service quality. *Journal of Air Transport Management,* **52**, pp.90-98.

Patterson, H. 1951. Change-over trials. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp.256-271.

Perline, R., Wright, B. D. and Wainer, H. 1979. The Rasch model as additive conjoint measurement. *Applied Psychological Measurement,* **3**(2), pp.237-255.

Peryam, D. R. and Pilgrim, F. J. 1957. Hedonic scale method of measuring food preferences. *Food technology*.

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,* **18**(1), pp.25-34.

Pett, M. A., Lackey, N. R. and Sullivan, J. J. 2003. *Making sense of factor analysis: The use of factor analysis for instrument development in health care research.* Sage.

Pohlert, T. 2018. R package PMCMRplus. (Version 1.4.0). [Software].

Prim, M., Gustafsson, I. B. and Hall, G. 2007. The appropriateness of ready meals for dinner. *Journal of Foodservice,* **18**(6), pp.238-250.

R Core Team. 2018. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rasch, G. 1960/1980. *Probabilistic Models for Some Intelligence and Attainment Tests.* Danmarks Paedagogiske Institut.

Rasch, G. 1977. On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. *Danish Yearbook of Philosophy,* **14**, pp.58-94.

Raubenheimer, J. 2004. An item selection procedure to maximize scale reliability and validity. *SA Journal of Industrial Psychology,* **30**(4), pp.59-64.

Reed, Z., Mcilveen ‐ Farley, H. and Strugnell, C. 2003. Factors affecting consumer acceptance of chilled ready meals on the island of Ireland. *International Journal of Consumer Studies,* **27**(1), pp.2-10.

Reed, Z., Mcilveen, H. and Strugnell, C. 2000. The retailing environment in Ireland and its effect on the chilled ready meal market. *Journal of Consumer Studies & Home Economics,* **24**(4), pp.234-241.

Reed, Z., Mcilveen, H. and Strugnell, C. 2001. The chilled ready meal market in Northern Ireland. *Nutrition & Food Science,* **31**(2).

Revelle, W. 2015. R package 'psych'. (Version 1.5.8). [Software].

Reynolds, E. 2002. Marmite develops' love it or hate it'theme in new ads. *Marketing*, pp.24-24.

Robitzsch, A., Kiefer, T. and Wu, M. L. 2018. R package 'TAM'. (Version 3.0-21). [Software].

Roininen, K., Lähteenmäki, L. and Tuorila, H. 1999. Quantification of consumer attitudes to health and hedonic characteristics of foods. *Appetite,* **33**(1), pp.71-88.

Roininen, K., Lähteenmäki, L. and Tuorila, H. 2000. An application of means ‐ end chain approach to consumers’ orientation to health and hedonic characteristics of foods. *Ecology of Food and Nutrition,* **39**(1), pp.61-81.

Roininen, K. and Tuorila, H. 1999. Health and taste attitudes in the prediction of use frequency and choice between less healthy and more healthy snacks. *Food Quality and Preference,* **10**(4-5), pp.357-365.

Roininen, K., Tuorila, H., Zandstra, E., De Graaf, C., Vehkalahti, K., Stubenitsky, K. and Mela, D. J. 2001. Differences in health and taste attitudes and reported behaviour among Finnish, Dutch and British consumers: a cross-national validation of the Health and Taste Attitude Scales (HTAS). *Appetite,* **37**(1), pp.33-45.

Rossiter, J. R. 2002. The C-OAR-SE procedure for scale development in marketing. *International journal of research in marketing,* **19**(4), pp.305-335.

Rost, J. 1990. Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement,* **14**(3), pp.271-282.

Royal, K. D. 2016. The Impact of Item Sequence Order on Local Item Dependence: An Item Response Theory Perspective. *Survey Practice,* **9**(5), p2797.

Rubin, D. B. 1976. Inference and missing data. *Biometrika,* **63**(3), pp.581-592.

Sailer, M. O. 2013. R package 'crossdes'. (Version 1.1-1). [Software].

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,* **57**(3), p377.

Salzberger, T., Holzmüller, H. H. and Souchon, A. 2009. Advancing the understanding of construct validity and cross-national comparability: Illustrated by a five-country study of corporate export information usage. *In:* R. R. SINKOVICS and P. N. GHAURI, eds. *New Challenges to International Marketing.* Emerald Group Publishing Limited, pp.321-360.

Sarstedt, M., Diamantopoulos, A. and Salzberger, T. 2016. Should we use single items? Better not. *Journal of Business Research,* **69**(8), pp.3199-3203.

Scheuneman, J. 1979. A method of assessing bias in test items. *Journal of Educational Measurement,* **16**(3), pp.143-152.

Scheuneman, J. D. 1975. A new method of assessing bias in test items. *In: the annual meeting of the American Educational Research Association, Washington, DC*.

Schutz, H. G. and Cardello, A. V. 2001. A labeled affective magnitude (lam) scale for assessing food liking/disliking 1. *Journal of Sensory Studies,* **16**(2), pp.117-159.

Scott, B. 2001. Rasch vs. Tradition. *Rasch Measurement Transactions,* **15**(1), p809.

Shapiro, S. S. and Wilk, M. B. 1965. An analysis of variance test for normality (complete samples). *Biometrika,* **52**(3/4), pp.591-611.

Shealy, R. and Stout, W. 1993. A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika,* **58**(2), pp.159-194.

Shen, X. 2015. R package 'vGWAS'. (Version 2015.01.08). [Software].

Smirnov, N. 1948. Table for estimating the goodness of fit of empirical distributions. *The annals of mathematical statistics,* **19**(2), pp.279-281.

Smith, E. V. 2002. Understanding Rasch measurement: Detecting and evaluating the impact of multidimenstionality using item fit statistics and principal component analysis of residuals. *Journal of applied measurement*.

Smith, R. M. 1996. A comparison of methods for determining dimensionality in Rasch measurement. *Structural Equation Modeling: A Multidisciplinary Journal,* **3**(1), pp.25-40.

Smith, R. M. 2000. Fit analysis in latent trait measurement models. *Journal of Applied measurement*.

Soutar, G. N. and Cornish-Ward, S. P. 1997. Ownership patterns for durable goods and financial assets: a Rasch analysis. *Applied Economics,* **29**(7), pp.903-911.

Soysal, S., Arikan, Ç. A. and Inal, H. 2016. Impact Of Missing Data On Rasch Model Estimations. *TOJET: The Turkish Online Journal of Educational Technology,* (Special Issue for INTE 2016).

Spearman, C. 1904. " General Intelligence," objectively determined and measured. *The American Journal of Psychology,* **15**(2), pp.201-292.

Spearman, C. 1910. Correlation calculated from faulty data. *British Journal of Psychology, 1904 ‐ 1920,* **3**(3), pp.271-295.

Steptoe, A., Pollard, T. M. and Wardle, J. 1995. Development of a measure of the motives underlying the selection of food: the food choice questionnaire. *Appetite,* **25**(3), pp.267-284.

Streiner, D. L. 2003. Starting at the beginning: an introduction to coefficient alpha and internal consistency. *Journal of personality assessment,* **80**(1), pp.99-103.

Sullivan, G. M. and Artino Jr, A. R. 2013. Analyzing and interpreting data from Likert-type scales. *Journal of graduate medical education,* **5**(4), pp.541-542.

Swaminathan, H. and Rogers, H. J. 1990. Detecting differential item functioning using logistic regression procedures. *Journal of Educational measurement,* **27**(4), pp.361-370.

Tanner, C., Kaiser, F. G. and Wöfing Kast, S. 2004. Contextual conditions of ecological consumerism: A food-purchasing survey. *Environment and Behavior,* **36**(1), pp.94-111.

Tavakol, M. and Dennick, R. 2011. Making sense of Cronbach's alpha. *International journal of medical education,* **2**, p53.

Tennant, A. and Conaghan, 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? *Arthritis Care & Research,* **57**(8), pp.1358-1362.

Tennant, A., Penta, M., Tesio, L., Grimby, G., Thonnard, J.-L., Slade, A., Lawton, G., Simone, A., Carter, J. and Lundgren-Nilsson, Å. 2004. Assessing and adjusting for cross-cultural validity of impairment and activity limitation scales through differential item functioning within the framework of the Rasch model: the PRO-ESOR project. *Medical care*, pp.I37-I48.

Teresi, J. A., Ramirez, M., Lai, J.-S. and Silver, S. 2008. Occurrences and sources of Differential Item Functioning (DIF) in patient-reported outcome measures: Description of DIF methods, and review of measures of depression, quality of life and general health. *Psychology Science Quarterly,* **50**(4), p538.

Thissen, D. 1982. Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika,* **47**(2), pp.175-186.

Thompson, M. J. 2003. The Application of Rasch Scaling to Wine Judging. *International Education Journal,* **4**(3).

Thurstone, L. L. 1931a. The measurement of social attitudes. *The journal of abnormal and social psychology,* **26**(3), p249.

Thurstone, L. L. 1931b. Multiple factor analysis. *Psychological Review,* **38**(5), pp.406-427.

Thurstone, L. L. 1934. The vectors of mind. *Psychological review,* **41**(1), pp.1-32.

Thurstone, L. L. 1947. Multiple-factor analysis; a development and expansion of The Vectors of Mind.

Toland, M. D. and Usher, E. L. 2016. Assessing mathematics self-efficacy: How many categories do we really need? *The Journal of Early Adolescence,* **36**(7), pp.932-960.

Torchiano, M. 2018. R package 'effsize'. (Version 0.7.4). [Software].

Tukey, J. W. 1949. Comparing individual means in the analysis of variance. *Biometrics*, pp.99-114.

Van Den Wollenberg, A. L. 1982. Two new test statistics for the Rasch model. *Psychometrika,* **47**(2), pp.123-140.

Van Der Horst, K., Brunner, T. A. and Siegrist, M. 2011. Ready-meal consumption: associations with weight status and cooking skills. *Public health nutrition,* **14**(2), pp.239-245.

Vanvoorhis, C. R. W. and Morgan, B. L. 2007. Understanding power and rules of thumb for determining sample sizes. *Tutorials in quantitative methods for psychology,* **3**(2), pp.43-50.

Verlegh, P. W. and Candel, M. J. 1999. The consumption of convenience foods: reference groups and eating situations. *Food Quality and Preference,* **10**(6), pp.457-464.

Villanueva, N. D., Petenate, A. J. and Da Silva, M. A. 2000. Performance of three affective methods and diagnosis of the ANOVA model. *Food Quality and Preference,* **11**(5), pp.363-370.

Von Hippel, E. 1986. Lead users: a source of novel product concepts. *Management science,* **32**(7), pp.791-805.

Wakeling, I. N. and Macfie, H. J. 1995. Designing consumer trials balanced for first and higher orders of carry-over effect when only a subset of k samples from t may be tested. *Food Quality and Preference,* **6**(4), pp.299-308.

Wang, W.-C., Cheng, Y.-Y. and Wilson, M. 2005. Local item dependence for items across tests connected by common stimuli. *Educational and Psychological Measurement,* **65**(1), pp.5-27.

Waterbury, G. T. 2019. Missing Data and the Rasch Model: The Effects of Missing Data Mechanisms on Item Parameter Estimation. *Journal of applied measurement,* **20**(2), pp.154-166.

Welch, B. L. 1947. The generalization ofstudent's' problem when several different population variances are involved. *Biometrika,* **34**(1/2), pp.28-35.

Whitley, E. and Ball, J. 2002. Statistics review 6: Nonparametric methods. *Critical care,* **6**(6), pp.509-513.

Wilcoxon, F. 1945. Individual comparisons by ranking methods. *Biometrics bulletin,* **1**(6), pp.80-83.

Williams, E. 1949. Experimental designs balanced for the estimation of residual effects of treatments. *Australian Journal of Chemistry,* **2**(2), pp.149-168.

Willse, J. T. 2014. R package 'mixRasch'. (Version 1.1). [Software].

Wilson, E. B. and Hilferty, M. M. 1931. The distribution of chi-square. *Proceedings of the National Academy of Sciences,* **17**(12), pp.684-688.

Wilson, K. L., Lizzio, A. and Ramsden, P. 1997. The development, validation and application of the Course Experience Questionnaire. *Studies in higher education,* **22**(1), pp.33-53.

Wilson, M. 2004. *Constructing measures: An item response modeling approach.* Routledge.

Wright, B. D. 1991. Factor analysis versus Rasch analysis of items. *Rasch Measurement Transactions,* **5**(1), pp.134-135.

Wright, B. D. 1992. Raw scores are not linear measures: Rasch vs. classical test theory CTT comparison. *Rasch Measurement Transactions*

6(1), p208.

Wright, B. D. 1994a. Data analysis and fit. *Rasch Measurement Transactions,* **7**(4), p324.

Wright, B. D. 1994b. Unidimensionality coefficient. *Rasch Measurement Transactions,* **8**(3), p385.

Wright, B. D. 1996a. Comparing Rasch measurement and factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal,* **3**(1), pp.3-24.

Wright, B. D. 1996b. Reliability and separation. *Rasch Measurement Transactions,* **9**(4), p472.

Wright, B. D. 1999. Fundamental measurement for psychology. *In:* S. E. EMBRETSON and S. L. HERSHBERGER, eds. *The new rules of measurement: What every psychologist and educator should know.* Psychology Press, pp.65-104.

Wright, B. D. and Linacre, J. M. 1989. Observations are always ordinal; measurements, however, must be interval. *Archives of physical medicine and rehabilitation,* **70**(12), pp.857-860.

Wright, B. D. and Linacre, J. M. 1994. Reasonable mean-square fit values. *Rasch Measurement Transactions,* **8**(3), p370.

Wright, B. D. and Masters, G. N. 1982. *Rating scale analysis.* Mesa Press.

Wright, B. D. and Masters, G. N. 1990. Computation of OUTFIT and INFIT Statistics. *Rasch Measurement Transactions,* **3**(4), pp.84-85.

Wright, B. D. and Masters, G. N. 2002. Number of Person or Item Strata. *Rasch Measurement Transactions,* **16**(3), p888.

Wright, B. D. and Panchapakesan, N. 1969. A procedure for sample-free item analysis. *Educational and Psychological measurement,* **29**(1), pp.23-48.

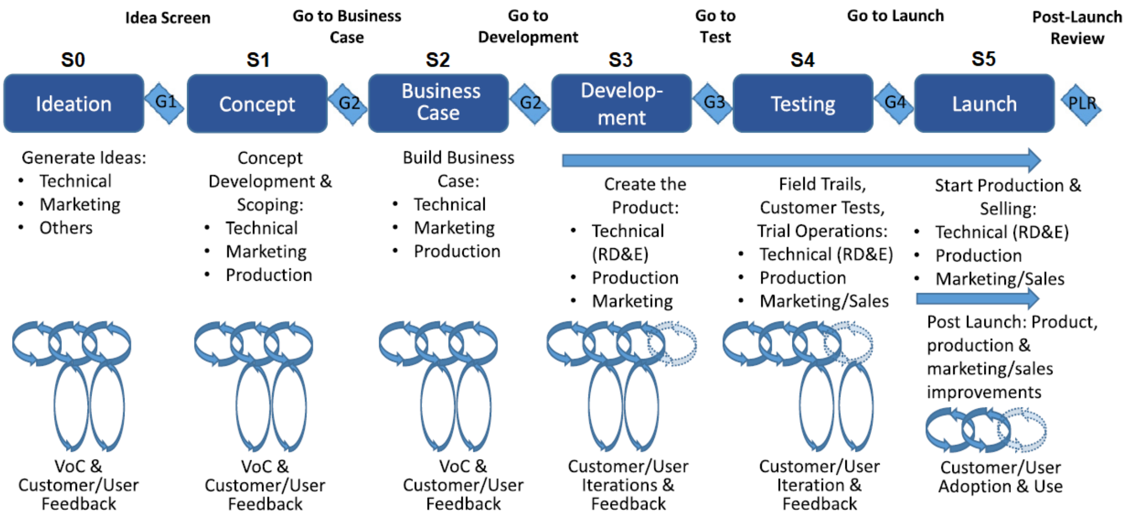
Wu, M. L., Adams, R. J. and Wilson, M. 2015. ACER ConQuest: Generalised item reponse modelling software. (Version 4). [Software].

Yen, W. M. 1984. Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement,* **8**(2), pp.125-145.

Yen, W. M. 1993. Scaling performance assessments: Strategies for managing local item dependence. *Journal of educational measurement,* **30**(3), pp.187-213.

Zandstra, E., De Graaf, C. and Van Staveren, W. 2001. Influence of health and taste attitudes on consumption of low-and high-fat foods. *Food Quality and Preference,* **12**(1), pp.75-82.

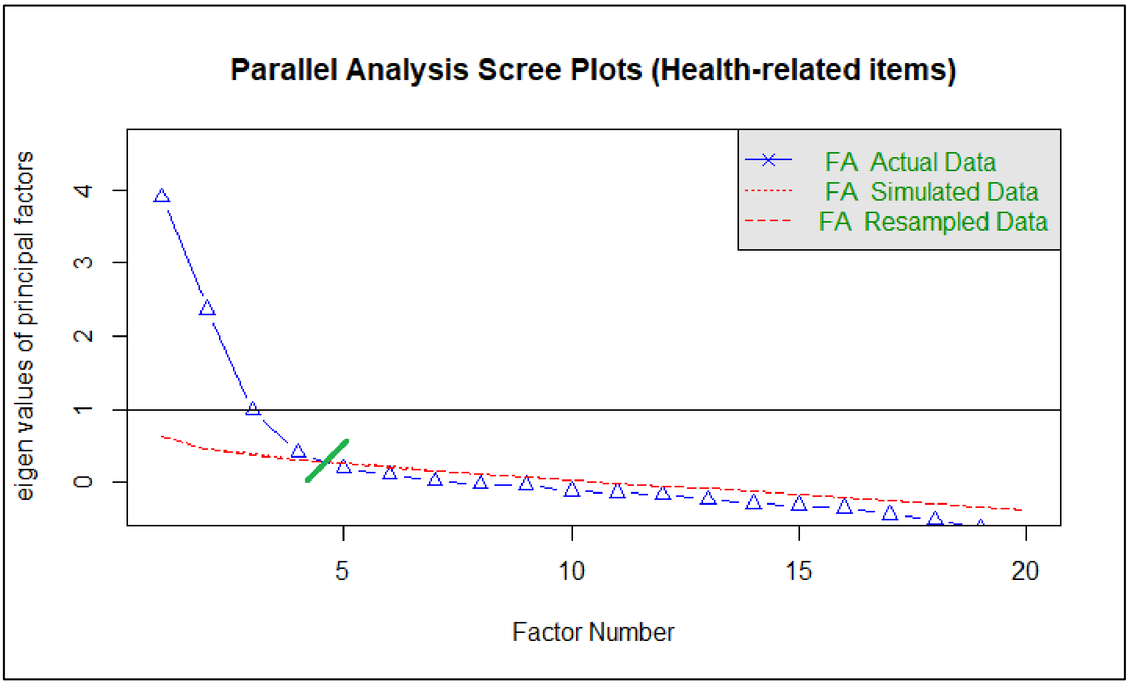
Appendix A: The latest version of Stage-Gate® model adapted from (Cooper and Sommer, 2016)



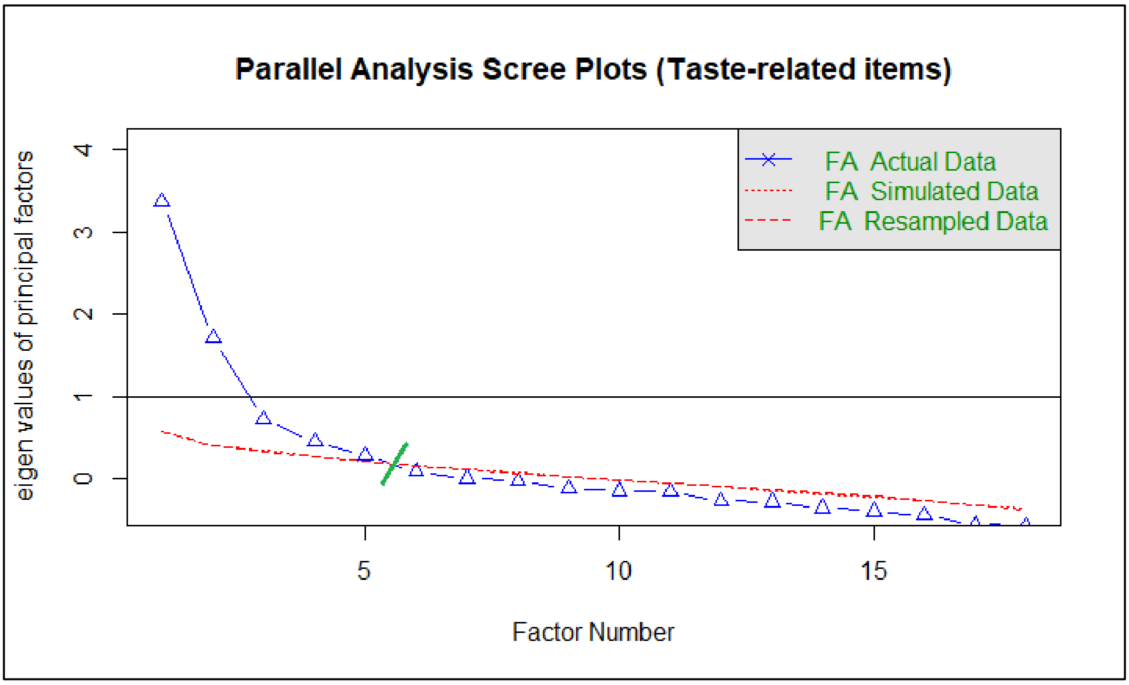
Appendix B: The main estimation methods of Rasch model parameters

Estimation Method	Iteration	Software
Pairwise Conditional likelihood Estimation (PAIR) (Andrich and Luo, 2003; Choppin, 1968; Garner and Engelhard Jr, 2000)	Non-iterative	RUMM2030 (Andrich *et al.*, 2010), R package ‘pairwise’ (Heine, 2017)
Normal Approximation Algorithm (PROX) (Cohen, 1979; Linacre, 1994a; Linacre, 1995b)	Non-iterative or Iterative	WINSTEPS (Linacre, 2014d), Facets (Linacre, 2014a), Ganz Rasch (Alexandrowicz, 2012)
Joint Maximum Likelihood Estimation (JMLE) (Wright and Panchapakesan, 1969)	Iterative	WINSTEPS (Linacre, 2014d), Facets (Linacre, 2014a), Ganz Rasch (Alexandrowicz, 2012), R package "TAM" (Robitzsch *et al.*, 2018), R package ‘mixRasch’ (Willse, 2014)
Marginal Maximum Likelihood Estimation (MMLE) (Bock and Aitkin, 1981; Thissen, 1982)	Iterative	ConQuest (Wu *et al.*, 2015), R package ‘TAM’ (Robitzsch *et al.*, 2018)
Conditional Maximum Likelihood Estimation (CMLE) (Andersen, 1973; Rasch, 1960/1980)	Iterative	Ganz Rasch (Alexandrowicz, 2012), R package ‘eRm’ (Mair *et al.*, 2018)

Appendix C: The scree plots obtained from parallel analysis (case study I)



Parallel analysis on health-related items



Parallel analysis on taste-related items

Appendix D: The initial attribute pool (case study IV)

Appearance related attributes

Attributes	Source
Distribution of browning	Farley and Reed, 2005
Meat sauce colour	AL-OBAIDY *et al.* 1984; Farley and Reed, 2005
Visibility of vegetable	Farley and Reed, 2005
Visibility of oil	Farley and Reed, 2005
Visibility of herb	Original*
Visibility of meat (amount)	Original*
Visibility of cheese	Original*
Colour of surface	Original*
Moistness	AL-OBAIDY *et al.* 1984
Fat separation	AL-OBAIDY *et al.* 1984
Overall consistency**	Farley and Reed, 2005
Meat particle size**	Farley and Reed, 2005
Firmness - Lasagne sheet**	Original*
Consistency of sauce**	Original*
Thickness (watery)**	Original*

*****According researchers' own consumption experience

**Visual perception about the texture

Aroma related attributes

Attributes	Source
Cheese aroma	Farley and Reed, 2005
Tomato aroma	Farley and Reed, 2005
Herb aroma	Farley and Reed, 2005
Meat aroma	Original
Onion aroma	Original

*****According researchers' own consumption experience

Taste-Flavour related attributes	
Attributes	**Source**
Meat flavour	Farley and Reed, 2005
Herb flavour	Farley and Reed, 2005
Tomato flavour	Farley and Reed, 2005; Landy *et al.* 2002
Vegetable flavour	Farley and Reed, 2005
Level of cheese flavour	Farley and Reed, 2005
Sweetness	Landy *et al.* 2002
Sourness	Landy *et al.* 2002
Saltiness	Landy *et al.* 2002
Bitterness	Landy *et al.* 2002
Onion flavour	Original*
Cream flavour	Original*
Intensity of spice flavour in the sauce	AL-OBAIDY *et al.* 1984
Blended flavour in the sauce	AL-OBAIDY *et al.* 1984
Intensity of spice flavour in the meat	AL-OBAIDY *et al.* 1984
Intensity of beefy flavour in the meat	AL-OBAIDY *et al.* 1984
Intensity of off-flavour	AL-OBAIDY *et al.* 1984
Malty	Landy *et al.* 2002
Old frying oil	Landy *et al.* 2002
Dry	Landy *et al.* 2002
Chemical	Landy *et al.* 2002
Smoky	Landy *et al.* 2002
Green	Landy *et al.* 2002
Old cloth	Landy *et al.* 2002
Cardboard	Landy *et al.* 2002
Earthy	Landy *et al.* 2002
Pasta	Landy *et al.* 2002
Applesauce	Landy *et al.* 2002
Rosebud	Landy *et al.* 2002
Maggi	Landy *et al.* 2002
Metal	Landy *et al.* 2002
Bitterness	Landy *et al.* 2002

*****According researchers' own consumption experience

Texture and mouthfeel related attributes	
Attributes	**Source**
Meat chewiness	AL-OBAIDY *et al.* 1984; Farley and Reed, 2005
Smoothness	Original*
Consistency - Béchamel sauce	Arocas *et al.* 2010
Resilience - Béchamel sauce	Arocas *et al.* 2010
Graininess	Arocas *et al.* 2010
Thickness	Arocas *et al.* 2010
Heterogeneity	Arocas *et al.* 2010
Creaminess	Arocas *et al.* 2010
Mouth coating	Arocas *et al.* 2010
Dryness	AL-OBAIDY *et al.* 1984
Greasiness	AL-OBAIDY *et al.* 1984
Particle Size	Original*
Body	Original*

Firmness - Lasagne sheet	Larmond and Voisey, 1973;
	Olivera and Salvadori, 2006;
	Haraldsson, 2010

Gumminess	Larmond and Voisey, 1973

Adhesiveness	Haraldsson, 2010;
	Larmond and Voisey, 1973;
	Olivera and Salvadori, 2006

Chewiness - Lasagne sheet	Larmond and Voisey, 1973
Starchiness	Larmond and Voisey, 1973
Cohesiveness	Olivera and Salvadori, 2006
Consistency - Lasagne sheet	Olivera and Salvadori, 2006
Springiness	Haraldsson, 2010 Olivera and Salvadori, 2006
Masticability	Olivera and Salvadori, 2006
Resilience - Lasagne sheet	Haraldsson, 2010
Meat chewiness	AL-OBAIDY *et al.* 1984; Farley and Reed, 2005
Consistency - Béchamel sauce	Arocas *et al.* 2010
Resilience - Béchamel sauce	Arocas *et al.* 2010

*****According researchers' own consumption experience

Appendix E: The refined attribute pool (case study IV)

Appearance (24 attributes)		
Attribute	**Count**	**Proportion[1]**
Visibility of meat (amount)	34	75.56%
Visibility of cheese	27	60.00%
Colour of surface	27	60.00%
Overall height	23	51.11%
Distribution of browning	22	48.89%
Overall firmness	22	48.89%
Visibility of vegetable chunks	19	42.22%
Meat sauce colour	14	31.11%
Amount of fillings	13	28.89%
Visibility of oil	10	22.22%
Number of layers	6	13.33%
Meat particle size	5	11.11%
Visibility of herb	4	8.89%
Thickness of lasagne sheet	4	8.89%
Consistency of sauce	3	6.67%
Juiciness	2	4.44%
Visibility of onion	2	4.44%
Overall consistency	2	4.44%
Colour of lasagne sheet	1	2.22%
Colour of white sauce	1	2.22%
Crispness of the cheese topping	1	2.22%
Moistness	1	2.22%
Type of cheese as topping	1	2.22%
Visibility of carrot	1	2.22%

[1] Total=45 participants

Aroma (18 attributes)		
Attribute	**Count**	**Proportion[1]**
Cheese aroma	43	95.56%
Tomato aroma	34	75.56%
Meat aroma	34	75.56%
Herb aroma	21	46.67%
Onion aroma	7	15.56%
Burned aroma	3	6.67%
Cream aroma	3	6.67%
Wine aroma	3	6.67%
Black pepper aroma	2	4.44%
Garlic aroma	2	4.44%
Butter aroma	1	2.22%
Cardboard aroma	1	2.22%
Carrot aroma	1	2.22%
Egg aroma	1	2.22%
Freshness	1	2.22%
Plastic aroma	1	2.22%
Wheat flour aroma	1	2.22%
White sauce aroma	1	2.22%

[1] Total=45 participants

Taste-flavour (29 attributes)		
Attribute	**Count**	**Proportion[1]**
Meat flavour	43	95.56%
Cheese flavour	42	93.33%
Tomato flavour	39	86.67%
Herb flavour	30	66.67%
Saltiness	24	53.33%
Cream/milk flavour	15	33.33%
Onion flavour	14	31.11%
Spice flavour	10	22.22%
Sweetness	8	17.78%
Sourness	6	13.33%
Garlic flavour	5	11.11%
Mushroom flavour	4	8.89%
White sauce flavour	4	8.89%
Blended flavour in the sauce	3	6.67%
Butter flavour	2	4.44%
Freshness	2	4.44%
Wheat flour flavour	2	4.44%
Burned flavour	1	2.22%
Egg flavour	1	2.22%
Other vegetable flavour	1	2.22%
Nutmeg flavour	1	2.22%
Oil flavour	1	2.22%
Pasta flavour	1	2.22%
Persistence of flavour	1	2.22%
Sour cream flavour	1	2.22%
Spiciness (aftertaste)	1	2.22%
Starch flavour	1	2.22%
Cardboard	1	2.22%
Wine flavour	1	2.22%

[1] Total=45 participants

Texture and mouthfeel (29 attributes)		
Attribute	**Count**	**Proportion**[1]
Firmness of lasagne sheet	39	86.67%
Chewiness of meat	38	84.44%
Thickness of sauce	28	62.22%
Creaminess of sauce	26	57.78%
Body	22	48.89%
Chewiness of lasagne sheet	14	31.11%
Heterogeneity	12	26.67%
Mouth coating	12	26.67%
Crispness of cheese topping	8	17.78%
Smoothness	8	17.78%
Perceived amount of particle in mouth	7	15.56%
Graininess	6	13.33%
Greasiness (mouthfeel)	5	11.11%
Extensibility of cheese	4	8.89%
Dryness of meat	3	6.67%
Gumminess of lasagne sheet	3	6.67%
Overall dryness (mouthfeel)	3	6.67%
Adhesiveness of lasagne sheet	2	4.44%
Consistency of lasagne sheet	2	4.44%
Perceived particle size in mouth	2	4.44%
Consistency of sauce	1	2.22%
Crunchiness of onion	1	2.22%
Dryness of lasagne sheet	1	2.22%
Firmness of the meat	1	2.22%
Fracturability of lasagne sheet	1	2.22%
Overall firmness	1	2.22%
Overall thickness	1	2.22%
Stickiness of white sauce	1	2.22%
Toughness of meat (aftertaste)	1	2.22%

[1] Total=45 participants