Calibrating Scales for Affective Responses to Physical Features of Products Using Rasch Measurement Theory

Fabio Ribeiro de Camargo

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

University of Leeds

School of Mechanical Engineering

June, 2013

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

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

Chapter 1 and Chapter 2

- Camargo, F.R. and Henson, B., 2012. The Rasch probabilistic model for measuring affective responses to product features. *Int. J. Human Factors and Ergonomics*, 1 (2), pp.204 – 219.

Chapter 3

- Camargo, F.R. and Henson, B., 2010. Measuring the specialness of confectionery: a Rasch model approach in affective engineering. *In*: *International conference on probabilistic models for measurement in education, psychology, social science and health*, 13 – 16 June, Copenhagen.
- Camargo, F.R. and Henson, B., 2011. Measuring affective responses for human-oriented product design using the Rasch model. *Journal of Design Research*, 9 (4), pp.360 – 375.
- Camargo, F.R. and Henson, B., 2012. Improving Kansei measurement using the Rasch model. *In*: *International conference on kansei engineering and emotion research*, 22 – 25 May, Penghu, Taiwan.

Chapter 4

- Camargo, F.R. and Henson, B., 2012. A rationale for comparing affective responses to stimulus objects using the faceted Rasch model. *In*: *International conference on probabilistic models for measurement in education, psychology, social science and health*, 23 –25 January, Perth, Australia.

Chapter 6

- Camargo, F.R. and Henson, B., 2012. Invariant comparisons in affective design. *In*: Y.G. Ji, ed. *Advances in affective and pleasurable design*, Boca Raton: CRC Press, pp.490 – 499.

Acknowledgements

This thesis would not happen to be possible without the encouragement and unconditional support of my family. I share the accomplishment of this work with my adorable wife Elizabeth who not only cheered me up during my difficulties and frustrations but also managed to keep me on the path. I wish to thank my two sons Juliano and Gustavo who have forgiven me for the frequent unavailability and my lovely daughter Julie who makes me smile all the time.

I would like to express my immense gratitude to Dr Brian Henson who fuelled my research with his ceaseless energy and motivation. I am very proud to have strengthened my academic ideals by his guidance. I shall also share the merits of a number of jointly-authored publications with Dr Henson, in which I have learned from his experience.

I gratefully acknowledge the support given by University of Leeds through a grant associated with the Research Mobility Programme of the Worldwide Universities Network that allowed a two-month research at the University of Western Australia. I would like to thank Dr Anoushka Kulikowski from the International Office of the University of Leeds for the support throughout the process.

I am indebt to Professor Alan Tennant from the Department of Rehabilitation Medicine of the University of Leeds for introducing me to Rasch measurement theory and to Professor David Andrich for making me very welcome in Australia and for providing a rich and fertile environment to study and explore new ideas. I would also like to thank Dr Ida Marais from the UWA, Dr Barry Sheridan from RUMM laboratory and Professor Tim Dunne from the University of Cape Town for the enjoyable time and stimulating conversations in Australia and South Africa.

I am very grateful to Toyota –Boshoku Corporation and University of Leeds for the generous opportunity of a three-month internship in Japan. The opportunity allowed an application of kansei methodology along with Rasch measurement theory in the company’s processes. I owe my gratitude to Mr Hiroshi Shibata, Mr Kenji Kawano, Mr Kasuki Hayashi and a special thanks to Mr Kazuyuki Motohata from the Evaluation and Engineering Division, Ms Yuka Osaki and Mr Shunsuke Tsuboi from the division of Global Human Resources Development. I am most grateful to Dr Jacqui Brown and Ms Aiko Mizumori from the University of Leeds for the support and confidence.

I am indebted to my many colleagues and all volunteers who contribute to the empirical studies of the research. My special thanks to Mr Alexis Lefreve for your genuine friendship.

I would like to thank my dear sister Cristine and my beloved mother for their extraordinary support and great faith in my perseverance. I will be grateful forever for your love.

I dedicate this work in memory of my father, a wise man.

Abstract

Affective engineering applies mathematical models to convert the information obtained from persons' feelings to product features into an ergonomic design. However, the methods commonly used to elicit persons' responses can present inaccuracies if measurement principles are violated. Consequently, empirical studies cannot be easily replicated and results cannot reliably be compared.

This research aimed to overcome the problem by establishing a novel approach in affective engineering using probabilistic models underpinned by Rasch measurement theory. The Rasch model verifies whether the observations meet the assumptions necessary for quantifying the numerical validity of the data employing the tools of standard statistics.

Initially, the research examined how well the data from affective responses would fit the expectations of the Rasch model to create a scale of specialness for four pieces of wrapped confectionery. Anomalies in the data were investigated to determine their potential impact on measure interpretation. A second empirical study investigated the stability of a measurement structure. Affective responses were compared with the physical properties related to compliance of a collection of product containers. A cross-validation strategy contrasted calibrations of the scale using different groups of respondents.

The results indicated that the differences between person locations on the measurement continuum from different calibrations were statistically non-significant. This provided evidence that the use of a Rasch-calibrated scale can systematically refine and generalise its frame of reference without loss of measurement properties.

The contribution of the research for the advancement of knowledge is established by transforming affective responses to physical elements into objective measures. A rationale was developed to achieve measurement properties in scales used in affective engineering, adapting the Rasch model's taxonomy used in other domains. Furthermore, the stability of a scale using different samples for calibration is demonstrated to be a property of Rasch-based scales. As a consequence of the stability, the association between affective responses and sensory information was realised and further variables were incorporated in the calibrated metric to refine the understanding of users' experience. Consequently, reliable results can be obtained from small samples, which will reduce time and costs of quantitative consumer research.

Table of contents

Chapter 3

Chapter 4

Chapter 5

Chapter 6

List of tables

List of figures

List of acronyms and symbols

AE	Affective engineering
ANOVA	Analysis of variance
CAD	Computer aided design
CAT	Computerized adaptive testing
CCC	Category characteristic curves
CTT	Classical test theory
DIF	Differential item functioning
DSF	Differential stimulus functioning
EPA	Evaluation, potency and activity dimensions
FA	Factor analysis
ICC	Item characteristic curve
INFIT	Inlier-sensitive fit
IRT	Item response theory
JMLE	Joint maximum likelihood estimation
KE	Kansei engineering
MANOVA	Mutivariate analysis of variance
MFRM	Many-facet Rasch model
MLE	Maximum likelihood estimation
OUTFIT	Outlier-sensitive fit
PCA	Principal component analysis
PML	Pairwise maximum likelihood algorithm
PROX	Normal approximation estimation algorithm
PSI	Person separation index
RM	Rasch model
RMP	Research Mobility Programme
RMT	Rasch measurement theory
SCC	Stimuli characteristic curve
SD	Semantic differential
UCON	Unconditional maximum likelihood procedure
UK MHRA	United Kingdom Medicines and Healthcare products Regulatory Agency
US FDA	United States Food and Drug Administration
WLE	Weighted likelihood estimation
WUN	Worldwide Universities Network
CI	Confidence interval
SE	Standard error
$RMSE$	Root mean square error
β	Person parameter
δ	Item parameter
σ	Standard deviation
ς	Stimulus parameter
τ	Threshold parameter
Pr	Theoretical probability
CVR	Covariance ratio

Immersion in water makes the straight seem bent; but reason, thus confused by false appearance, is beautifully restored by measuring, numbering and weighing; these drive vague notions of greater or less or more or heavier right out of the minds of the surveyor, the computer, and the clerk of the scales. Surely it is the better part of thought that relies on measurement and calculation.

Plato (The Republic)

CHAPTER 1

Introduction

In this chapter it is argued that measures from affective responses to physical elements produced in the domain of product design are absent of objectivity. To overcome the problem a novel approach is introduced in the domain based on a non-typical paradigm. The approach is underpinned by Rasch measurement theory, which allows independence between persons' estimates and the measurement instrument. Accordingly, a rationale is pursued throughout the research to transform observations into objective measures. Relevance to industry and academic research settings have been demonstrated by the outcomes of the research, strengthening this doctoral thesis as a contribution for the advancement of knowledge.

1.1 PERSONS' ATTITUDE TO PRODUCTS AND THE DOMAIN OF ENGINEERING

The value of consumers' attitude with regard to products and the transient nature of their desires have been acknowledged by scholars and explored by practitioners. Attitude is a hypothetical construct that has been defined as the persons' evaluation of particular objects and events that expresses some degree of positive or negative response with certain consistency (Fazio and Olson, 2003). The bond for positive responses and long-term attachment to a product is established by different dynamics. Design is one of the factors, playing a key role as a differential aspect of products' success or failure. If a product does not fulfil the consumers' expectations, disappointment and a negative impact on users' experience will certainly take place.

Eliciting consumer's attitude to a product is not straightforward as it is typically to measure the physical properties of the design elements that aggregate the product. Attitude to a product is frequently idiosyncratic, culturally located in the consumer's values and dependent on the influence of social groups. The role of an engineer or a designer is therefore expanded in a multidisciplinary field of knowledge that provides means to transform individuals' latent expression to a product into an improved design.

In the domain of engineering the behaviour of physical systems is commonly analysed through mathematical models. These models are simplified representations of a determined system embodying a set of assumptions and constraints on the values of variables. The magnitude of a variable is a measure of some quantifiable property within the modelled system. The quantifiable property is exemplified by differences in degree, such as the mass of a body, the displacement of a point on the body and the stress at a particular point in a deformed surface. Magnitude contrasts with qualitative properties that cannot be represented through different degrees, such as a brick, a proton and a sheet of steel (i.e., it is not possible to tell if a brick is more brick than another one). Therefore, if the physical system satisfies determined conditions, measurement can be performed. The theoretical foundations for those conditions were explicitly formulated by Helmholtz (1887) and by Campbell (1920).

However, quantifiable properties have been a bottleneck in the emerging area of engineering that seeks to integrate established mechanical engineering topics, such as design, with human factors (Schütte and Eklund, 2010; Laurans et al., 2009; Elokla and Hirai, 2012). In this domain physical elements are connected with one another and with them are associated dispositions of mind such as feelings and preferences. As such, observations are characteristically discrete, formalised when individuals respond to determined stimuli. In contrast to physical systems, discrete observations can just be measured indirectly if they are transformed by statistical models and if the observations meet measurement assumptions.

1.2 AFFECTIVE RESPONSES TO DESIGN ELEMENTS OR PRODUCT FEATURES

Affective engineering has been acknowledged as the domain of science that collects information about persons' feelings to products, identifies those aspects of the product to which people are responding and then uses the information to improve the design of the product (Barnes et al., 2004).

The topic has drawn attention of Japanese scholars and practitioners since the 1970s when Mitsuo Nagamachi was pointed by the Hiroshima University to the engineering management group with a briefing to develop an emotional ergonomics for product design. In 1974 Nagamachi published his paper titled "*A study of emotional technology*" (Nagamachi, 1974), emphasizing a consumer-oriented technology for new product development, called later *kansei engineering.* The Japanese kansei engineering (KE) has been used as a systematic approach that captures consumers' affective responses to products and employs mathematical models to convert the information into an ergonomic design (Nagamachi, 1989, 1995; Schütte and Eklund, 2001; Schütte, 2005).

Since Nagamachi's pioneering work, many other similar terms have been employed for identifying peoples' affective interaction with products or systems. Some examples are

Picard (1997) who has used the term *affective computing* and Karwowski (2005) who has employed the term *affective ergonomics.* Other terms with similar conceptual approaches have been introduced in the literature, such as emotional design, affective design, design for experience, pleasurable products and sensorial design (Schütte, 2005). The term *affective engineering* was applied for the first time during the affective human factors design conference in 2001 (Childs, 2010) in substitution to the word *emotional* and for representing the Western world view[1] of the Japanese *kansei.* The term *affective* is related to persons' feelings. Thus, the expression *affective responses* is, in this thesis, connected to persons' attitudes associated with the physical properties of products, following a definition adopted for practical purposes by Picard (1997) and by Schütte (2005).

1.3 THE PROBLEMATIC ISSUE OF MEASUREMENT OF LATENT VARIABLES IN AFFECTIVE ENGINEERING

The engagement of consumers in product development processes addresses characteristics beyond functionality and usability. However, there is a gap between knowing how well a certain product performs and understanding what a person actually feels when interacting with the product. Making inferences on phenomena that are not directly observable may be error-prone to some extent and may lead one to make a hasty generalisation. Thus, moving from functional design to human-centred design requires effective measurement methods that support the understanding of human-product interaction and that allow the sharing of outcomes with manufacturers, designers, researchers, and consumers.

A number of affective engineering (AE) studies have, however, presented results that are not stable in different samples and their procedures have not assured that the responses are assessed at least on an interval scale. As such, a number of applications in the domain have violated measurement principles and inaccuracies have been introduced in the process. For example, the inconsistent use of response options by respondents found in Chen et al. (2009) and statistical inferences using non-transformed data obtained from a continuous scale as in Laurans et al. (2009), to mention but a few. In some cases it is doubtful to assume that the mathematical operations needed to calculate means and

[1] Although it is possible to identify slight differences in the methodologies developed and used by kansei engineering and by affective engineering (Henson et al., 2006; Barnes et al., 2004), both approaches present similar concepts for human-centred design and therefore, they will not be distinguished throughout the thesis.

standard deviations support the assumption that the relevant latent attribute of a product is quantitative. Consequently, studies cannot be easily replicated and different results cannot reliably be compared. Thus, a theory-driven approach is required in the domain to overcome measurement problems in the design process.

1.3.1 A Solution Based on Fundamental Measurement

Fundamental measurement is a refined concept, but while discussing its most elementary understanding, it means merely to allow mathematical operations on measures (Andrich, 1988a). Thus, the notion of applicable mathematics can be extended to the general concept of meaningfulness (Finkelstein and Leaning, 1984) and to the concept of specific objectivity (Rasch, 1977), where comparisons between individuals ought to be generalised beyond the particular conditions under which they were observed.

Invariant comparison is an approach toward these concepts and it is a key consideration in fundamental measurement (Luce and Tukey 1964; Andrich, 1988a). As a result of the additive correspondence, comparisons can be made by the difference between the numbers associated with the persons' responses, where a particular difference has the same interpretation across a scale continuum.

The work in this thesis uses Rasch measurement theory for the development of scales for affective responses to design elements. Rasch measurement theory (RMT) fulfils the principles based on additivity, unit and invariant comparisons underpinned by the concept of fundamental measurement (Andrich 1988a; Embretson and Heise 2000). The Rasch model (RM) generically refers to a family of probabilistic models and provides procedures to assess measurement properties, increasing the information available about a scale. The measurement is obtained from a combination between persons' responses and independent variables, termed *items* in more recent psychometric approaches, and which will be used throughout this thesis. The RM's property of parameters' separability allows the design of a range of items on a scale and to distinguish individuals of a sample in different levels of attitude for each item. The model's procedures indicate a range of indirect tests for the hypothesis that the observations meet the necessary assumptions for quantifying the numerical validity of the data (Andrich, 1988a). Such procedures are denoted *calibration*, a term coined by Wright and Panchapakesan (1969), referring to measurement scales that are independent of the sample of persons used to estimate parameters of items and independent of the set of items used to obtain scale scores.

1.4 A PARADIGM SHIFT IN HUMAN-CENTRED DESIGN

Analysis of user-product interaction has typically used classical test theory (CTT). Although the classical approach is accepted as a well-established paradigm in the field of human-centred product design, some of their measurement hypotheses are rarely tested. In contrast, the RM is a probabilistic approach that examines the alignment of the measures against scientific measurement principles (Bond and Fox, 2007).

One of the measurement principles is concerned with the additivity of a rating scale. Usually, measurement of affective responses using classical approach obtains scores by counting the ordinal position of the response possibilities on a scale. However, a measurement process is achieved if the positions of the responses meet the assumption of interval properties (Wright and Stone, 1979). Therefore, the assumption ought always to be tested and not just assumed in every classical measurement process. On the other hand, the RM does not require presupposing such a property. The reason is that the RM provides theory and procedures to examine how well the data fit together and cooperate to define the attribute being measured.

Reliability is another key factor in a measurement process. In classical theory, reliability depends on the characteristics of a sample (Traub, 1994) and therefore, its stability is not expected across samples. Furthermore, classical reliability estimation fails when conveying information about different sources of errors, providing just one estimate of standard error for all respondents (Hays et al., 2000). Another consideration is that residuals (i.e., fitting errors) are not always independent, normally distributed or additive as required for some statistical methods. Frequently, residuals hold complex dependencies to preserve monotonicity with a scale. Differently, the RM distinguishes systematic measurement errors from random errors. This allows an analyst to characterise the sources of systematic error and calibrate the measurement structure subsequently.

Although persons' attitude originates from a multidimensional experience, such multidimensionality cannot be addressed as a whole in terms of measurement (Wright and Stone, 1979). The RM deals with this problem by decomposing the variable into a single dimension. The model's assumption of unidimensionality differs from AE applications that use methods for data reduction, such as factor analysis (FA) and principal component analysis (PCA). Those methods take into account that items hold equal difficulty of endorsement, clustering them solely on the basis of their correlations. However, the inherently multidimensional space originated from clustering variables can yield shortcomings because the properties of the affective attribute are usually complex and interconnected. Differently, the assessment of dimensionality within the context of the RM uses discrepancies between the observed responses and the expected responses by the model. Although a variety of psychological processes are involved when responding to a set of items, each item is affected by the same processes and in the same manner. That is, if data fit the model, items are said to be part of a unidimensional structure (Smith, 2002).

1.4.1 Rasch Measurement Theory and Item Response Theory

George Rasch, a Danish mathematician and statistician, developed his mathematical model emphasizing its measurement properties. The model was elaborated independently of item response models, comprising unique properties for successful measurement (Andrich, 1989), making clear distinction against the unidimensional item response theory (IRT) models. Nevertheless, the RM, called a one-parameter logistic model by some authors, has arguably been considered as the simplest model of the IRT approaches.

In the unidimensional IRT approaches, a model is fitted to a set of observed data. This allows the inclusion of additional parameters to represent the data. For example, the three-parameter model (Lord, 1980) and the two-parameter model (Lord, 1952; Birnbaum, 1968) add a parameter for discrimination of items (Embretson and Reise, 2000). However, the parameterisation complicates estimation and prevents a clear interpretation of items' hierarchy (Wright, 1995). The three-parameter model also includes a lower asymptote parameter for accommodating guessing responses. In this case, the model is fitted to the data in detriment of the specific objectivity property (Wright, 1999). Specific objectivity or measurement invariance is a unique property of the RM (Rasch, 1977; Bond and Fox, 2007; Fischer, 1995) (see Section 2.4.3). Furthermore, the addition of parameters in a measurement model will not result in linearity because they frequently compromise estimation of person and item parameters.

The approach for measurement of affective responses to product features using the RM is therefore to design appropriate items (i.e., adjectives or statements) that explain the attribute being measured and that produce a structure where data fit the model. In Rasch analysis, anomalies in the data to fit the model are identified and examined in order to validate the measurement structure. Consequently, the unique properties of the RM (e.g., linearity) can be obtained and the variance can mostly be explained (Liu, 2010).

1.4.2 Divergent Voices

Although the RM has supporters in different fields of knowledge, there are also discordant opinions regarding the model. Nunnally (1978), for example, stated there are few differences, if any, between scales designed by classical methods and those by the RM. Kline (1986, 1998) argued that a large sample has to be assessed for reliability without considering the characteristics of the distribution and that if there is enough calibration of items, nearly no items fit the model. However, classical regression models, for example, describe the data. This is not the purpose of the RM. The aim when using the RM is to establish quantifiable relations of a latent variable. These measures might not be necessarily a good description of the data (Perline et al., 1979). If there are just guesses, data will badly fit the model and most of the items will present misfit. This originates from the characteristic of lack of additivity in the data, rather than a problem associated with the model itself.

Goldstein (1979, 1980, 2012) has also been a vehement critic regarding the RM in the domain of education. One of Goldstein's claims is that the RM unrealistically supposes a unidimensional latent space. Linacre and Fisher's (2012) rebuttal to that claim emphasised that only when one dimension is isolated (even though there are no empirical data strictly unidimensional) it is possible to understand the meaning of the measure, and then study how that measure relates to measures on other dimensions. In a more general perspective, Sijtsma (2010) has argued that additive conjoint measurement and item response theory, including in this case the RM, do not succeed when addressing the nature of the measurement problems as a result of absence of well-established theories of psychological attributes. Nevertheless, Rasch theory is the sole approach that addresses and satisfies measurement principles in the domain of social sciences. In the traditional paradigm, the expertise of an analyst in statistics or data analysis will identify a model that accounts better for the given data although other problems can be found in those data. On the contrary, the challenge in RMT is that the expertise should come from those who are involved in the relevant field of application to understand the statistical misfit and where possible to generate new data that conform to the model, validating the latent variable (Andrich, 2004).

1.4.3 Previous Works

Despite the divergent voices, examples of RM applications regarding affective and sensory responses to products have mainly grown on consumption experience such as brand attractiveness, consumer behaviour and consumer satisfaction. Bechtel (1985) used the model for brand-attribute measurement related to preference, sweetness and fizziness of soft drinks. Garcia et al. (1996) measured sensorial quality of Iberian ham. Alvarez and Blanco (2000) applied in the model data from tasting panellists as part of a sensory evaluation of virgin olive oil. Alvarez and Galera (2001) established a hierarchy of attributes related to product and price evaluated by users of tractors after their acquisition. Ganglmair and Lawson (2003), using the model, developed a scale to measure affective response to consumption, conceptualising it as consumer satisfaction. Klöcker et al. (2012) have used the model to establish a scale of pleasantness when touching different materials. The RM has also been used to measure attributes based on user's experience related to services, such as anxiety with regard to train services (Cheng, 2010), analysis of users' satisfaction concerning hospital services (Lucadamo, 2010), airlines (Chang and Yang, 2008), and analysis of difficulties of elderly users of bus services (Chang and Wu, 2010). Soutar et al. (1990), Ewing et al. (2005) and Salzberger (2009) have presented measurement solutions using the RM in marketing, including cross-cultural issues.

1.5 A NOVEL APPROACH FOR MEASURING OBSERVATIONS IN PRODUCT DESIGN

In the field of human-centred product design, measurement approach using the RM is novel. Published applications of the model with implications in the domain, namely in AE, have practically been absent except those based on this doctoral thesis. The first application of the RM known in the domain was reported by Henson (2009) when using AE techniques at the UK arm of an international confectionery company in 2005. Henson's results suggested that further research would be necessary for an interpretation of the measures obtained when applying the RM in AE studies.

1.5.1 Anticipated Benefits

The novel approach developed throughout this thesis using RMT represents a potential advancement in measurement techniques in human-centred product design. Differently from other approaches in AE, a collection of relevant observations can be transformed into objective measures. This transformation will help analysts to exploit new design variables derived from compilations of consumers' experience when interacting with a product.

Objective measures can be achieved through of a rationale for applying the RM in the domain. The rationale will directly benefit product manufacturers because specific, calibrated scales could be developed for *off-the-shelf* administration. Furthermore, objective measurement in AE will allow the diagnosis of the best design for each particular individual affective response. This is possible because the rationale will provide a basis for the development of computer-aided assessments with reliable results from small samples which may reduce time and costs of quantitative consumer research.

1.6 RESEARCH AIM

The aim of the research is to provide a scientific rationale to establish valid and reliable metrics for quantifying differences between individuals associated with physical characteristics of design elements using probabilistic models underpinned by Rasch measurement theory in the domain of engineering, namely in product design. The purpose is to offer a contribution to overcome unresolved issues on generalisation of measurements in the domain.

1.6.1 Specific Objectives: Test of the Research Hypotheses

i. The primary research hypothesis formulates that the data obtained from persons' affective responses to design elements fit together and cooperate to define a quantitative structure established on the fundaments of RMT. To test this hypothesis an empirical study was designed for eliciting affective responses from a sample of persons to specialness of a set of wrapped confectionery.

ii. The test of the primary research hypothesis demonstrated that although part of the data fitted the RM, each stimulus ought to be considered in a different frame of reference for measurement. This finding raises the hypothesis that different continuums can be compiled in a sole scale if a derivation of the RM is used. The test of this hypothesis requires adapting a more complex model from the Rasch family of measurement models, namely the faceted Rasch model.

iii. Based on the findings from the tests mentioned in i and ii, three more hypotheses are formulated. One of the hypotheses is that the measurement of affective responses through calibrated structures using the RM does not vary within a same context even if different groups of persons are used.

iv. Another hypothesis is that if a calibrated set of items is kept as the mainstay of the measurement structure, further items could be calibrated and accommodated into the structure to convey further information about the attribute.

v. The final hypothesis formulates that if a stable measurement structure is obtained through calibration, then it is possible to model a design element of a product for particular affective responses. The rationale to test those three last hypotheses is to design three associated empirical studies.

1.6.2 Contributions to Knowledge

i. *Transformation of affective responses into objective measures*

In this thesis, affective responses to physical elements of products are demonstrated to have a quantitative structure when measurement properties are achieved. The approach uses RMT to transform raw scores obtained from self-report questionnaires to determined physical stimuli into objective measures. This transformation overcomes many of the problems of the current approaches in the domain, allowing the comparison between results from different studies and the generalisation of research findings.

ii. *Adaptation of the model's taxonomy for applications in the domain*

The taxonomy of the multi-facet Rasch model has been adapted to take into account the requirements of affective engineering process. The RM has mainly been applied in education, health and social sciences that use particular definitions and notations. One of the contributions of this thesis to AE is, therefore, to adjust the concepts, terms, definitions and equations used to elicit affective responses in the domain with those used in other fields of knowledge.

iii. *Identification of issues associated with local dependence in the faceted structure*

Local dependence in Rasch modelling prevents a data set to hold a quantitative structure and therefore, meaningful algebraic operations cannot be realised. An alternative technique to identify anomalies in a data set that originate local dependence is developed in this thesis. The solution is a novel contribution, applied when data from affective responses are analysed through the faceted Rasch model.

iv. *Demonstration of stability of a measurement structure across different samples*

The stability of a measurement structure is demonstrated with the application of the rationale developed throughout the research. In this thesis, the stability of a scale is shown to be a property of Rasch-based measurement structures. Therefore, no further demonstration of stability is considered necessary in other studies if Rasch-calibrated scales are established.

v. *Demonstration of the association between affective responses and sensory information*

As a consequence of the RM's property of independent parameterisation of estimations of persons, items and stimuli, the association between affective responses and sensory information can objectively be established through a shared physical element. The association is demonstrated to be invariant on the scale continuum. As a consequence, the research contributes to a more detailed understanding of connection between sensory and affective information for each respondent individually.

vi. *Establishment of the basis for the development of item banks*

Another contribution to human-centred product design is the demonstration of how an analyst can incorporate additional variables in a Rasch-calibrated scale to fit different persons' inclinations of endorsement. This is the basis for the development of an item bank, which can minimise costs of consumer research.

1.6.3 Impact of Research

i. *Knowledge creation*

In this thesis a novel approach is proposed to bridge the gap between objective measurement of latent variables and the current methods in the domain of AE. The approach contributes to strengthen the quality of studies in the domain through a rationale that supports the validity and stability in measurement of affective responses associated with physical elements, as demonstrated throughout the thesis. This creates a new perspective other than a description of the available observations, but one of generalisation of an AE outcome based on a different paradigm.

ii. *Collaboration with groups of academic research*

Measurement has been a critical topic of research in different fields of academic knowledge. Although RMT is a novel approach in product design, it has been vastly discussed and developed in education, health and social science. This has provided fruitful opportunities for critical and reflexive thinking as a key part to play in the product design applications. Those opportunities have turned into a reciprocal contribution, where other domains have also obtained benefits from a different perspective on non-typical applications of the RM, such as those reported in this thesis. This has, for example, established a mutual collaboration between the School of Mechanical Engineering of the University of Leeds and the Graduate School of Education of the University of Western Australia. Further collaboration has been established with the UK Rasch users group and the Rasch working group, which congregates experts from more than 15 world class universities.

iii. *Information to the professional practice*

The gap in current knowledge associated with the lack of objective measurement of latent variables in consumers quantitative research and the difficulty of conveying consumer information to support product design have frequently been a request from the professional practice. For example, Procter & Gamble Innovation Centre has provided written expression of interest with regard to the findings of this research for potential applications in its processes. Furthermore, Toyota-Boshoku Corporation provided a three-month internship in Japan to apply the approach along with its team of the Evaluation and Engineering Division. Those opportunities have allowed closing the loop between the contribution to academic knowledge and practical outcomes.

iv. *Knowledge dissemination and spillover effects*

The different aspects of the research produced six jointly academic publications and two submissions[2]. The focus has been on international journals in the domain of product design, engineering design and ergonomics (two publications and one submission). Peer-reviewed conferences papers have targeted audiences from the domain of human factors, engineering and measurement science (four publications and one submission).

The research has capitalized on interdisciplinary knowledge and therefore, it has also produced spillover effects. For example, a prospective application of the rationale proposed in this thesis was examined by scholars of computing linguistics of the University of Leeds to improve information in AE systems combined with data mining methods.

1.7 THESIS FRAMEWORK

The thesis is divided into nine chapters. These chapters follow the reasoning when adapting the RM in the domain of AE. Thus, they interplay between theoretical approaches and empirical studies (Figure 1.1).

In Chapter 2 a review of the literature provides an overview of AE processes focused on data analysis. The RM is introduced as a solution to overcome measurement limitations in the domain. In the chapter, an exposition on the underpinnings of the RM principles and the model’s properties are delineated. Furthermore, the procedures for calibrating scales are discussed.

Chapter 3 is concerned with an empirical approach to determine whether data from affective response to a set of wrapped confectionery would cooperate to establish a quantitative structure using the RM. The results indicate that part of the data fit the model although each stimulus presents a different frame of reference.

[2] A list of publications associated with this doctoral thesis can be seen on page ii.

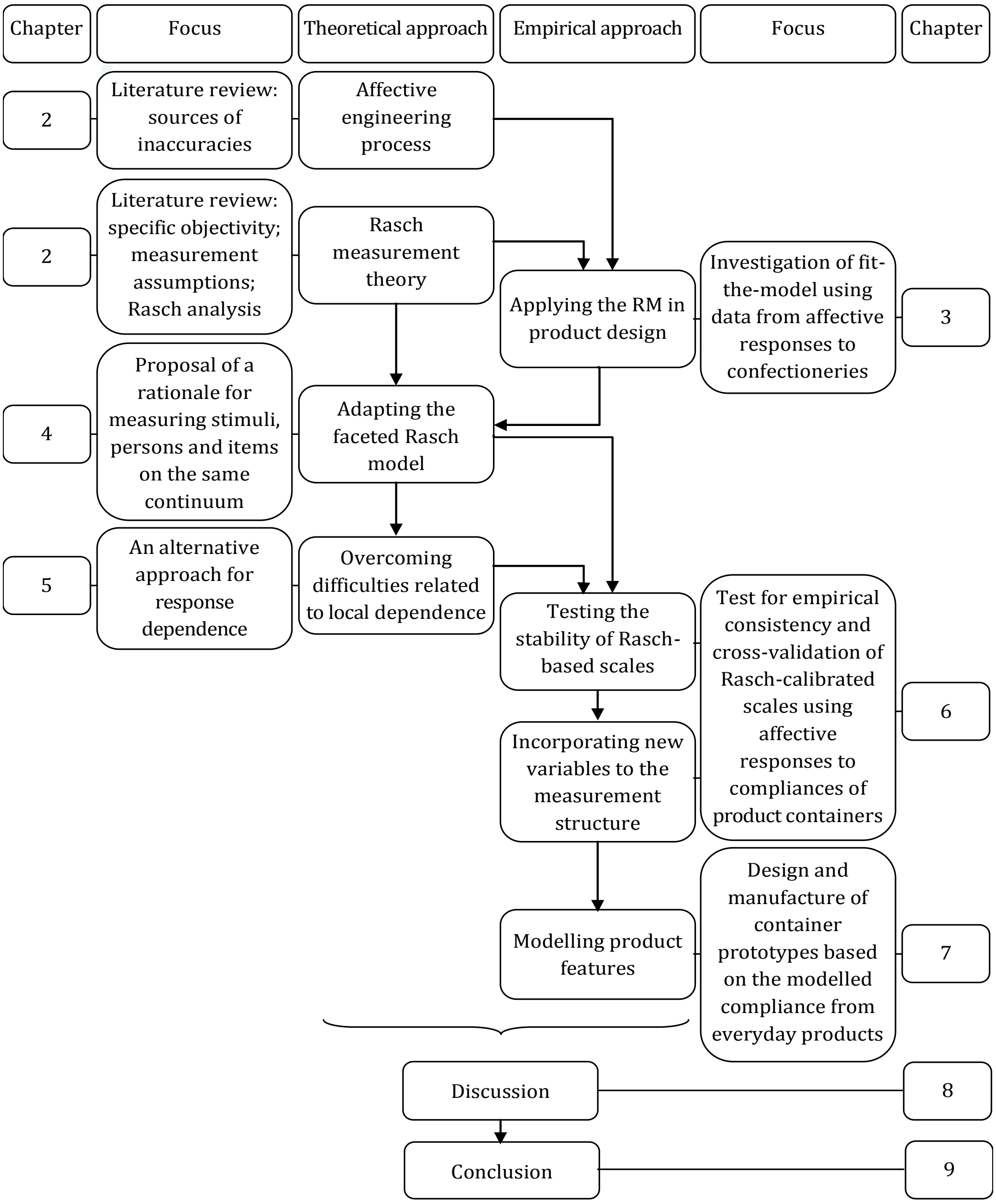

Chapter
Focus
Theoretical approach
Empirical approach
Focus
Chapter
2
Literature review: sources of inaccuracies
Affective engineering process
2
Literature review: specific objectivity; measurement assumptions; Rasch analysis
Rasch measurement theory
Applying the RM in product design
Investigation of fit-the-model using data from affective responses to confectioneries
3
4
Proposal of a rationale for measuring stimuli, persons and items on the same continuum
Adapting the faceted Rasch model
5
An alternative approach for response dependence
Overcoming difficulties related to local dependence
Testing the stability of Rasch-based scales
Test for empirical consistency and cross-validation of Rasch-calibrated scales using affective responses to compliances of product containers
6
Incorporating new variables to the measurement structure
Modelling product features
Design and manufacture of container prototypes based on the modelled compliance from everyday products
7
Discussion
8
Conclusion
9

Figure 1.1 - Graphic representation of the thesis framework**.**

In Chapter 4, on the basis of the results from the previous chapter, a derivation of the RM, namely the multi-faceted Rasch model is explored. The adoption of the model in the domain requires adaptation of its taxonomy. Furthermore, a novel test to determine whether the stimuli in a study present statistically distinguishable characteristics is proposed.

Chapter 5 is concerned with analysis of the data from the confectionery study using the multi-facet Rasch model, which indicated some peculiarities with regard to local independence. An alternative technique to test for local dependence is proposed without covering signals of anomalies if they do exist in the data.

Chapter 6 is related to an empirical study that examines the measurement stability when applying the faceted approach using data from affective responses to a collection of product containers. Stability is assessed by the replication of calibrated items across two different samples. Additionally, a cross-validation strategy compares calibrations using different groups of respondents. It is also shown that it is possible to incorporate items into the calibrated scale without loss of comparability.

In Chapter 7 through an empirical approach is demonstrated that the correspondence between sensory information and affective responses can be modelled, connecting them with a shared physical element. In the study, the modelled compliance of product packaging specified a set of container prototypes to stimulate particular responses.

Chapter 8 refers to the discussion of the outcomes from the research. The benefits of constructing item banks and computer-aided assessment are presented and further studies are also suggested. Finally, in Chapter 9 the conclusions are drawn from the outcomes of the research.

CHAPTER 2

Rasch measurement theory for eliciting affective responses in product design

Affective engineering uses mathematical models to convert the information obtained from persons' feelings to product features into an ergonomic design. However, the methods commonly used to elicit persons' responses can present shortcomings if measurement principles are violated. In this chapter a review of the main sources of inaccuracies in the process associated with measurement is presented. Furthermore, it is shown that although the measurement of attitude has been investigated in many domains of knowledge, the methods in affective engineering have evolved on a parallel path. One potential approach to overcome the problem is the Rasch model. The main assumptions of Rasch measurement theory that underlie a family of probabilistic models are reviewed. Based on the literature, support is provided for enabling an analyst to develop a frame of reference for affective responses with measurement properties[3].

2.1 AFFECTIVE ENGINEERING PROCESS

2.1.1 Affective Engineering and the Assessment of Attitude

Different approaches have been suggested in the literature for eliciting affective responses to design elements or product features (Table 2.1). The most commonly used method is to identify adjectives that people use to describe the product and embody them into a self-report, semantic differential (SD) questionnaire (Osgood et al., 1957; Schütte, 2005). A number of consumers are asked to rate the degree to which each word describes a range of product stimuli or a sample of potential components of a product usually through scales containing five, seven, nine or eleven categories.

The responses to the questionnaires are turned into a measure of affective response using statistical techniques such as PCA, clustering the responses against the words to a small number of constructs. This process creates qualitative semantic spaces against which to correlate measures of the physical properties or features of the products (Henson et al., 2006).

[3] Publication based on this chapter is found in Camargo and Henson (2012a).

Although adjectives and scales embedded in self-report questionnaires are common practices to capture people's feelings when interacting with products or design components, the violation of measurement assumptions introduce imprecision in the process.

Table 2.1 – Current approaches to analyse data from affective responses to stimuli.

Approach	Stimuli	Data collection	Data analysis	Source
AE	Adjectives and object concepts.	Self-report questionnaire.	Cluster analysis, ranking.	Barnes et al., 2008.
AE	SD, physical components or CAD models.	Self-report questionnaire.	PCA, multivariate regression and correlation.	Barnes and Lillford, 2009; Henson et al., 2006.
KE type II - hybrid kansei	SD and product component images.	Self-report questionnaire.	Factor analysis and multivariate regression.	Matsubara and Nagamachi, 1997; Chen et al., 2008.
KE	SD and product samples.	Self-report questionnaire.	PCA and partial least square.	Nagamachi et al., 2008
KE type I	SD or sentences, product samples, product images and prototypes.	Self-report questionnaire.	Quantification theory types I, II, III.	Nagamachi, 1995; Komazawa and Hayashi, 1976; Hirata et al., 2004; Nagamachi, 2008.
KE type II	Images obtained from morphological analysis. Data bank based on word pairs.	Results are compared to responses from a questionnaire.	Neural network and genetic algorithm.	Su et al., 2008.
KE	SD and product samples.	Self-report questionnaire.	PCA and neural network.	Ishihara et al., 1995, 1997.
KE type III	Adjectives and product samples.	Pseudo-data generated by neural network.	Fuzzy set theory.	Shimizu and Jindo, 1995; Hotta and Hagiwara, 2005.
KE type VIII	SD and product samples.	Self-report questionnaire.	PCA and rough set theory.	Nishino et al., 2001; Nagamachi et al., 2006; Nagamachi, 2008; Okamoto et al., 2007.
Self-confrontation	Visual information.	Continuously through a special device.	Not specified.	Laurans et al., 2009.
PrEmo	Visual cartoons and images of products.	Discrete responses using a scale.	Cluster analysis and MANOVA.	Desmet, 2004.
Affective design	Adjectives and images of product components.	Self-report questionnaire.	Ordinal logistic regression.	Zhou et al., 2008
Consumer-product attachment	Questions on relevant products.	Structured questionnaires.	Confirmatory factor analysis, PCA and regression.	Schifferstein and Zwartkruis-Pelgrim, 2008.

2.2 INACCURACIES IN THE AFFECTIVE ENGINEERING PROCESS

2.2.1 The Semantic Differential Method

SD is a rating methodology developed originally for the investigation of connotative or metaphorical meaning of objects or concepts. The procedure was introduced by Osgood (1952) and compiled by Osgood et al. (1957) and since then its use as a psychological research instrument has been firmly established. The SD measures people's reaction to stimulus words in terms of rating on bipolar scales defined with contrasting adjectives at each end. Originally, the method involved discriminating different adjectives on a scale with seven divisions to which scores can be allocated and correlations can be established between the scales. Using factor analysis, Osgood and Suci (1955) were able to reduce the scales to three primary dimensions of connotative meaning called semantic space.

In the 1970s the SD method was used as a well-known technique for assessing peoples' attitude (Heise, 1969). Some of the reasons for the popularity of the method were lower costs than other methods, instant readiness and the assumption of various objects measured on the same scale could reliably be compared. The popularity of the SD method had as a consequence applications in a number of domains including Nagamachi's studies.

2.2.2 Issues in the Semantic Differential Method

Many applications of SD have distinguished themselves radically from the purposes for which the method was originally devised and many of these pose special problems for data analysis. Some studies were designed on particular characteristics being researched, rather than designed on adjectives from the Osgood's original theoretical framework (McKennell and Bynner, 1969), such as some studies of attitudes to companies' brand and responses to advertisement (Mindak, 1961). Those bespoken constructs used words and phrases based on the content analysis of the companies' tests with their consumers. However, Heise (1969) stated that tailoring the SD to a new area of application (other than that by Osgood) requires a rigorous research design, and some substitutes can yield instruments which are distorted and may stand only a metaphorical relationship to the concepts.

Furthermore, Coxhead and Bynner (1981) raised the question of whether the application of factor-analytic techniques might yield quite different factor structures for different concepts. A single factor-structure produced by some kind of averaging procedure may have no validity for some concepts and so turn the comparison of the concepts suspect or meaningless in terms of factor scores obtained for the factor dimensions. This problematic issue could be associated with different applications than those originally presented by Osgood, in which their assumptions were verified.

Heise (1969) re-examined some of the key methodological procedures and pointed out four main sources of concerns for SD: bipolarity, scale intervals, sources of variance and dimensionality. Those concerns are indeed legitimate not only for the SD method but also any construct that aims to measure differences between individuals.

2.2.3 Bipolarity

Most of the adjectives used in SD are true linguistic contrasts (Deese, 1964). It is assumed in SD studies that linguistic opposites offer means for forming scales which define basic affective contrasts. In other words, if the scales fulfil the assumptions of functional antonyms, then two contrasting adjectives plotted in the SD space would be equidistant from the origin point and they also would be opposite one another so that a line passing between them would pass through the centre of the plot.

However, the bipolarity assumption has not been justified for scales that use SD research in AE. In many cases the scales might not meet the assumption of true bipolarity and their use can distort measurements of the SD structure. Mordkoff (1963), for example, evaluated the functional antonyms of 16 adjective-pairs and found that some of them were not true affective contrasts, such as masculine-feminine, hard-soft and complex-simple. According to Mordkoff, the accuracy of the method is related to the degree to which the underlying assumptions are fulfilled. Therefore, the difficulty for analysts to make inferences might not be completely due to inadequacies of the methodology, but may to some extent reflect the failure of the data to fulfil some fundamental measurement assumption.

2.2.4 The Assumption of Equal Intervals of Categorical Scales

A further problem concerning SD is that data are coded numerically assuming equal interval on scales which pass through the origin of the SD space. The difficulties of assuming equal intervals (i.e., that categorical scales produce interval data) have been known since the middle 1940s (Stevens, 1946) (see Section 2.3).

Likert (1932) proposed the well-known five category agreement scale. These are displayed equally spaced and equally sized on a response form. Likert worked out the scale for the summation and averaging of the scale responses. The intention was to convey to the respondent that these categories are of equal importance and require equal attention.

However, Linacre (2002a) has emphasised that since an analyst is always uncertain of the exact manner in which a particular rating scale will be used by a particular sample, it is always worth investigating its functioning.

2.2.5 Sources of Rating Variance

Sources of variance could also raise inaccuracies (Borsboom, 2006). Differences are desired when eliciting affective responses to characterise a product. Nevertheless, it is necessary to distinguish what differences are relevant for eliciting the theoretical attribute. Actual variations in affective responses stemmed from individual differences and temporal changes are expected. However, variance can also be derived from other factors such as biased items and imprecision of a measurement structure. Biased errors are related to important differences between persons in scale-checking styles. For example, some respondents use the end points of scales more often and avoid the intermediate discriminatory positions or contrariwise. Furthermore, systematic errors can raise variance as a consequence of differences between sex, age groups or cultural groups. Random errors can be related to the imprecision of the instrument, such as reliability, factor scores and group means.

2.2.6 Dimensionality

Osgood et al. (1957) understood that true verbal opposites are defined by straight lines that cross through the origin. If the straight lines were at orthogonal angles to each other, then they are considered independent. Osgood (1964) stated that it is necessary to determine the interrelations of a large and representative sample of qualitative dimensions defined by verbal opposites and then observe if they do fall into natural clusters or factors which can serve as reference coordinates.

However, such factors are subject to the circumstances of sampling (see Section 8.2). Thus, the analysis becomes a hypothesis whose confirming procedure is its replicability at the same domain with rules of sampling not influenced by the factors previously found and nevertheless, the same factor structure occurs (Osgood, 1964). Following this assumption, the original Osgood’s factor analyses of the SD pointed to three major dimensions of rating response and denoted by him as evaluation, potency and activity (EPA). Some studies took those EPA dimensions into account as though additional dimensions do not exist or even they demonstrated through factor analysis that other dimensions accounted just a little for the structure’s overall variance. However, that fact can be associated with the sample size and the sources of rating variance.

Regarding the rating variance, the dimensionality of the semantic space can vary as a function of the individuals who take part in an experiment (Heise, 1969). Heise stated that the affective responses of individuals do vary along dimensions of EPA, but some persons engage in more affective differentiation and some persons less than the simple three-factor structure indicates. Therefore, there could be more meaningful dimensions due to respondent differences than due to stimuli differences. In other words, the instrument could not be stable enough for its purpose if it falls out of the EPA dimensions.

Additionally, the meanings of the scale words (e.g., kansei words) change depending on the environment and context provided by the attribute and by the stimuli, and since the meanings are different, the scale's factorial composition might be as well. Thus, a mistaken choice of affective (kansei) words can lead to completely misleading factor-analytic results.

On the other hand, Oskamp (1977) recommended using the first dimension denoted evaluation as the most important indicator of attitude toward the object since such a dimension is clearly an affective dimension. Osgood (1970), based on previous writings on attitude studies, proposed just the evaluation dimension be sufficient for measuring attitudes[4]. However, this issue is distant from a consensual understanding. For example, Laurans et al. (2009) consider that self-report instruments for measuring affective responses based on simple dimensional instruments often lack inspirational value for design-oriented research, especially in early phases of the design process. Indeed, the problem goes beyond it. The empirical relationships between design elements and consumer's affective response based on the assumption of linear correlations mislead their inferences and conclusions.

From a different point of view, Heise (1969) suggested although the basic metric assumptions for the SD method are not accurate, violations of the assumptions could not be serious enough to interfere with many applications of the method. Heise assumed some metric errors would be expected to counteract one another when ratings on several different scales are added together to form factor scores. This assumption, however, goes against that of independence of the instrument used for measuring with regard to the phenomenon being measured.

[4] Heise (1969) stated evaluation is only sometimes a pure dimension, but it is very often a compound dimension embracing evaluation, potency and activity. Thus, Heise understood that these three major dimensions yield much more information about the character of responses than alternative measures that depend on unidimensionality.

2.2.7 Non-Linearity of Persons' Affective Responses

Concerns about the lack of linearity in scales used by AE analysts and its consequences have been reported in some publications. Ishihara et al. (1995) reported that statistical procedures in KE based on multiple regression methods have presented shortcomings since they assume that the predictors hold linear correlations with all of the other predictors.

However, the matter has led analysts to different discernments. Nagamachi (2008) and Nagamachi et al. (2006) have stated that since affective responses have in general non-linear characteristics, methods that support non-normal distribution should be used in the analysis. In effect, regression models, which are commonly applied in kansei analysis, do not assume that the predictors have to be normally distributed even though they assume that the residuals in the model are random, normally distributed variables with mean equal to zero. This means that the differences between the observed data and the expected data by the model shall be very close to zero. Therefore, the concerns lie on the residuals rather than on the predictors themselves.

Different solutions have arisen from the diverse understandings on the matter. Research on the acquisition of kansei responses to products has been disposed towards techniques to deal with uncertainty and non-linearity in data analysis, such as neural networks, fuzzy logic and rough set theory mentioned in Table 2.1. However, the current methods to overcome the problems related to non-linearity are merely sophisticated solutions for clustering data but not for measuring consumers' responses to products. Therefore, some measurement system has to be devised and be subject to metrological rules, which shall, according to Rossi (2007), ensure traceability and some control under uncertainty.

2.3 AFFECTIVE ENGINEERING AND MEASUREMENT THEORY: A PARALLEL EVOLUTION

The evolution of measurement of attitudes and the evolution of emotional approach to products, which is described in this chapter mainly in terms of AE, have trailed parallel paths. This is illustrated by the timeline in Figure 2.1. Starting with the Thurstone's publication in 1927, the timeline follows the evolution of attitudes measurement at its left-hand side. The evolution of the emotional approach to products is depicted at the right-hand side. Although the timeline does not intend to be a comprehensive representation of the entire universe of studies on the subjects, it represents enough milestones for identifying their developments.

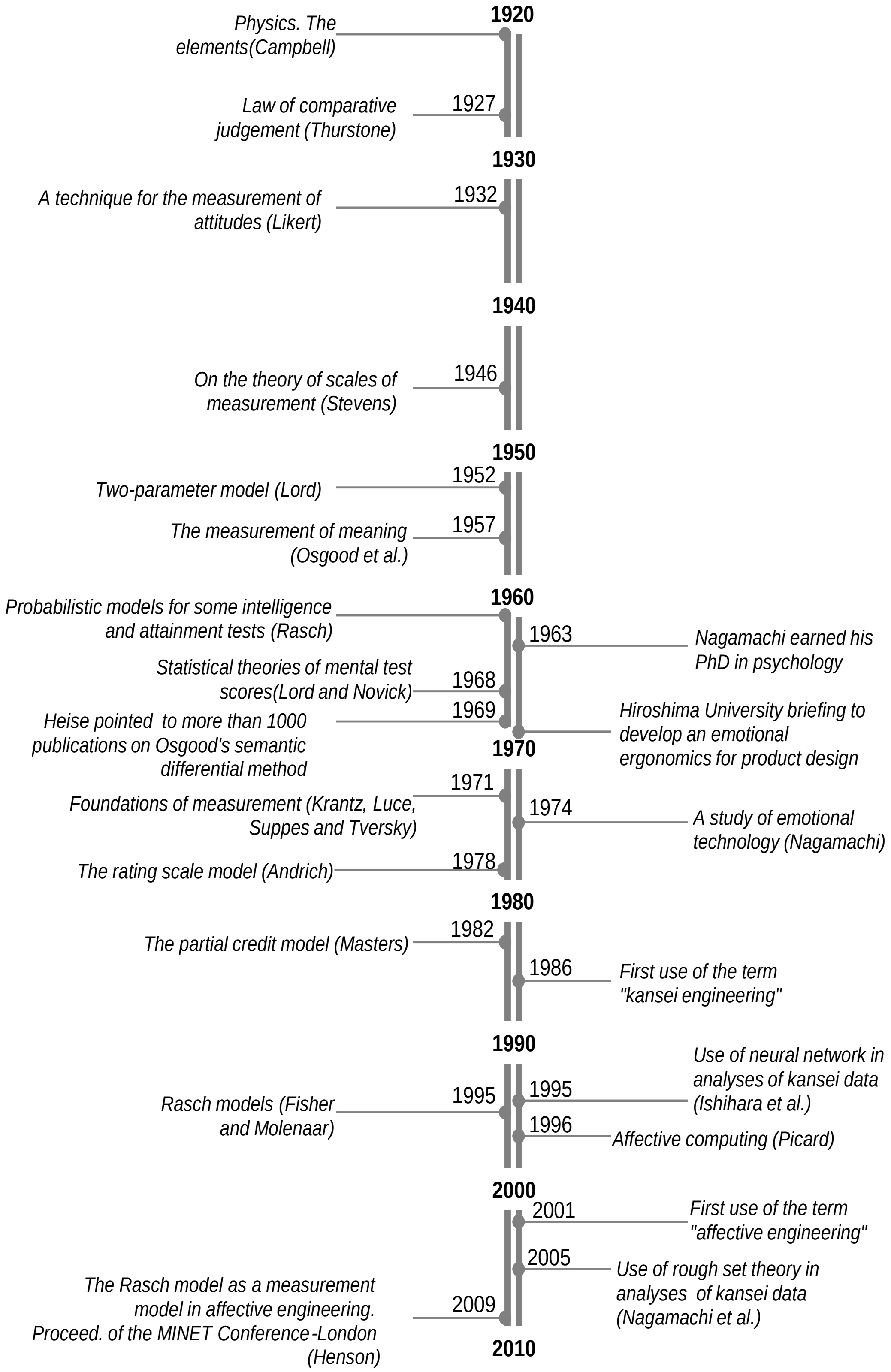

1920
Physics. The elements(Campbell)
1927
Law of comparative judgement (Thurstone)
1930
1932
A technique for the measurement of attitudes (Likert)
1940
1946
On the theory of scales of measurement (Stevens)
1950
1952
Two-parameter model (Lord)
1957
The measurement of meaning (Osgood et al.)
1960
Probabilistic models for some intelligence and attainment tests (Rasch)
1963
Nagamachi earned his PhD in psychology
1968
Statistical theories of mental test scores(Lord and Novick)
1969
Heise pointed to more than 1000 publications on Osgood's semantic differential method
Hiroshima University briefing to develop an emotional ergonomics for product design
1970
1971
Foundations of measurement (Krantz, Luce, Suppes and Tversky)
1974
A study of emotional technology (Nagamachi)
1978
The rating scale model (Andrich)
1980
1982
The partial credit model (Masters)
1986
First use of the term "kansei engineering"
1990
1995
Rasch models (Fisher and Molenaar)
1995
Use of neural network in analyses of kansei data (Ishihara et al.)
1996
Affective computing (Picard)
2000
2001
First use of the term "affective engineering"
2005
Use of rough set theory in analyses of kansei data (Nagamachi et al.)
2009
The Rasch model as a measurement model in affective engineering. Proceed. of the MINET Conference -London (Henson)
2010

Figure 2.1 - Concise timeline of affective (kansei) engineering and measurement of attitude.

Many authors have credited the studies of Louis Thurstone as precursors of the modern psychometrics. Thurstone proposed measuring the separation between two opinions on an attitude scale and then, testing the validity of the scale continuum by means of its internal consistency (Thurstone, 1927). The main point in Thurstone's idea is the possibility to measure attitude associated with a collection of sensory stimuli based on a series of pairwise comparisons.

However, the work of Norman Campbell on the theory of measurement for the physical sciences (Campbell, 1920) had a vast influence on scientists at that time. Campbell stated that a necessary condition for measurement was that the attributes must be additive and therefore, non-additive psychological attributes were essentially impracticable. Stevens (1946) tried to conciliate the understanding on that necessary condition for measurement in his paper "*On the theory of scales of measurement*" asserting that

> *"Perhaps agreement can better be achieved if we recognize that measurement exists in a variety of forms and that scales of measurement fall into certain definite classes. These classes are determined both by the empirical operations invoked in the process of 'measuring' and by the formal (mathematical) properties of the scales. Furthermore - and this is of great concern to several of the sciences - the statistical manipulations that can legitimately be applied to empirical data depend upon the type of scale against which the data are ordered (Stevens, 1946)."*

Later, in a more general theory, namely conjoint measurement, Luce and Tukey (1964), demonstrated that non-geometric properties could be quantified, including psychological attributes (see Section 2.4.1). A comprehensive study on measurements was elaborated by Krants, Luce, Suppes and Tversky (1971). They stated measurement theory deems that the instruments for measuring and the attributes being measured are distinct entities. In order to draw conclusions, one must take into account the nature of the associations between the attribute and the measurements.

Such a statement addresses the idea of fundamental measurement. The elementary concept of fundamental measurement is purely to allow mathematical operations of addition and subtraction on measures (Andrich, 1988a) (see Section 2.4). That is, a scale must show valid evidence for a one-to-one relationship between the structure of mathematical operations on real numbers and the properties of the attribute that is measured.

2.3.1 Overcoming the Measurement Limitations in Affective Engineering Using the Rasch Model

Such as it has been argued in this thesis, a model based on CTT has limitations to meet the assumption for measurement structures (see Section 1.4). Nevertheless, the core concepts of CTT are well-known by scientists and by practitioners (e.g., true scores, random error and reliability) as a consequence of an influential treatise on psychological measurement elaborated by Lord and Novick (1968). However, the popular statistical procedures in CTT are prone to misinterpretation; for example, the mistaken understanding that true score is the construct score, that random error is associated with irrelevant variance and that reliability is a characteristic of the measurement structure. Borsboom (2005) has argued that a meaningful interpretation of true scores as a stable property and random error demonstrating unsystematic variance is philosophically unsustainable.

A key consideration in any measurement structure is the property of invariant comparisons (Andrich, 1988a). Invariant comparisons are established when the comparison between two items are independent of the particular sample taken as instrumental and the relative difference between any two persons is independent of the relevant items taken as elements for comparison (Rasch, 1961). Although Thurstone (1928) formerly expressed these properties of measurement structures of latent variables in his work, it was George Rasch who derived such properties as a probabilistic model.

The RM presents unique properties that fulfil the requirement for an empirical test of quantitative framework. The concept of specific objectivity proposed by George Rasch (see Section 2.4.3) gives independence of measurement from the investigator (Rasch, 1960, 1980; Andrich, 1988a) and from the instrument used to measure the latent variable. Such unique properties operationalised by the rationale for calibrating items in questionnaires could solve problems related to scales of measurement designed to interpret results from affective responses based on multivariate analysis.

2.4 RASCH MEASUREMENT THEORY

2.4.1 Fundamental Measurement and the Rasch Model

Luce and Tukey (1964) proposed an axiomatic approach denoted conjoint measurement[5] to adjust non-physical objects or phenomena to Campbell's concept of fundamental

[5] Luce and Tukey axiomatic approach has also been referred to as additive conjoint measurement.

measurement, which is based on the concatenation of the objects measured (Campbell, 1920, 1928)[6]. They replaced the concatenation operation by axioms that provide simultaneous measurement on interval scales for each kind of quantity separately and for their joint effects. They drew the conclusion from their axioms that it is possible to quantify effects from or responses to latent variables in a two-way matrix such that the observed ordering of the cells is preserved by the natural ordering of the numbers assigned to the responses. Luce and Tukey (1964) stated the measure for any cell is the sum of a function of its row element and another function of its column element, and those functions are unique up to the positive linear transformation of interval measurement.

The applicability of conjoint measurement has remained as a subject of discussion although it is also acknowledged as an important theoretical contribution in a range of different fields of knowledge (Perline et al, 1979). The reason for the difficulty of application is that conjoint additivity axioms, representing interval measurement, are deterministic formulations. On the other hand, the RM expresses a stochastic formulation of additive conjoint measurement, employing the available tools of standard statistics (Karabatsos, 2001).

Nevertheless, Karabatsos argued that while the RM conveniently uses standard fit statistics, its context of application is data dependent and therefore, prone to absorb data containing measurement disturbances. However, this seems to be a consequence of the limitations of addressing the nature of the measurement problems, such as the absence of well-established theories of psychological attributes (Sijtsma, 2010), rather than essentially being a consequence of the statistical properties of the RM.

The RM has been referred to as the operationalisation of the Luce and Tukey's approach (Fischer, 1995; Scheiblechner, 1999). For example, Brodgen (1977) stated the RM is a special case of additive conjoint measurement. Given that the probability of a person *n* to give a rating on item i (π_{ni})provides a required rank ordering, then the additive conjoint measurement will be possible if $T(\pi_{ni}) = \beta_n - \delta_i$, where T is an order-preserving transformation ofπ_{ni}, β is a person parameter and δ is an item parameter. If items and persons are thoroughly graded and the double cancellation axiom[7] and conjoint

[6] A summary of Campbell's fundamental measurement can be found in Reese (1943).

[7] The double cancellation axiom is concerned with a class of relations in a two-way matrix where the common terms of two antecedent inequalities cancel out to produce a third inequality (Luce and Tukey, 1964; Michell, 2009). Michell (1988) stated that upon existence of weak orders in 3 × 3 sub-matrices of an M x N matrix solely independence and the fulfilment of the double cancellation axiom are necessary and sufficient conditions for additivity in a measurement structure. However, Van der Linden (1994) argued that those conditions are necessary but not sufficient for the underlying model to hold. He based his argument on the deterministic characteristic of the conjoint measurement, though.

measurement is satisfied (Luce and Tukey, 1964; Krantz et al., 1971), then a transformation satisfying $T(\pi_{ni}) = \beta_n - \delta_i$ can exist for all *n* and *i* (Brodgen, 1977). Following this concept, Perline et al. (1979) corresponded indicators of fit the RM with the results of certain conjoint measurement axiom tests, demonstrating that the RM satisfies the concept of a form of fundamental measurement and therefore, holds the measurement property of additivity.

2.4.2 Separability of Parameters

Fundamental measurement requires independence between the estimates of items parameters and the characteristics of the sample. Thurstone (1927) noted that one of the requirements to construct a rational assignment of values for measuring attitude is that the scale ought to be entirely independent from the actual opinions of persons. This statement just follows the measurement process in physics where an instrument does not depend on either the objects or phenomena being measured or the agent that is measuring them. Thus, Thurstone concluded, the estimates in social sciences for measuring attitude ought to be free from the transient characteristics of the samples used for calibration of the instrument (Thurstone, 1928).

The separability of parameters is an essential property of the RM which allows the independence between objects (i.e., persons) and agents (i.e., items). This property is a determinant condition for achieving objective measurement, which other IRT models have failed in demonstrating (Fisher and Molenaar, 1995). In practice, the independence between persons' responses and the measurement instrument is manifested statistically if the parameters present sufficient statistics[8] (Andersen, 1973). From this basis, linearity can be constructed and algebraic operations can meaningfully be realised.

2.4.3 Specific Objectivity

George Rasch developed his model based on the multiplicative Poisson model, profoundly detailed in his seminal book (Rasch, 1960, 1980). Rasch applied the Poisson distribution to determine the probability that a person *n* in a given time reads a_{ni} words of a text *i*, such that

$$\Pr(a_{ni}) = e^{-z_n E_i} \frac{(z_n E_i)^{a_{ni}}}{a_{ni}!} \tag{2.1}$$

[8] The meaning of term *sufficient statistics* in Rasch measurement theory is explored in Section 2.4.5.2.

where $\lambda_{ni} = z_n E_i$ "*indicates that many words are read in a given time*," for a high value of z_n representing "*high reading speed*" [9] of the person n and a high value of E indicating a text which can be "*read quickly*" (Rasch, 1977). Similarly, the probability that a person reads a_{nj} words of a test j in a similar context is given by

$$\Pr(a_{nj}) = e^{-z_n E_j} \frac{(z_n E_j)^{a_{nj}}}{a_{nj}!} \tag{2.2}$$

The multiplication of probabilities implies that the outcomes a_{ni} and a_{nj} of the two tests is such that

$$\Pr\{a_{ni}, a_{nj}\} = \Pr(a_{ni})\Pr(a_{nj}) = e^{-z_n(E_i+E_j)} z_n^{(a_{ni}+a_{nj})} \frac{E_i^{a_{ni}} E_j^{a_{nj}}}{a_{ni}!a_{nj}!} \tag{2.3}$$

As a property of the Poisson distribution, the sum of the two Poisson distributed variables is also Poisson distributed with a parameter which is the sum of the two parameter values (Rasch, 1977). Attributing the notation $a_{n+} = a_{ni} + a_{nj}$, then

$$\Pr\{a_{n+}\} = e^{-z_n(E_i+E_j)} \frac{z_n^{a_{n+}} (E_i + E_j)^{a_{n+}}}{a_{n+}!} \tag{2.4}$$

If the total of words read a_{n+} has an established value, the probability of the outcomes a_{ni} and a_{nj}, conditional on this total, is given by dividing Equation 2.4 into Equation 2.3 (Rasch 1977). This conditional probability cancel out those two factors and therefore, it does not contain the person parameter z_n. "*Thus the conditional probability is governed by the observed numbers a_{ni} and a_{nj} and by the ratio between the difficulty parameters of the two tests, while it is not influenced at all by which person is involved* (Rasch, 1977)." Thus,

$$\Pr\{a_{ni}, a_{nj} | a_{n+}\} = \binom{a_{n+}}{a_{ni}} \left(\frac{E_i}{E_i + E_j} \right)^{a_{ni}} \left(\frac{E_j}{E_i + E_j} \right)^{a_{nj}} \tag{2.5}$$

Generalising, the conditional probability only holds for a certain class Ω of objects κ_i interacting with a class Ψ of agents ν_n, being a a fixed number of responses. The interaction of classes of objects and agents is denoted *frame of reference*. As a consequence of the model's property of separability of parameters and within the frame of reference, invariance holds for any set of elements $\kappa_1, \kappa_2, ..., \kappa_i \in \Omega$ and for any set of elements $\nu_1, \nu_2, ..., \nu_n \in \Psi$, qualifying it as *specific objectivity* (Rasch, 1968, 1977).

[9] Terms used by Rasch (1977).

George Rasch understood that if the objectivity is to be specific, this specificity ought to be fully fulfilled and the parameters ought to be in the same dimension (Rasch, 1968). In other words, "…*if a set of empirical data cannot be described by…* [the] *model* [,] *then complete specifically objective statements cannot be derived from them* (Rasch, 1968)." Thus, if a data set fails to fit the model, then any set of person parameters will depend on which other persons are compared and therefore, it will not be invariant.

2.4.4 The Assumption of Local Independence

Local independence is a core assumption of the RM. This assumption is met when the probability of a response to any item is independent of the results with any other item after accounting for person estimate (Smith, 2005). Estimates of parameters can be misleading in the presence of local dependence (Chen and Thissen, 1997). Marais and Andrich (2008a) distinguish the violation of the assumption in two situations. One of them is the violation of unidimensionality, called trait dependence. Another situation is a type of specific statistical dependence denoted response dependence.

Trait dependence, also called multidimensionality, is identified in scales containing items developed for measuring a single attribute although there are sub-sets of items measuring somewhat different aspects of the attribute. In the domain of product design, for example, an analyst could be interested in the persons' impression about different materials used as stimulus objects although respondents are influenced by the experiment conditions, such as intensity of light on the objects. In addition, items out of the considered context could stimulate different aspects of the users' experience other than the relevant attribute which an analyst wants to know about. A questionnaire could, for example, contain items associated with visual elements of a product while the analyst is exclusively interested in the persons' tactile interaction.

Response dependence is identified when a person's response to an item in a scale interferes with his or her response to another item within the same scale. In RMT each item (i.e., statement, adjective or question in the domain of product design) is valid as an independent item. According to Marais and Andrich (2008b) response dependence can, for example, be found in satisfaction questionnaires where a positive rating of a respondent depends on the responses to the preceding items and where that rating will interfere in the way that the responses on the following items are rated (Wilson et al., 1997). One of the sources of response dependence is the redundancy of items (Smith, 2005). Redundant items can mislead inferences or decision made on account of means and standard deviations. Redundant items have the effect of inflating reliability indices and item discrimination estimates. Response independence is formalized by Marais and Andrich (2008b) as follows:

$$\Pr\{((X_{ni})|\beta,\delta)\} = \prod_n \prod_i \Pr\{x_{ni}\} \tag{2.6}$$

where $(X_{ni})$is the matrix of responses for $n = 1,...,N, i = 1,...,I$.

2.4.5 The Rasch Model

2.4.5.1 The Rasch dichotomous model

The RM expresses the probability that a person will endorse an item with two-category responses (e.g., yes or no, true or false, agree or disagree) as a logistic function of the difference between the person's location[10] (β) and the item's location (δ) on a linear continuum (Rasch 1960, 1980), represented by Equation 2.7,

$$\Pr\{X_{ni} = x|\beta,\delta\} = \frac{\exp[x(\beta_n - \delta_i)]}{\gamma_{ni}} \tag{2.7}$$

given that $\gamma_{ni} = 1 + \exp(\beta_n - \delta_i)$, where $x \in \{0,1\}$, taking *1* as a positive response and *0* otherwise. $\Pr\{X_{ni} = x \ = 1|\beta,\delta\}$ is the probability that a person *n* will endorse an item *i*, such that $0 \le \Pr\{x_{ni} = 1\} \le 1$ and $-\infty \le (\beta_n - \delta_i) \le +\infty$. The relationship between the difference in person locations on the continuum and the probability of a positive response, denoted in RMT as item characteristic curve (ICC), is indicated in Figure 2.2.

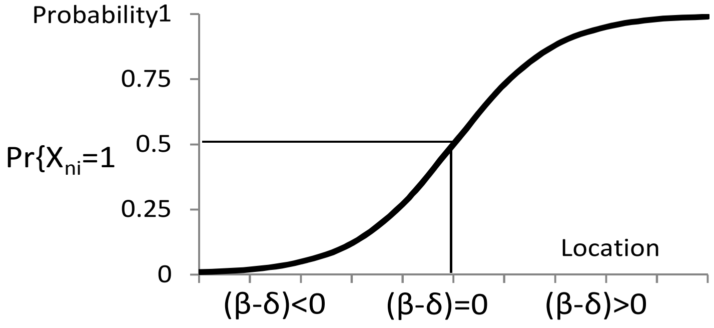


Figure 2.2 – Probability of a positive response associated with persons' locations on the continuum.

[10] The parameter β is denoted person ability in current Rasch literature. These terms were adapted to facilitate its application and understanding in the domain of product design (see Section 2.5). The parameter δ is denoted item difficulty in Rasch literature. Its adaptation follows the precedent justification.

Taking the inclination of endorsement of a person *n* as β_n and the difficulty of endorsement of an item *i* as δ_i and letting a positive response be represented by *1* and a negative response be represented by *0*, then the probability of responses of the person *n* to give a rating to three dichotomously scored items, *i* = 1, 2 and 3 can be given such as in Table 2.2.

Table 2.2 – Joint probability for three dichotomous items.

Item 1	Item 2	Item 3	Total score $X_n = x_{n1} + x_{n2} + x_{n3}$	Joint probability
1	1	1	3	$\frac{1}{\gamma_{n1}\gamma_{n2}\gamma_{n3}}\left(e^{(\beta_n-\delta_1)}\right)\left(e^{(\beta_n-\delta_2)}\right)\left(e^{(\beta_n-\delta_3)}\right)$
1	1	0	2	$\frac{1}{\gamma_{n1}\gamma_{n2}\gamma_{n3}}\left(e^{(\beta_n-\delta_1)}\right)\left(e^{(\beta_n-\delta_2)}\right)$
1	0	1	2	$\frac{1}{\gamma_{n1}\gamma_{n2}\gamma_{n3}}\left(e^{(\beta_n-\delta_1)}\right)\left(e^{(\beta_n-\delta_3)}\right)$
0	1	1	2	$\frac{1}{\gamma_{n1}\gamma_{n2}\gamma_{n3}}\left(e^{(\beta_n-\delta_2)}\right)\left(e^{(\beta_n-\delta_3)}\right)$
1	0	0	1	$\frac{1}{\gamma_{n1}\gamma_{n2}\gamma_{n3}}\left(e^{(\beta_n-\delta_1)}\right)$
0	1	0	1	$\frac{1}{\gamma_{n1}\gamma_{n2}\gamma_{n3}}\left(e^{(\beta_n-\delta_2)}\right)$
0	0	1	1	$\frac{1}{\gamma_{n1}\gamma_{n2}\gamma_{n3}}\left(e^{(\beta_n-\delta_3)}\right)$
0	0	0	0	$\frac{1}{\gamma_{n1}\gamma_{n2}\gamma_{n3}}$

2.4.5.2 Sufficient statistics

The RM establishes the significance of the simple total score for a person as a consequence of the specification of the interaction between a person and an item. For example, based on Table 2.2, let the responses of a person to the items be *1*, i.e., one responds just to one item positively, then the probability does not depend on the inclination of the person, but only the relative difficulty of endorsement of the items. This is demonstrated by Equation 2.8 (Rasch, 1960, 1980).

$$\Pr\left\{\left(x_{n1} = 1, x_{n2} = 0, x_{n3} = 0\right)\middle| X_n = 1\right\} = \frac{\frac{1}{\gamma_{n1}\gamma_{n2}\gamma_{n3}} e^{(\beta_n-\delta_1)}}{\frac{1}{\gamma_{n1}\gamma_{n2}\gamma_{n3}}\left(e^{(\beta_n-\delta_1)} + e^{(\beta_n-\delta_2)} + e^{(\beta_n-\delta_3)}\right)}$$

$$= \frac{e^{(\beta_n-\delta_1)}}{e^{(\beta_n-\delta_1)}+e^{(\beta_n-\delta_2)}+e^{(\beta_n-\delta_3)}} = \frac{e^{\beta_n}e^{-\delta_1}}{e^{\beta_n}\left(e^{-\delta_1}+e^{-\delta_2}+e^{-\delta_3}\right)}$$

$$= \frac{e^{-\delta_1}}{e^{-\delta_1}+e^{-\delta_2}+e^{-\delta_3}} \tag{2.8}$$

Similarly,

$$\Pr\{(x_{n1}=0, x_{n2}=1, x_{n3}=0) | X_n = 1\} = \frac{e^{-\delta_2}}{e^{-\delta_1}+e^{-\delta_2}+e^{-\delta_3}} \tag{2.9}$$

and

$$\Pr\{(x_{n1}=0, x_{n2}=0, x_{n3}=1) | X_n = 1\} = \frac{e^{-\delta_3}}{e^{-\delta_1}+e^{-\delta_2}+e^{-\delta_3}} \tag{2.10}$$

Thus, if data fit the RM, the total score of a person is a sufficient statistic to provide all the information about the persons' degree of endorsement of the items. By a symmetrical argument, it can be demonstrated that the total score is a sufficient statistic for all the information about the item difficulty.

Similarly, the equations can be extended for a total score of $X_n = 2$, such that

$$\Pr\{(x_{n1}=1, x_{n2}=1, x_{n3}=0) | X_n = 2\} = \frac{e^{-\delta_1-\delta_2}}{e^{-\delta_1-\delta_2}+e^{-\delta_1-\delta_3}+e^{-\delta_2-\delta_3}} \tag{2.11}$$

$$\Pr\{(x_{n1}=1, x_{n2}=0, x_{n3}=1) | X_n = 2\} = \frac{e^{-\delta_1-\delta 3}}{e^{-\delta_1-\delta_2}+e^{-\delta_1-\delta_3}+e^{-\delta_2-\delta_3}} \tag{2.12}$$

and

$$\Pr\{(x_{n1}=0, x_{n2}=1, x_{n3}=1) | X_n = 2\} = \frac{e^{-\delta_2-\delta_3}}{e^{-\delta_1-\delta_2}+e^{-\delta_1-\delta_3}+e^{-\delta_2-\delta_3}} \tag{2.13}$$

The argument that the total score of a person is sufficient statistic emerges from a property of the model. Therefore, to obtain all the relevant information for estimating a person's inclination of endorsement to a latent attribute, it is an essential condition that the data fit the model.

2.4.5.3 *The Rasch polytomous model*

George Rasch understood that the possibility of separating two parameters is a fundamental property of a class of models[11] (Rasch, 1977). He proposed an extension of the

[11] A model is usually considered to be part of the Rasch family of models if it holds the properties of separability of parameters, specific objectivity, statistical sufficiency and additivity. One of the models in

dichotomous model for polytomously-scored responses on the basis of a unidimensional form (Rasch, 1961), such that the probability of person *n* responding in the category *x* to item *i* is given as

$$\Pr\{X_{ni} = x|\beta,\delta,k,\phi\} = \frac{\exp[k_x + \phi_x(\beta_n - \delta_i)]}{\sum_{k=0}^{m}\exp[k_x + \phi_k(\beta_n - \delta_i)]} \tag{2.14}$$

where $x \in \{0,1,...m\}$, k_x is a parameter associated with the category *x*. Rasch denoted ϕ as a non-parametric scoring coefficient.

Andersen (1977) understood that for the raw score to be a sufficient statistic for β in Equation (14) the coefficients $\phi_o, \phi_1, ..., \phi_m$ ought to be taken as successive integers. Later, Andrich (1978) interpreted the category parameters k_x, denoting it as threshold parameter. He considered that the responses' categories are successive alternatives on a rating scale, such that a threshold is the transition between two consecutive categories. For example, on a scale with three categories; disagree, neutral and agree, let τ_0 represent the first threshold of item *i*, i.e., the location on the continuum at which a person is equally likely to choose the options disagree or neutral. The second threshold τ_1 represents the location at which a person is equally likely to choose the option neutral or agree. In the case of three categories the model is given by

$$\Pr\{x_{ni} = 0\} = \frac{e^{0(\beta_n - \delta_i)}}{e^{0(\beta_n - \delta_i)} + e^{-\tau_{1i} + 1(\beta_n - \delta_i)} + e^{-\tau_{1i} - \tau_{2i} + 2(\beta_n - \delta_i)}} \tag{2.15}$$

$$\Pr\{x_{ni} = 1\} = \frac{e^{-\tau_{1i} + 1(\beta_n - \delta_i)}}{e^{0(\beta_n - \delta_i)} + e^{-\tau_{1i} + 1(\beta_n - \delta_i)} + e^{-\tau_{1i} - \tau_{2i} + 2(\beta_n - \delta_i)}} \tag{2.16}$$

and

$$\Pr\{x_{ni} = 2\} = \frac{e^{-\tau_{1i} - \tau_{2i} + 2(\beta_n - \delta_i)}}{e^{0(\beta_n - \delta_i)} + e^{-\tau_{1i} + 1(\beta_n - \delta_i)} + e^{-\tau_{1i} - \tau_{2i} + 2(\beta_n - \delta_i)}} \tag{2.17}$$

Generalising to any number of categories gives the representation of the Rasch-Andrich model, frequently denoted in literature as the rating scale model (Andrich, 1978; Embretson and Reise, 2000; Bond and Fox, 2007). Thus,

this family is the many-facet Rasch model (Linacre, 1989), which will be further explored in Chapter 4. Other Rasch models are the log-linear Rasch model (Fisher, 1973), the item-bundle model (Wilson and Adams, 1995), the graphical Rasch model (Kreiner, 2007; Kreiner and Christensen, 2004) and the multi-dimensional Rasch model (Briggs and Wilson, 2003). Nevertheless, it is not part of the scope of this thesis to discuss their particularities.

$$\Pr\{X_{ni}=x|\beta,\delta,k\}=\frac{\exp\left[x(\beta_n-\delta_i)-\sum_{k=1}^{x}\tau_k\right]}{\sum_{k=0}^{m}\exp\left[k(\beta_n-\delta_i)-\sum_{k=1}^{x}\tau_k\right]} \tag{2.18}$$

The rating scale model is conditional with regard to distance between category intervals. The model assumes that the thresholds between three or more categories are equally spaced on the continuum. However, it is very likely that most of the cases in applications of the RM in the domain of product design will indicate thresholds with different distances between adjacent categories (see Chapter 3 and Chapter 6).

Masters (1982) developed a model based on Rasch's simple logistic model, which does not parameterise thresholds in terms of equal spaces between adjacent categories on the continuum (Wright and Masters, 1982). He assumed that when the thresholds within an item are taken in sequence, it is possible to infer from a person's response the steps that the person must have taken to arrive at his or her response. Thus, the difficulty of the *kth* step within an item governs the probability of a person responding in category *k* instead of responding in category *k-1*, such that

$$\phi_{nik}=\frac{\pi_{nik}}{\pi_{ni(k-1)}+\pi_{nik}}=\frac{e^{(\beta_n-\delta_{ik})}}{1+e^{(\beta_n-\delta_{ik})}} \tag{2.19}$$

where π_{nik} is the probability of a person n to respond in category k to item i, β_n is the inclination of a person n of endorsement, and δ_{ik} is the difficulty of the k^{th} step in item *i*. Masters denoted his model as the partial credit model, represented by Equation 2.20.

$$\Pr\{X_{nij}=x_{nij}|\beta,\delta\}=\frac{\exp\sum_{j=0}^{x}(\beta_n-\delta_{ij})}{\sum_{k=0}^{m_i}\exp\sum_{j=0}^{k}(\beta_n-\delta_{ij})} \tag{2.20}$$

where $x=0,1,\ldots,m_i$.

In the partial credit model the sufficient statistic is a consequence of the separability of the parameters. This allows independence of sample distribution estimates of step difficulty because the model eliminates the parameter of persons from the estimation equations for the items.

2.4.6 Estimation of Parameters

In Rasch modelling the calibration of items and the persons measures are based on estimation of parameters (Wright and Master, 1982). This is accomplished using inverse probability described originally by Bernoulli (1713). Estimates of the person locations and item locations are preliminarily made according to a rating scale or partial credit structure and then compared with the observations. Estimates are then revised and new estimates are computed. This process of iteration is carried out until the changes of the estimates are smaller than a stopping rule controlled by a convergence criterion[12]. After the estimates have been made, the data are evaluated to determine the extent to which they fit the model. Most of the estimation procedures are based on the method of maximum likelihood (Fisher, 1922). The estimates obtained from this method point to the values of parameters which maximize the likelihood that the observed data would have generated. A benefit of this method is the calculation of the standard error for each estimate through a second derivative of a likelihood function (Linacre, 1999).

There are different procedures of estimation which present benefits and drawbacks. All of them have been widely examined in the literature[13]. Therefore, in this section is solely presented an overview of the general principles of four different procedures. The reason is that two of them are used in the software package Winsteps® and other two are used in the software package RUMM2030®. Both programs (see Section 2.4.7.1) are commercially available for Rasch-dedicated analysis. Although those estimates are statistically equivalent, their approaches are slightly different.

The normal approximation estimation algorithm (PROX)[14] procedure was developed for dichotomous data (Cohen 1979) and extended to polytomous data (Linacre, 1994a, 1995). The procedure assumes that the effects of the sample on calibration of items and person measures can be summarised by means and standard deviations. However, the PROX method assumes that persons and items are normally distributed. This assumption causes some degree of bias because mismatches between the distributional assumption and the data can skew the PROX estimates (Linacre, 1999). PROX is used in Winsteps® as a rough estimate followed by a refinement using joint maximum likelihood estimation (JMLE).

The JMLE method, also known as unconditional maximum likelihood procedure (UCON) (Wright & Panchapakesan, 1969), takes into account that all estimates can be

[12] There are standard convergence criteria which are suitable for most data sets. The convergence criterion used in an estimate procedure is usually associated with extreme scores found in the data set and not with misfit to the model.

[13] Mathematical derivations of the parameter estimates and further information can be found in Wright and Douglas (1977), Wright and Stone (1979), Wright and Masters (1982), Linacre (1999), Andrich (1988b), Andrich and Luo (2003), Luo and Andrich (2005) to mention but a few. A practical demonstration of the algebraic operations based on the JMLE procedure can be seen in Moulton (2003).

[14] Originally, PROX was developed as a non-iterative method used when the data set is complete. Linacre (1994a) derived PROX equations for supporting missing data.

computed concurrently because raw scores are sufficient statistics for persons and items. The estimate of the Rasch parameter take places when the observed raw score for the parameter matches the expected raw score. Linacre (2011) points to some advantages of the JMLE method, such as robustness when dealing with missing data and same measures for all persons with the same total raw score on the same items and for items across the same persons. However, Linacre has also pointed to some disadvantages. For example, measures for extreme scores (i.e., zero score and perfect score) for persons and items require post-hoc estimation and in cases of small samples or few items the estimation can inflate logit distances although, according to Linacre, this rarely exceeds the model's standard error of the measures.

The computer program RUMM2030® has embodied the pairwise maximum likelihood algorithm (PML) to estimate item parameter. In the method, person parameters are cancelled when grouping persons by their total score on each pair of items along with the distribution of persons' rates within those groups to estimate the differences in the difficulties of the two items (Choppin, 1968; Wright and Masters, 1982; Zwinderman, 1995). However, some disadvantages have been pointed out. Linacre (2011), for example, states the method presents asymmetric analysis of person and item parameters. As a consequence, transposing rows and columns changes the estimates.

Andrich and Luo (2003) have employed in RUMM2030® a pairwise technique of re-parameterisation of centralised thresholds (Andrich, 1985a; Andrich, 2010) into their principal components[15]. The first step of the estimation is to retain the parameters of thresholds provisionally. The implicit equations are solved by iteration using the Newton-Raphson algorithm (Andrich, 1978). The second step is to estimate the location of the items, which is the first principal component. Location is distinguished from the thresholds. Thus, RUMM uses two separate constraints. One constraint is that the sum of item locations is equal to zero. Another constraint is that the sum of threshold estimates to be zero[16].

[15] The term principal component is associated with Guttman's scaling (Guttman, 1950) that rearranged ordered thresholds in successive principal components. It is analogous to orthogonal polynomial in regression (Andrich and Luo, 2003) and, therefore, the term does not refer to the commonly used principal component analysis (PCA).

[16] The RUMM algorithm refers to such a constraint as centralised thresholds. When the threshold estimates are derived by adding the location estimate to each centralised threshold, they are then referred to as uncentralised thresholds. The mean of the set of uncentralised thresholds for an item is therefore the location estimate for that item (RUMM2030, 2012).

The second component is generated when there are at least three ordered categories (Andrich, 1985a). In RUMM (Andrich and Luo, 2003) this component is denoted *spread*, representing a half distance between the thresholds. The third component is termed skewness obtained when there are four categories, identifying any deviation from the equidistance structure of the spread between successive thresholds. Pedler (1987) derived a fourth component when at least five ordered categories are present. This component is identified as *kurtosis* (Andrich and Luo, 2003). A general equation that represents the method is given by

$$\Pr\{X_{ni} = x\} = \frac{e^{x\beta_n + \sum_{l=1}^{m_i} f_{li}(x)\omega_{li}}}{\gamma_{ni}} \tag{2.21}$$

where $\omega_i, \{i = 1,\ldots,m_i\}$ represents the principal components and $f_{li}, \{l = 1,\ldots,4\}$ represents their coefficients which are successive polynomials in x. Thus,

$$f_{1i}(x) = -x_i \tag{2.22}$$

$$f_{2i}(x) = x_i(m_i - x_i) \tag{2.23}$$

$$f_{3i}(x) = x_i(m_i - x_i)(2x_i - m_i) \tag{2.24}$$

$$f_{4i}(x) = x_i(m_i - x_i)(5x^2 - 5xm_i + m_i^2 + 1) \tag{2.25}$$

One of the algorithm's properties is that its statistic is a function of the frequencies of all response categories for estimating each threshold, rather than only a function of the frequency of the corresponding category (see Section 2.5.2). This property optimises the stability of the estimates, including cases of few responses in some categories, counts for missing data and generalises to different numbers of categories for different items.

It is noteworthy that once the estimates of items have been obtained a third step taken by RUMM is to obtain the estimates of persons (β_n) based on the total scores without taking into account the extreme scores. The new parameters of persons are kept fixed, repeating the first step and the second step. When two successive sets of solutions for the items and parameters of thresholds differ by less than an arbitrary convergence value (see Note 12), the iterative process is terminated (Andrich, 1978). These estimates are obtained using the weighted likelihood estimation (WLE) (Warm, 1989)[17], which corrects for the bias inherent in the direct maximum likelihood estimation (MLE). The

[17] RUMM2030® provides WLE as the default and MLE as an option.

asymptotic variance of MLE and WLE estimates are the same, meaning that the estimates have the same model standard errors.

2.4.7 Calibration of a Measurement Structure through Rasch Analysis: Developing the Validity of a Metric

2.4.7.1 Computer aided solutions

In contrast with the classical approach for analysing data (e.g., multivariate statistical methods) that uses straightforward computation (Embretson and Reise, 2000), the response pattern in RMT is achieved by an iterative process. This iterative process is laborious (see Section 2.4.6), requiring a computer-intensive solution for practical purposes. Among the commercially available options of software are WINSTEPS® (Linacre, 2011), ConQuest® (Wu et al., 2007) and RUMM2030® (Andrich et al., 2012). Additionally, there are computational programs free of charge. One of them is the software R, which is an integrated suite of software facilities for data analysis that uses add-on packages (currently ~2,500) (Mair and Hatzinger, 2007; Rizopoulos, 2006). The framework of an analysis can have slight differences according to the software used. Nevertheless, Rasch analysis is carried out to identify the degree of discordance between the observed response pattern and the expected pattern, and to verify whether the assumptions of the model are met.

2.4.7.2 Expected response pattern

Rasch analysis examines to what extent the probabilistic form of the Guttman pattern has been achieved in a measurement structure (Guttman, 1950; Andrich, 1985a). The Guttman hierarchical scaling arranges items in an order such that the person who endorses a particular item also endorses items of a lower difficulty level. The RM uses the pattern in a more tractable way, i.e., it considers that there is high probability that a person who endorsed an item also endorsed easier items.

Take as an example a scale with three items and dichotomous responses (e.g., agree or disagree) to a stimulus object where responses were ordered according to the Guttman pattern (Table 2.3). The easiest item (i.e., the item endorsed by comparatively most respondents) is at the left side of the table. Persons were ordered according to their inclination to endorse an item. The least inclined person is on the top of the table. A hypothetical ideal Guttman pattern constitutes a unidimensional set of items such that a given item can predict the responses to all previous items in the set. It is noteworthy that although with empirically independent items, it is very likely that a deterministic

Table 2.3- Example of a Guttman-like scaling for dichotomous items emerged from affective responses.

		Easy endorsement		Difficult endorsement
		Item 1	**Item 2**	**Item 3**
Least inclined to endorse	**Person 1**	**Disagree**	**Disagree**	**Disagree**
	Person 2	**Agree**	**Disagree**	**Disagree**
	Person 3	Disagree	Agree	Disagree
	Person 4	Agree	Disagree	Agree
Most inclined to endorse	**Person 5**	**Agree**	**Agree**	**Disagree**
	Person 6	**Agree**	**Agree**	**Agree**

Note: The deterministic Guttman pattern is represented by responses in bold.

limiting Guttman structure will not be observed in data. This is the case of Person 3 and Person 4 in the example. Then to locate items on a continuum, the probabilistic RM ought to be used (Andrich, 1985a).

2.4.7.3 The arbitrary origin

One of the characteristics of the RM is that parameter estimates are always located on the scale continuum on the basis of an arbitrary zero of the measurement scale. Usually, the arbitrary zero is established as a default of the method applied in an analysis and afterwards as a decision made by the analyst (Wright and Masters, 1982).

The value of zero, in that sense, does not mean an overall lack of attitude to the relevant affective attribute of a product. It merely means that as a person increases his or her inclination to endorse the attribute, the difference between β and δ can pass through zero. Therefore, the origin is nothing else than a convenient point on the continuum of an instrument to measure an object by calculating its relative position through calibration.

2.4.7.4 Expected value

Polytomous items are the most common case when measuring attitudes to an affective attribute of an object. In that case, the general expression for the expect value, or in other words the theoretical mean, is given as follows:

$$E[X_{ni}] = \sum_{x=0}^{m_i} x \Pr\{x_{ni}\} \tag{2.26}$$

where $P\{x_{ni}\}$ is the probability of a response x of a person n on item i determined from Equation 2.18, and m is the maximum score for item i. In the presence of polytomous items the thresholds are incorporated into the model (see Section 2.4.5.3). Thus, the slope of the

ICCs (see Figure 2.2) will be dependent, at least in part, on the distribution of the thresholds along the continuum (see Section 2.4.7.7).

2.4.7.5 Test of fit

Test of fit examines the degree to which the observed responses match in probabilistic terms with the Guttman pattern (see Section 2.4.7.2) in a structure based on the expected values. Thus, the residual is then computed as

$$y_{ni} = x_{ni} - E[x_{ni}] \tag{2.27}$$

such that x_{ni} is the rating of a person n on item i and E_{ni} is the expected value. The variance of x_{ni} can be expressed as

$$V(x_{ni}) = E[x_{ni}^2] - E[x_{ni}]^2 \tag{2.28}$$

such that

$$E[x_{ni}^2] = \sum_{x=0}^{m_i} x^2 \Pr\{x_{ni}\} \tag{2.29}$$

where m is the maximum score for item i. The standardised residual is then defined as

$$z_{ni} = \frac{y_{ni}}{\sqrt{V(x_{ni})}} \tag{2.30}$$

Data that do not fit the model are not automatically rejected, rather they are investigated to identify their source of misfit and to what extent they corrupt measurement (Smith, 1996; Smith et al., 1998). Tests of fit vary according to the software used to run the analysis. In WINSTEPS®, for example, there are the chi-squared based INFIT (*inlier-sensitive fit*) and OUTFIT (*outlier-sensitive fit*). The former (Equation 2.31) is a weighted fit statistic that reports overfit for Guttman patterns and is more sensitive when items have their difficulty level close to the persons' ability level. The latter index (Equation 2.32) reports overfit for responses and it is more sensitive to differences for items with difficulty far from a person's ability level (Linacre, 2002b). An approach to summarising the fit of an item using the weighted mean square and non-weighted mean square is given as follows (Wright and Masters, 1982):

$$v_i = \sum_{n=1}^{N} y_{ni}^2 \Big/ \sum_{n=1}^{N} V_{ni} \tag{2.31}$$

and

$$u_i = \sum_{n=1}^{N} z_{ni}^2 / N \tag{2.32}$$

where $n = 1,2,...N$.

In RUMM2030®, there are two overall statistics of item-person interaction that summarise items-fit and persons-fit and are distributed as a *z*-score. Residual statistics of items fit assess the degree of divergence between the expected value and the observed value for each person-item as summed over all items for a given person (Figure 2.3).

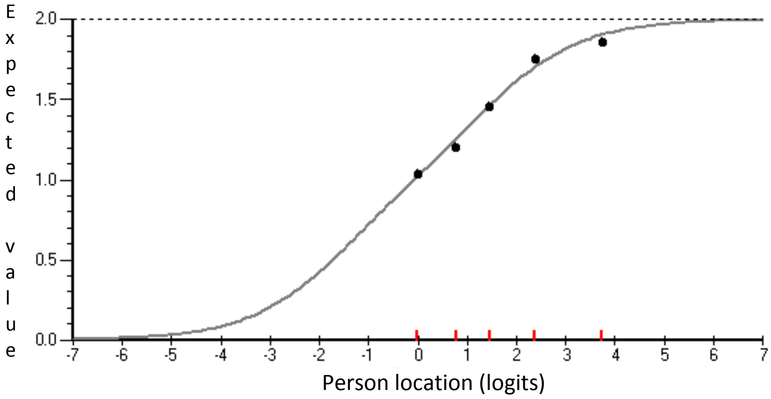


Figure 2.3 - Example of item characteristic curve (ICC) using RUMM2030. The dots represent the observed values of five person ability groups and the typical s-shaped curve represents the expected values.

To obtain the magnitude of the residuals, these are squared, giving a summary value for a person by summing over the items, such that

$$w_n^2 = \sum_{i=1}^{I} z_{ni}^2 \tag{2.33}$$

and a summary value for an item by summing over the persons, such that

$$w_i^2 = \sum_{i=1}^{N} z_{ni}^2 \tag{2.34}$$

Those values are then compared to their expected values where their expected values are their degrees of freedom. Because the same data are used to obtain the estimates of the parameters and to compute the residuals, the variance of residuals will be less than 1. Denoting it as degrees of freedom $\left(f_{ni} < 1\right)$ then the residual of a person *n* can be summarised for all the person-item residuals, such that

$$w_n^2 - E\left[w_n^2\right] = \sum_{i=1}^{I} z_{ni}^2 - \sum_{i=1}^{I} f_{ni} \tag{2.35}$$

and the residual of an item *i* can be summarised for all the person-item residuals as

$$w_i^2 - E\left[w_i^2\right] = \sum_{n=1}^{N} z_{ni}^2 - \sum_{i=1}^{N} f_{ni} \tag{2.36}$$

Standard form of Equations 2.35 and 2.36 gives respectively

$$Z_n = \frac{w_n^2 - E[w_n^2]}{\sqrt{V[w_n^2]}} \tag{2.37}$$

and

$$Z_i = \frac{w_i^2 - E[w_i^2]}{\sqrt{V[w_i^2]}} \tag{2.38}$$

2.4.7.6 Investigating item and person misfit

The score for item-person interaction, such as it was elaborated in the preceding section, indicates the degree of discrepancy of each person from the model. High positive fit-residual is observed by the flatter form of the dots than the expected curve indicating that the item is significantly under-discriminating (Figure 2.4). Negative fit-residual indicates that the item is significantly over-discriminating by observing that the dots form a steeper curve than the expected curve (Figure 2.5). Misfit has been indicated by the degree of divergence between the expected value and the true value for each person-item (Andrich, 1988b). In most of the cases when the residual value is within the theoretical interval of ±2.50, which represents approximately 99% of the confidence interval (CI), an item has been deemed as adequate fit to the model (Pallant and Tennant, 2007).

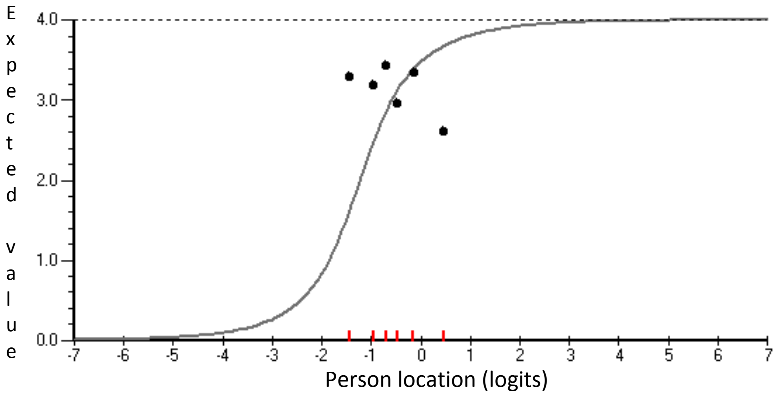


Figure 2.4 – Example of fit-residual with a high positive value.

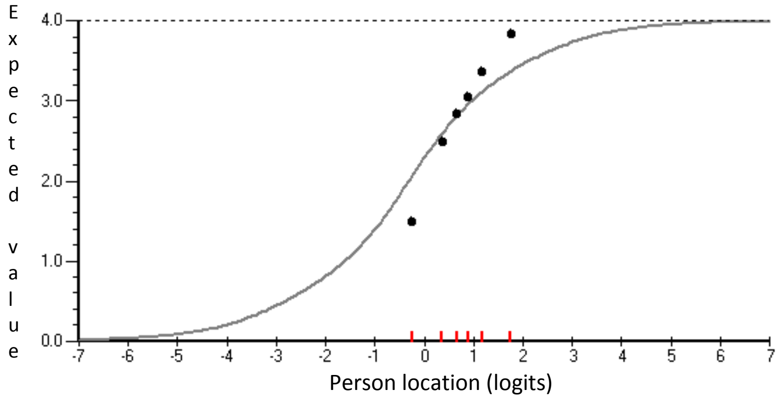


Figure 2.5 - Example of fit-residual with a high negative value.

Furthermore, item fit residuals and person fit residuals are compared against a reference value of their standard deviation (σ), suggested as ≤1.40 by Tennant and Conaghan (2007). However, this ought to be understood solely as an indication of a potential source of investigation for misfit, rather than a cut-off value.

2.4.7.7 Ordering of response categories

An important source of misfit is associated with the respondents' inconsistent use of the response categories when the scale has more than two response options. This involves the examination of the threshold patterns. Analysis of the transitions between categories can be interpreted as though there was an independent response for each of the thresholds. This allows identifying potential problems with the empirical order of categories (Smith and Plackner, 2009) (Figure 2.6).

If the response patterns are consistent, each response category has a point along the ability continuum where it is identified the most probable response (Figure 2.7) (Pallant and Tennant, 2007). As noted earlier in Section 2.4.7.4, the magnitude of the respective differences between adjacent threshold values has an effect on the slope of the ICC for an item (Figure 2.6).

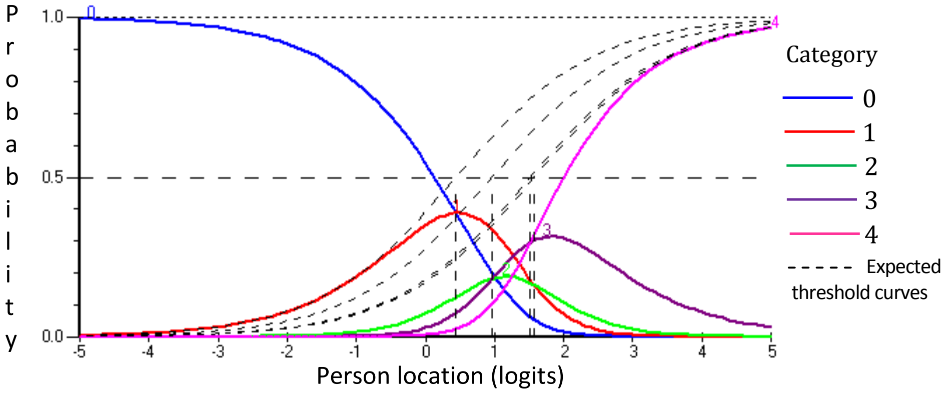


Figure 2.6 – Category characteristic curves for a five-category item with disordered thresholds.

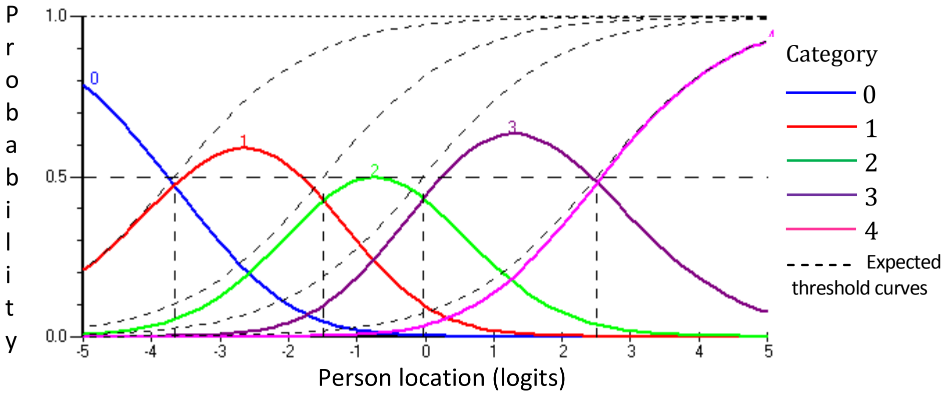


Figure 2.7 – Category characteristic curves for a five-category item with ordered thresholds.

The intersection of the line of probability of 0.50 with the ICC also establishes the location for the item. It is noteworthy that the degree of threshold separation for that item has an inverse relationship with the slope of an ICC.

If respondents do not use the category system such as it was designed by an analyst, i.e., the thresholds do not discriminate between adjacent categories, the analyst could consider combining the frequencies of those categories. This is not arbitrary, though (Andrich, 1978). If data fit the model for some number of categories, then summing their frequencies can promote misfit. An analyst could consider modifying the category system when the threshold estimates are reversed from their natural order or when the discrimination of adjacent thresholds might be close to zero (Figure 2.6).

2.4.7.8 Test for differential item functioning

Another source of misfit in the data with regard to the model is denoted differential item functioning (DIF) or item bias (Andrich and Hagquist, 2012; Broderson et al., 2007; Osterlind and Everson, 2009). The uniform DIF is indicated when a group demonstrates consistently greater ability to endorse an item than another group. This could, for example, be found in male and female groups, different age groups, cross-cultural investigations and cross-national studies. Non-uniform DIF is characterised when the ability differences to endorse an item are inconsistent amongst ability groups. Usually, residuals are analysed by a standard two-way analysis of variance (ANOVA) for detecting significant DIF on the resulting measures, allowing each element of the structure to be adjusted to any bias (Andrich and Hagquist, 2004). To detect DIF, the standardised residual expressed by Equation 2.37 can be identified by the target group. Equation 2.39 exemplifies it for sex group g and by ability group c for person n on item i, such that

$$z_{n_{cg}i} = \frac{x_{n_{cg}i} - E[x_{n_{cg}i}]}{\sqrt{V[x_{n_{cg}i}]}} \qquad (2.39)$$

Tests of statistical significance for DIF have been reported as the probabilities of single tests for each item i of a construct. The probability of incorrectly rejecting a true null hypothesis (h_i) (i.e., Type I error) for each item i is contrasted with a level of significance α adopted by an analyst. Another approach is to use Bonferroni adjustment. Bonferroni adjustment is a conceptualisation in which test of significance are, in effect, multiple independent tests of the same process (Bland and Altman, 1995). Bonferroni adjustment takes the probability for each comparison equal to α/n where n is the number of tests for h_i. However, the application of Bonferroni adjustment in Rasch analysis has been source of criticism (Perneger, 1998; Wolf, 2006). The concern is that the approach does not take into account the probability of Type II errors, i.e., the probability of incorrectly rejecting a true alternative hypothesis. This might allow the acceptance of items that present DIF. Nevertheless, the use of Bonferroni adjustment depends on the circumstances of analysis (Perneger, 1998). Significance at a α level can identify a source of misfit originated by DIF in item-by-item analysis. In this case the hypothesis is that there is no DIF for a determined item. On the other hand, if the whole group of items in a scale shall be considered (e.g., when testing the hypothesis that the data fit the model), then significance level may be adjusted accordingly.

2.4.7.9 Test for local dependence

Tests for local dependence identify anomalies that do not allow a scale to perform independently. Items that do not provide independent or relevant information can be a source of violations of local dependence, for example. Those anomalies can prevent the data set from fitting the model (see Section 2.4.4). One technique for assessing those violations is to identify patterns of high correlations amongst the standardised item residuals. The definition of the meaning of high correlation can vary according to the context of use of the scale. A typical approach is to examine items with an absolute correlation coefficient higher than or equal to 0.30 between their standardised residuals (Tennant and Conaghan, 2007).

Another class of local dependence in the RM refers to the items computed in the measure, constituting a unidimensional scale (see Section 2.4.4). Different tests of trait dependence have been presented in the literature (Andrich, 1985b; Smith and Miao, 1994; Smith, 1996; Tennant and Pallant, 2006; Kreiner and Christensen, 2004; Christensen et al., 2002)[18]. Wright (1996) and Linacre (1998) suggested carrying out PCA of the residuals. This implies that once the Rasch factor has been considered, there should not be any significant pattern in the residuals resultant of the relationship between items, excepting random associations. Smith (2002) proposed taking the factor loadings on the first residual through a PCA to identify the two most divergent subsets of items and then examine via paired *t*-test comparisons any difference in the estimates that have been generated. If the proportion of independent tests falls outside the boundaries of acceptable significance,

[18] Comparisons between tests of dimensionality of a scale based on CTT and RMT can be found in Waugh and Chapman (2005) and Ewing et al. (2009).

there might still be some degree of multidimensionality within the item set. The acceptable amount of deviating results is given by a binomial test (Horton and Tennant, 2010). This test verifies the statistical significance of deviations from a theoretically expected proportion. Typically, many tests are expected to fall outside of the *t*-range of ±1.96 for the confidence interval of 95% of the observations. Therefore, if this value is less than or equal to the 5% level, unidimensionality of the scale can be deemed acceptable when using the method (Pallant and Tennant, 2007).

2.4.7.10 Reliability indices

Reliability indices are obtained as the proportion of the variance (after computing measurement error) of a distribution of person estimates relative to the sum of this variance and the error variance in the estimates. Two indices are typically used in Rasch analysis. The person separation index (PSI) from Rasch theory and Cronbach's α from CTT. It is noteworthy that the reliability index represents essential information in CTT; however, in RMT the precision of the individual estimates is emphasised and therefore, the index is useful solely as an element of a comprehensive interpretation of a data set. Cronbach's α can directly be obtained in terms of the observed scores (Cronbach, 1951).

Letting $y_n = \sum_{i=1}^{I} x_{ni}$ and $y_n = \upsilon_n + \varepsilon_n$, such that $V[y] = V\left[\sum_{i=1}^{I} x_i\right]$. Thus,

$$\alpha = \frac{I}{I-1}\left[\frac{V\left[\sum_{i=1}^{I} x_i\right] - \sum_{i=1}^{I} V[x_i]}{V\left[\sum_{i=1}^{I} x_i\right]}\right] \tag{2.40}$$

where x_{ni} is the observed score of person *n* on item *i*, for $x_{ni} \in \{0,1,\ldots,m_i\}$ and $i \in \{1,\ldots,I\}$.

PSI $\left(r_{\beta\hat{\beta}}\right)$ is expressed in terms of the estimated locations of the persons (i.e., non-linear transformations of the raw scores). Thus, taking Equation 2.18 for the polytomous case of the RM with β indicating person location and letting $\hat{\beta}_n$ be the estimate of a person *n* and σ_{β}^{2} be the variance of the person locations and $\sigma_{\hat{e}}$ be the associate standard error of the estimate (Gulliksen, 1950), then

$$r_{\beta\hat{\beta}} = \frac{\sigma_{\beta}^{2}}{\sigma_{\beta}^{2} + \sigma_{e}^{2}} = \frac{\sigma_{\hat{\beta}}^{2} - \sigma_{\hat{e}}^{2}}{\sigma_{\hat{\beta}}^{2}} \tag{2.41}$$

Some differences between Cronbach's α and PSI can be identified. PSI can be computed in the presence of random missing data while it is necessary a complete data set for α . When a skewed distribution occurs with extreme raw scores, PSI will be more sensitive than α because there will be higher error variance close to the extreme while there is no effect in the construction of α .

2.5 OBJECTIVE MEASUREMENT IN AFFECTIVE ENGINEERING

The processes in affective engineering are not exclusively qualitative or quantitative. A measurement process always begins with a qualitative experience, which is reasoned out with quantitative methods (Thurstone, 1928). Although qualitative comparisons are necessarily part of the process to elicit the users' affective responses, they are not sufficient to provide a more fine-grained interpretation of users' interaction with physical elements of products.

In this chapter it is characterized through measurement theory that not every property can be numerically represented. The preceding literature review has pointed out that one of the most important differences between numbers assigned to affective responses and their quantitative property is that the order of the numerals is established by convention while the order of the system in respect of the quantity is determined by empirical operations, which is a property of the mathematical model used to the construction of a measurement instrument.

The idea of good measurement of the affective value of products' characteristics is associated with quantitative comparisons. In this chapter it has been shown that for comparisons in a relevant frame of reference with a useful range of generality, the mathematical model ought to contain mechanisms that

i. control the variance inherently connected to uncertainties of the users' experiences,
ii. preserve the order in the structure of observations,
iii. obtain independent estimates of any pair of persons and any pair of items,
iv. obtain a precise estimate of error variance, separating random and systematic errors and
v. construct a linear scale, preserving additivity on the continuum.

The principles and procedures of RMT presented throughout Section 2.4 have widely been discussed and assessed in different domains of knowledge (e.g., education, health and social sciences), endorsing the RM's properties for objective comparisons. The matter in affective engineering, therefore, lies on the decision of what sort of evidence is necessary to validate the affective interaction of users' with physical components of products. If clarification and refinement of the outcomes are necessary in a relevant context and empirical conditions in the domain, the RM can provide mechanisms to establish linear measures scientifically modelled.

2.6 REMARKS

2.6.1 Adapting the Taxonomy of Rasch Measurement Theory in the Domain of Product Design

Throughout Section 2.4 particular terms and notations well-established in IRT and RMT were used to provide the concepts underlying the Rasch models. Applications in the domains of education and rehabilitation of those theories employ the term *person ability* to the respondent location and *item difficulty* to denote item location (Wright and Panchapakesan, 1969). This represents the ability of a person to respond correctly an item at a certain degree of difficulty on a scale (DeVellis, 2006).

However, these terms are not usual in affective (kansei) engineering and might lead to misinterpretations. Thus, item difficulty is understood as the proportion of respondents who endorse an item. That is, an item that is endorsed by relatively many respondents is taken as easier than one endorsed by few respondents, which is considered more difficult. The term person ability means the readiness with which a person endorses an item and is interpreted in this thesis as indicative of their affective response to the product. In other words, person ability in terms of the application in product design means the degree to which a person is inclined to respond positively to a determined adjective or statement associated with the relevant affective attribute of a design element or of a product feature.

Furthermore, the term *item*, commonly used in test theory, indicates the independent variables in a measurement structure (see Section 1.3.1). In this thesis, the term has been adapted to the domain of AE, associating it with the adjectives or statements embedded in paper-based or in computer-based, self-report questionnaires. Other adaptations of equations and various notations typically used in Rasch taxonomy will be presented throughout the thesis.

2.6.2 Estimation of Parameters

Even though each method of parameters estimation presents benefits and drawbacks (see Section 2.4.6), they produce statistically equivalent results (Wright, 1988). Nevertheless, particular care ought to be taken when comparing results on the same measurement continuum from different computer programs which employ different estimation methods (Linacre, 1999). For this reason, to avoid any shortcomings originating from differences in estimation methods, in this thesis, just the PML and WLE methods embedded in the software package RUMM2030 will be used.

The software package RUMM2030 uses four principal components when estimating the item parameter in presence of a structure with five categories (see Section 2.4.6).

However, Andrich and Luo (2003) have suggested investigating the trade-off between analysis of structures with fewer parameters estimated than the maximum possible and the stability of the estimates when in the presence of more than five categories holding low frequencies.

2.6.3 Concluding Comments

RMT embodies a meticulous rationale to identify anomalies in the data through statistical misfit with regard to the model. The RM properties are sufficient to determine whether a measurement structure is additive. However, this ought to be seen contextually. The model has been applied in education where in many cases the interest lies on the performance of students, in which can intuitively be observed a quantitative structure. For example, a student who holds higher ability will likely respond correctly to most of the difficult questions in a test. Thus, different levels of ability could be measured through the probability of responding to an item correctly. In tests of reading and writing, mistakes could be counted, timing could be taken into account, and so forth. In the field of rehabilitation, patients who have efficiently recovered from a stroke after a treatment will demonstrate more ability to accomplish some tasks, such as climbing stairs, than those who have not. Thus, different levels of recovery can also be measured.

However, data from affective responses to product features could present many symptoms of anomaly to fit the model. In AE it is very likely that persons do not give a rating of their entire interaction with products objectively. There will be items preliminary established in a structure that clearly do not fit the model. Redundancy, misrepresentation, misinterpretation, bias and ambiguity are some sources of misfit in items. There will also be cases in which the whole data set presents poor fit because, under the RMT perspective, this is indicative of a structure that is not quantitative. In this case, inferences drawn from the statistical results cannot be generalised beyond the sample studied and the scores cannot be considered as an element of a measurement structure.

Nevertheless, there are different degrees of misfit in terms of a measurement structure originated from that physical interaction with products. These different degrees of misfit ought carefully to be analysed for items and for persons. For example, if one item is removed from a set because some degree of misfit is identified, the fit statistics will change for all other items. Further, several statistics presented in Section 3 are affected by sample size and by the characteristics of the sample. Therefore, theoretical cut-offs, such as those mentioned in Section 3.6, are useful benchmarks although they ought not to be taken as a basis to make a decision. In general, such as in physics, misfit ought to be considered an anomaly in the data and substantively investigated.

CHAPTER 3

Applying the Rasch model for affective responses to products

In Chapter 3 an empirical approach is reported. The research aimed to examine how well the data from affective responses would fit the expectations of the Rasch model to create a scale of specialness for four pieces of wrapped confectionery. A pool of items partially obtained from a previous study in the UK division of an international confectionery company was used in the investigation and the responses were analysed using the model. The research has shown that the model can enable the development of a frame of reference for a measurement structure. This was possible because Rasch analysis validated the scale through calibration, assessed invariance of items (independent variables), tested items for potential bias, validated the category scoring system and fitted items to the model. Nevertheless, the results indicated that participants rated each piece of confectionery in different patterns, indicating that they are on distinct continuums. Therefore, to compare stimuli on the same continuum will require adapting derivations of the Rasch model in the domain[19].

3.1 APPLYING THE RASCH MODEL

The research reported in this chapter explored the RM through an experiment involving 306 participants of different sex and age in order to verify whether the data from participants' affective responses to some wrapped chocolates would fit the RM. The preliminary pool of statements was obtained from a previous study using different confectionery. In addition, this research used Likert items (Likert, 1932) to elicit affective responses, rather than contrasting adjective pairs because the RM assumes unidimensional data, whereas the SD approach is inherently multidimensional (see Section 2.2). It might be difficult to propose enough adjective pairs to define a unidimensional structure along a desired construct. Furthermore, it is not possible to carry out a factor analysis of responses to adjective pairs to identify a desired construct because the constructs emerge empirically from the multivariate analysis, instead of being prescribed. Therefore, if one wishes to measure a specific construct, developing Likert items offers a more tractable approach than developing adjective pairs.

[19] Publications based on this chapter can be found in Camargo and Henson (2010, 2011, 2012b).

3.2 HYPOTHESIS OF THE EMPIRICAL APPROACH

The study tested the hypothesis that the observed data from affective responses to product stimuli fit the expectations of the Rasch measurement model (see Section 1.6.1). The empirical approach aimed to identify whether the RM would produce appropriate interval measures in affective-based experiments for comparing different characteristics of products.

3.3 METHOD

3.3.1 Participants, stimuli and preliminary pool of items

Data were collected at the School of Mechanical Engineering of the University of Leeds, in the affective engineering laboratory of the Institute of Engineering Systems and Design in January 2010. Ethical approval for the empirical study was obtained from the University of Leeds Research Ethics Committee (ethics reference number MEEC 09-005).

Three hundred and six participants took part in the study[20], 44.12% females and 55.88% males, who ranged in age from 18 to 25 (52.40%), from 26 to 35 (33.87%), from 36 to 45 (10.22%), and over 45 (3.51%). This size of sample can anticipate a proper level of confidence without primarily taking into account the distribution across the response options of each item (Linacre, 1994b). Participants received £5.00 as a compensation for taking part in the study.

Four pieces of wrapped confectionery were presented to each participant. The stimuli were chosen to provide a variety of distinguishable properties related to specialness according to the participants' point of view. Thus, four well-known brands of wrapped confectionery in the United Kingdom were chosen as follows: Caramel® and Milky Way® from a Mars Celebrations® assortment, Ferrero Rocher®, and Lindor® (Figure 3.1).

The data were collected through a self-report questionnaire that contained 24 statements based on the understanding of the product context to assess the required affective attribute for each stimulus (Table 3.1). Some of the statements were determined in a previous study in the UK unit of an international confectionery company (Henson, 2009) in which the target demographic was British women aged between 25 and 45 who like sharing chocolate informally with adult friends. Those statements were established through qualitative consumer research, expert panel and the company's requirements to

[20] Seven additional persons participated in a previous pilot study to adjust the format of questionnaires and to time the session.

determine whether the confectionery was considered special and good for sharing. In this study, however, solely statements obtained from the company's study associated with the context of the stimuli and with the attribute *specialness* were used. Furthermore, other statements were introduced to identify whether the adjectives might be related to specialness of the confectionery.

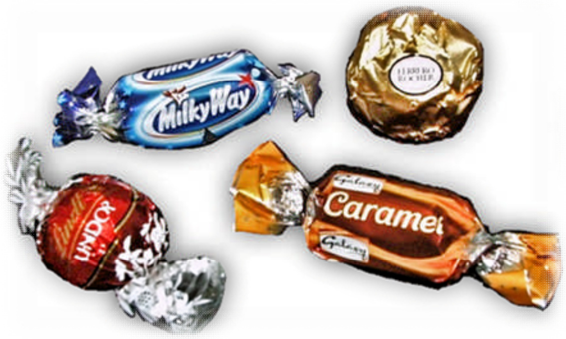


Figure 3.3.3.1.1 - Wrapped confectionery used as stimuli in the experiment.

Table 3.1 - Preliminary pool of items

Code	Statement
I01	With this chocolate, you feel as though you are getting more than just chocolate.
I02	Opening a box of these chocolates would really set the mood for a night in front the telly.
I03	This chocolate is for grown-ups.
I04	This chocolate looks expensive.
I05	This chocolate is a bit flash.
I06	This chocolate is special.
I07	This chocolate is mass-produced.
I08	This chocolate would show that someone took the time to choose just the right chocolate for the occasion.
I09	A box of these chocolates would make a lovely romantic gift.
I10	This is a premium chocolate.
I11	The chocolate in this wrapper is likely to exceed people's expectations.
I12	I would keep chocolates like this one for myself.
I13	This chocolate is like a little present for me.
I14	This chocolate would be good to enjoy with my loved-one on a quit night in.
I15	This chocolate would be nice during a break from housework.
I16	Eating one of these chocolates I would feel a little bit naughty.
I17	This chocolate is stylish.
I18	This chocolate is cheap.
I19	A box of these chocolates would be an appropriate "thank you" gift.
I20	A box of these chocolates would make a thoughtful gift.
I21	You could give someone a box of these if you wanted to say "sorry."
I22	This chocolate does not need to shout about how good it is.
I23	This chocolate would be nice at the end of a dinner party.
I24	This chocolate is for children.

The participants rated their endorsement on a five-point Likert-style scale (i.e., strongly disagree, disagree, neutral, agree, and strongly agree). Written information about the activity was provided in advance on the experiment website. A verbatim protocol was used for giving instructions before the test. The order in which participants were required to consider the pieces of confectionery was determined using a counterbalanced design. The order of the statements on the questionnaires was randomised. The data were independently double-entered, compared, and any transcription error was corrected.

3.3.2 Calibration of Items

Firstly, each piece of confectionery was analysed individually. A second step was to examine whether one scale could be employed as a basis of measurement for all stimuli. The Rasch analysis was carried out with the software package RUMM2030®, standard edition (2010).

3.3.2.1 Derivation of the Rasch model used in the analyses

The likelihood-ratio test verified which derivation of the RM should be applied because the scale had more than two options of response. The rating scale (Andrich, 1978) would be used if the outcome of the test was not significant. However, the test presented significance (i.e., p>0.05), indicating different intervals between categories and therefore, the partial credit model was adopted (Masters, 1982).

3.3.2.2 Verification of the score system, response pattern and item-person interaction

The category structure concerning disordered thresholds, which is an important source of misfit related to the respondents' inconsistent use of the response options (Tennant and Conaghan, 2007), was examined through the threshold patterns (see Section 2.4.7.7). The response pattern was examined through the residuals of distinctive person responses. Residuals with absolute values greater than 2.50 were sources of investigation (see Section 2.4.7.6). The item-person interaction indicated the degree of discrepancy of each person from the model. Residuals between the value of ±2.50 were assumed as random errors and residuals greater than the absolute value of 2.50 were carefully investigated (Andrich, 1988a).

3.3.2.3 Tests of fit

Two tests of fit were performed in the analysis. The first test of fit was to examine the degree to which the Guttman pattern had been achieved (see Section 2.4.7.5). Residual statistics of items fit assessed the degree of divergence between the expected value and the actual value for each person-item as summed over all items for a given person. This study used as an indicator of fit a σ between 0.70 and 1.40 based on empirical studies from others' research (Pallant and Tennant, 2007; Wright et al., 1994). The second test of fit was a formal test of invariance across the trait. A significant chi-square probability (i.e., p <0.05) indicated variance in the scale (Tennant et al., 2004).

The individual chi-square test-of-fit compared the difference between observed responses and those expected by the model over groups representing different ability levels through the trait to be measured (Elhan et al., 2008). The values lower than a Bonferroni-adjusted value (Bland and Altman, 1995) were indicators of data fit (Tennant and Conaghan, 2007). The Bonferroni-adjusted value was calculated as the ratio between the level of significance (α = 0.05) and the number of items (Bland and Altman, 1995).

3.3.2.4 Tests for DIF

DIF was detected through an ANOVA conducted for each item comparing scores across different class intervals and across each level of the person factor (see Section 2.4.7.8). This study focused on whether female group demonstrated consistently greater ability to endorse an item than the male group. Similarly, the person factor age group was also tested for DIF. Statistically significant uniform DIF was identified when p was lower than the Bonferroni-adjusted value.

3.3.2.5 Assumptions of response independence and unidimensionality

Response dependency between items was identified by observing high correlations in the residuals of the items. High correlations were assumed as an absolute value greater or equal to 0.30 in this study (Smith et al., 2003; Tennant and Conaghan, 2007).

Unidimensionality was tested through the method proposed by Smith (2002) through a PCA of the residuals and binomial test (see Section 2.4.7.9). Thus, if the value of the binomial test was less than or equal to 0.050 then the unidimensionality of the scale was deemed acceptable.

3.3.2.6 Power of fit and targeting

The power of test-of-fit was the indicator of internal reliability and was represented by the PSI (see Section 2.4.7.10). The level of adequacy was assumed to be greater or equal to 0.70. This value allows two groups of respondents (i.e., class intervals) to be differentiated (Fisher, 1992).

3.4 RESULTS

3.4.1 Analysis of the Preliminary Pool of Items and Score System

The fit of data to the model was examined from the preliminary pool of 24 items and sample of 306 persons. The likelihood-ratio test indicated that the outcomes for all of the stimuli were significant (p <0.05). Consequently, the partial credit model was used in the analysis.

The standard deviations of person-fit residuals and the standard deviations of item-fit residuals were higher than the expected value of $\sigma \leq 1.40$ (Table 3.2, Columns σ). Such values combined with p <0.05 for all of the stimuli, which indicates lack of the invariance across the trait, pointed to misfit to the model (Table 3.2, Column p). Statistics of the individual items interaction identified residuals of items with absolute value greater than 2.50. This was also verified by examining the ICC (see Section 2.4.7.5).

Table 3.2 – Fit statistics for the preliminary scales

Stimulus	Persons-fit residual		Items-fit residual		Item-trait interaction		
	Mean	σ	Mean	σ	df	χ^2	p
FerreroRocher®	-0.21	1.52	0.23	2.69	120	781.17	<0.05
Milky Way®	-0.25	1.75	0.17	3.25	120	993.05	<0.05
Caramel®	-0.32	1.85	0.42	3.41	120	928.13	<0.05
Lindor®	-0.22	2.06	0.41	4.92	120	1684.17	<0.05

3.4.2 Sources of Misfit

The transitions between categories indicated whether each response category had a point along the ability continuum indicating the most probable response. The double asterisks in the thresholds map indicate the items that held disordered thresholds (Figure 3.2 for Ferrero Rocher®, Figure 3.3 for Milky Way®, Figure 3.4 for Caramel® and Figure 3.5 for Lindor®). Those disordered thresholds are misfits associated with the respondents' inconsistent use of the response categories when the scale has more than two response options.

Analysis for detecting DIF through ANOVA indicated differences of responses between sexes after Bonferroni adjustment (p <0.002). The analysis of Ferrero Rocher®, Milky Way®, Caramel® and Lindor® indicated DIF for sex and age for the items presented in Table 3.3.

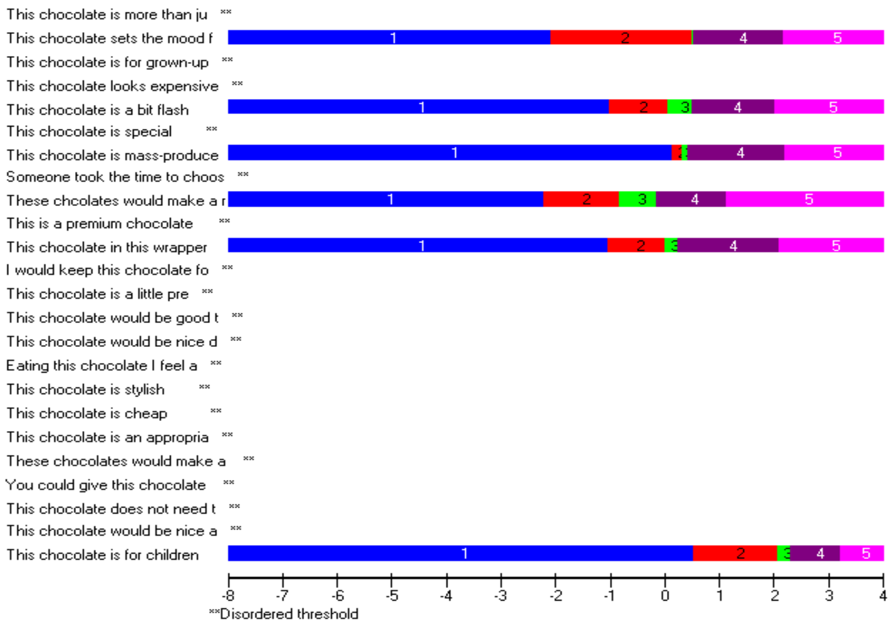


Figure 3.2 - Thresholds map for the preliminary analysis of the stimulus Ferrero Rocher®.

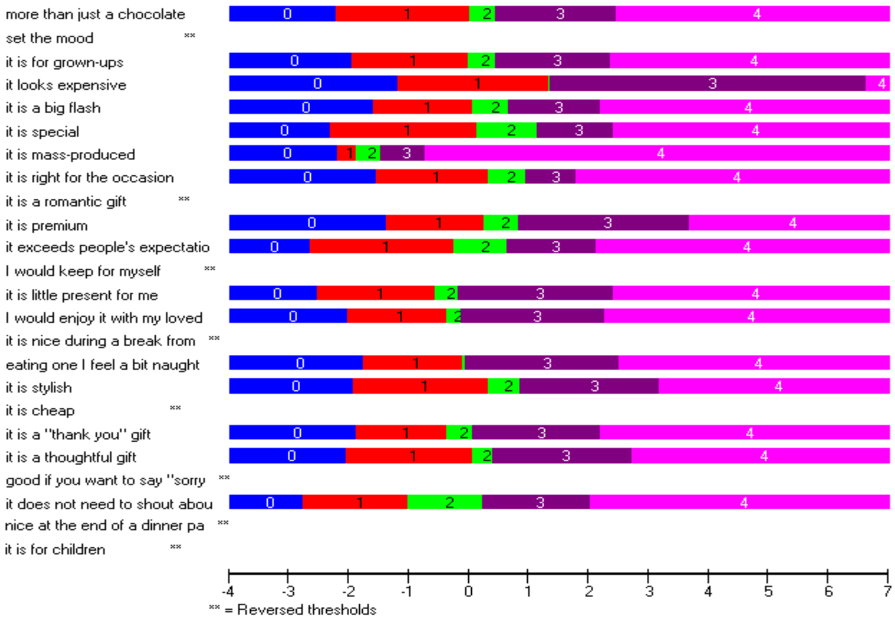


Figure 3.3 - Thresholds map for the preliminary analysis of the stimulus Milky Way®.

Examination of the pattern of residuals for every stimulus indicated that the preliminary set of items presented high correlations (i.e., residuals with correlation greater than or equal to ±0.30) (Table 3.4). Therefore, the preliminary scale for Ferrero Rocher®, Milky Way®, Caramel® and Lindor® violated the Rasch assumption of response independency.

The assumption of unidimensionality was tested through principal component analysis concerning fit-residuals for each item and for each person. An independent *t*-test was performed to test the assumption of unidimensionality. The outcome indicated the

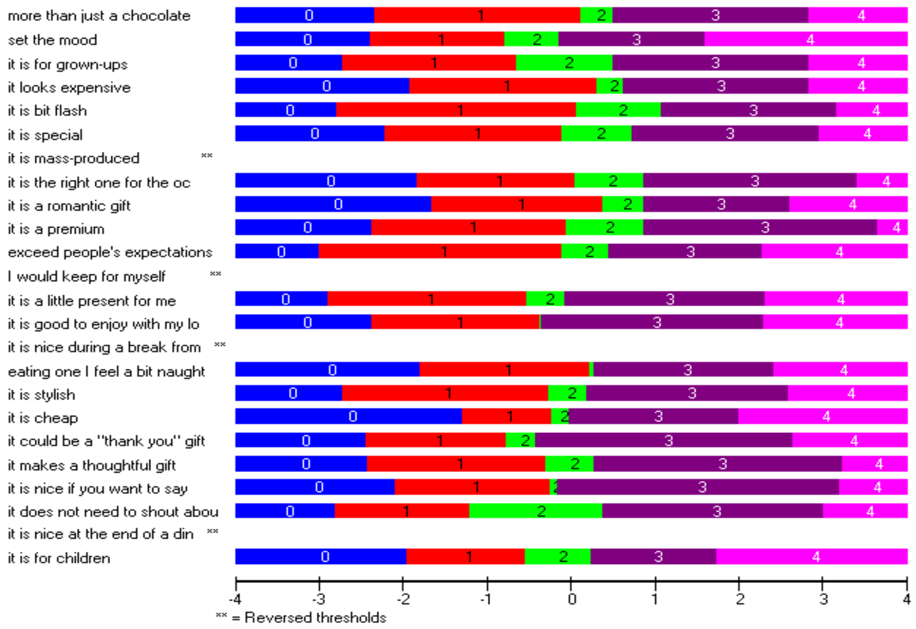


Figure 3.4 - Thresholds map for the preliminary analysis of the stimulus Caramel®.

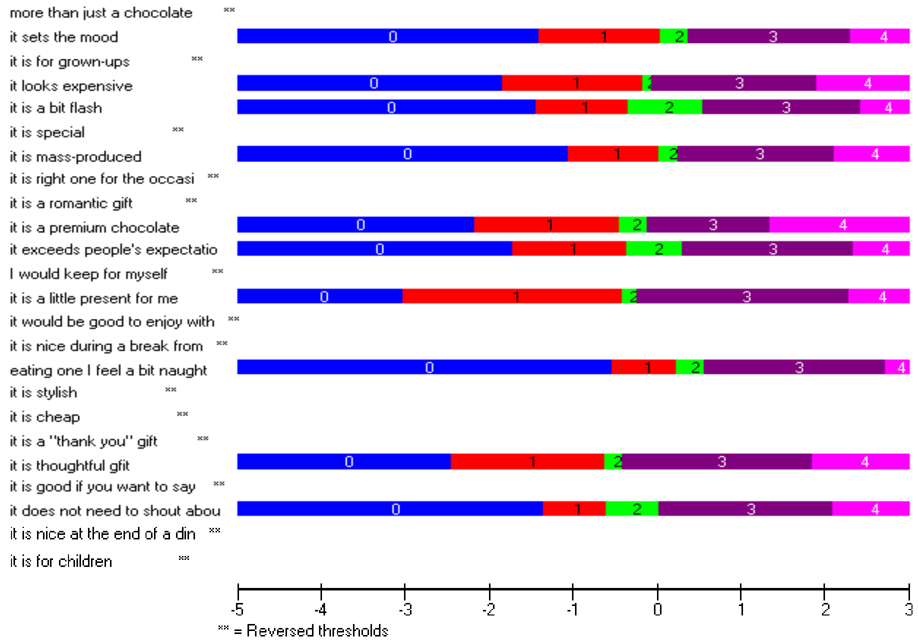


Figure 3.5 - Thresholds map for the preliminary analysis of the stimulus Lindor®.

Table 3.3 – Item that indicated presence of DIF for the confectioneries.

DIF	Items			
	Ferrero Rocher®	Milky Way®	Caramel®	Lindor®
Sex	I13	I02, I11	-	I05, I18, I24
Age	I02, I04, I15	I02, I05, I07, I08,I09, I10, I18, I24	I08	I04, I05, I09, I18, I24

Table 3.4 – Results from a correlation matrix of residuals for each stimulus

Stimulus	Item	Positive correlation (>0.30)	Negative correlation (>-0.30)
Ferrero Rocher®	I04	I06, I10 and I17	I18 and I24
	I07	I18	I08, I09 and I20
	I18	I07 and I17	I04, I06, I09 and I10
	I24	I18	I04 and I17
Milky Way®	I07	I18, and I24	I09, I10, and I23
	I18	I07	I04, I06, I09 and I17
	I24	I07 and I18	I08, I09, I19, and I23
Caramel®	I07	I18, and I24	I06, I08, I09, and I10
	I18	I07 and I24	I06, I08, I09, I10 and I20
	I24	I07 and I18	I09 and I10
Lindor®	I04	I06, I10 and I17	I15, I18, and I24
	I06	I09, I10, and I20	I07, I18, and I24
	I07	I18 and I24	I08, I09, and I20
	I18	I07	I01, I04, I06, I08, I09, I10, I14, I17, I19, and I20
	I24	I07, I15, and I18	I01, I04, I06, I09, I10, I14, I17, I19, and I20

proportion of *t*-tests that fell out of the limit of ±1.96 was 0.262 for Ferrero Rocher®, 0.283 for Milky Way®, 0.273 for Caramel® and 0.385 for Lindor®. These values were greater than 0.050 for an acceptable amount of deviating results, indicating multidimensionality across the trait.

3.4.3 Calibration of Items

Initially, the score system of each stimulus was recoded by collapsing responses categories (Table 3.5). In addition, reversed order was applied for Items I07, I16, I18, and I24. After collapsing each transition between categories, a point along the ability continuum was identified, indicating the position of the most probable response (see Section 2.4.7.7).

Analysis of individual item-fit indicated significant *p* after Bonferroni adjustment. Analysis of person-item correlation combined with analysis of individual item-fit identified 10 items that presented misfit to the model for Ferrero Rocher® (Table 3.6, indicated in italic). Thus, they were removed from the preliminary scale. Similarly, the procedure removed six items from the initial set for Milky Way®, eight items for Caramel®, and 11 items for Lindor®. The set of remaining items for every stimulus met the Rasch assumption of response independency (Table 3.6, indicated in bold).

Table 3.5 – Re-codification of the score system for each stimulus

Stimulus	Items collapsed to four categories	Items collapsed to three categories	Reversed order
Ferrero Rocher®	I01, I03, I06, I07, I08, I10, I12, I15, I16, I20, I21, and I22	I04, I14, I17, I18, I19, I23, and I24	I07, I16, I18, and I24
Milky Way®	I02, I04, I07, I09, I12, I15, I16, I18, I21, I23, and I24		I07, I16, I18, and I24
Caramel®	I07, I12, I15, I16, and I23		I07, I16, I18, and I24
Lindor®	I03, I06, I08, I09, I12, I14, I15, I16, I19, I21, I23 and I24	I03 and I18	I07, I16, I18, and I24

Table 3.6 – Remaining items and items removed after calibration.

Stimulus	Items																							
Ferrero Rocher®	**1**	**2**	*3*	*4*	*5*	*6*	*7*	**8**	*9*	**10**	**11**	**12**	**13**	**14**	**15**	*16*	*17*	*18*	**19**	**20**	**21**	**22**	**23**	*24*
Milky Way®	**1**	**2**	*3*	**4**	*5*	**6**	*7*	**8**	**9**	**10**	**11**	**12**	**13**	**14**	*15*	*16*	**17**	**18**	**19**	**20**	**21**	**22**	**23**	*24*
Caramel®	**1**	*2*	*3*	**4**	*5*	**6**	*7*	**8**	*9*	**10**	**11**	**12**	**13**	**14**	**15**	*16*	**17**	**18**	**19**	**20**	*21*	**22**	**23**	*24*
Lindor®	**1**	*2*	*3*	**4**	*5*	*6*	*7*	**8**	*9*	**10**	**11**	**12**	**13**	**14**	*15*	*16*	**17**	**18**	**19**	**20**	*21*	**22**	**23**	*24*

Note: Items that presented misfit are in italic and remaining items are in bold.

Furthermore, after re-scoring the system and removing items, analyses of the correlation between residuals did not identify associations greater than or equal to the absolute value of 0.30 for the stimuli. This was taken as evidence of response independency between items.

Analysis of the individual person-fit indicated that the pattern of responses from 31 participants for Ferrero Rocher®, 28 participants for Milky Way®, 33 participants for Caramel®, and 34 participants for Lindor® held high residuals. Those responses were identified and, in some cases, removed from each analysis.

Although DIF was tested with Bonferroni adjustment for sex and age after re-scoring and removing items, no significant item bias was identified except for Milky Way® that presented significant DIF for age in Item I02. Thus, this item was split in four sub-items by age group (see Section 2.4.7.8).

The fit of data to the model was examined from the set of the remaining items and the sample distributed into five class intervals for each and every stimulus. Rasch analysis identified a non-significant item-trait interaction, indicating that the data fitted the model (Table 3.7, Column p). Additionally, the residual of person-fit and residual of item-fit presented σ <1.40 (Table 3.7, Columns σ). The PSI of 0.84 for Ferrero Rocher®, 0.91 for Milky Way®, 0.91 for Caramel®, and 0.91 for Lindor® pointed to enough power to differentiate amongst respondents.

The results presented an expected proportion of *t*-tests of 0.040 for Ferrero Rocher®, 0.047 for Milky Way®, 0.038 for Caramel®, and 0.048 for Lindor®, suggesting unidimensionality across the measurement structure.

Table 3.7 - Summary of fit statistics for the calibrated scales.

Stimulus	Persons-fit residual		Items-fit residual		Item-trait interaction		
	Mean	σ	Mean	σ	df	χ^2	p
Ferrero Rocher®	-0.35	1.30	-0.22	1.21	70	86.24	0.06
Milky Way®	-0.23	1.31	-0.21	1.19	90	110.74	0.07
Caramel®	-0.23	1.32	0.45	1.23	64	60.90	0.59
Lindor®	-0.24	1.24	0.04	1.36	52	52.06	0.47

3.4.4 Co-calibration of Items for All Stimuli

The preceding calibration established a particular set of items for each stimulus. The remaining items were independent, the data fitted the model and the scales were deemed unidimensional. However, the stimuli held distinct item set characteristics. Thus, a further calibration established unified scales for all of the stimuli based on the individual scales.

Taking into account that the thresholds had been ordered, the first step was to remove or to add items from the individual scales to convey a solution for a common scale without considering persons misfit, whose analysis took place as a second step. Finally, an analysis of fit was performed. Table 3.8 presents a summary of the fit statistics of the co-calibrated scales. Table 3.9 indicates the 12 remaining items after co-calibration.

The PSI of 0.82 for Ferrero Rocher®, 0.87 for Milky Way®, 0.89 for Caramel®, and 0.90 for Lindor® indicate that the co-calibrated scales hold enough power for differentiating amongst respondents. Although the PSI of the co-calibrated scales was

Table 3.8 - Summary of fit statistics for the co-calibrated scales.

Stimulus	Persons-fit residual		Items-fit residual		Item-trait interaction		
	Mean	σ	Mean	σ	df	χ^2	p
Ferrero Rocher®	-0.32	1.22	-0.12	1.02	48	42.01	0.72
Milky Way®	-0.21	1.18	0.45	0.92	48	57.82	0.16
Caramel®	-0.24	1.26	0.50	1.19	48	38.37	0.84
Lindor®	-0.27	1.22	-0.07	1.19	48	42.64	0.69

Table 3.9 - Remaining items and items removed after co-calibration.

Stimulus	Items
Ferrero Rocher® Milky Way® Caramel® Lindor®	**1** *2* *3* *4* *5* *6* *7* **8** *9* **10** **11** **12** **13** **14** *15* *16* **17** *18* **19** **20** *21* **22** **23** *24*

Note: Items removed from the scale are in italic and remaining items are in bold.

reduced for every stimulus in comparison with the calibrated scales, it still remained at a level enough for the reliability of the fit statistics (Table 3.10).

Independent *t*-tests were performed for each stimulus to determine whether the assumption of unidimensionality had been met. The outcomes indicated the expected proportion of *t*-tests were 0.046 for Ferrero Rocher®, 0.048 for Milky Way®, 0.037 for Caramel® and 0.040 for Lindor®, which point to an acceptable amount of deviating results, indicating unidimensionality across the trait. Therefore, the co-calibrated scales met the Rasch assumption of unidimensionality. Although DIF was tested with Bonferroni adjustment for sex and age after removing and adding items, statistically significant item bias was not identified in the measurement structure.

Table 3.10 - Summary of the sample size and PSI for the analyses.

Measurement	Ferrero Rocher®		Milky Way®		Caramel®		Lindor®	
	n	PSI	*n*	PSI	*n*	PSI	*n*	PSI
Preliminary	306	0.83	306	0.86	306	0.86	306	0.87
Calibrated	275	0.84	278	0.91	273	0.91	272	0.91
Co-calibrated	278	0.82	270	0.87	271	0.89	272	0.90

The graphical representation of item-person location indicates the ability level of respondents and item difficulty on the linear scale in logit (Figures 3.6, 3.7, 3.8 and 3.9). The groups of respondents and ability levels are on the upper part. The item locations and their distribution are on the lower part of the graph.

The results from the co-calibrated scales indicated that the average of mean location for persons of 1.59 logit for the Stimulus Ferrero Rocher® (Figure 3.6) and 1.71 logit for Lindor® (Figure 3.7) are greater than zero and thus, are understood as respondents' *endorsement to specialness*. The average mean location for persons of zero logit is understood as respondents' *neutrality to specialness* of the Stimulus Caramel® (Figure 3.8). The average of mean location for persons of -1.01 is understood as respondents' *non-endorsement to specialness* of the Stimulus Milky Way® (Figure 3.9).

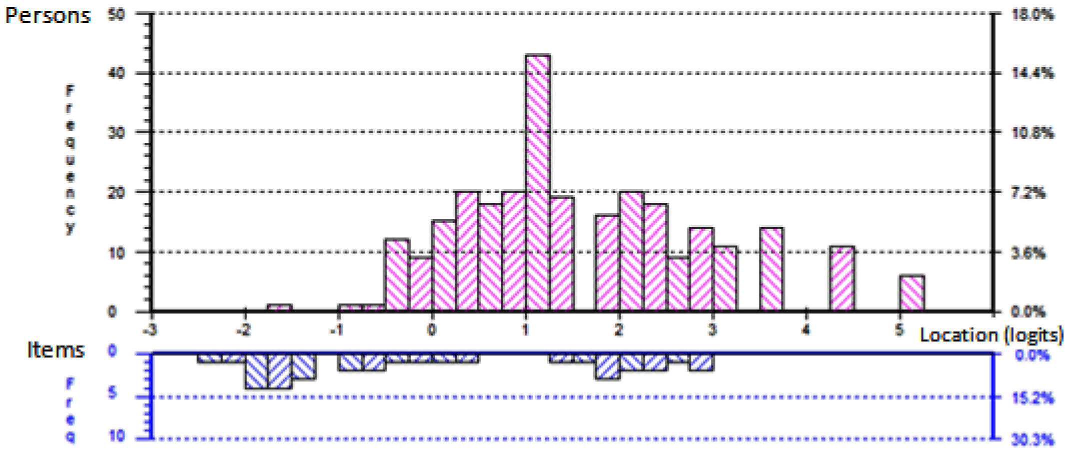


Figure 3.6 – Items-persons thresholds distribution for Ferrero Rocher®.

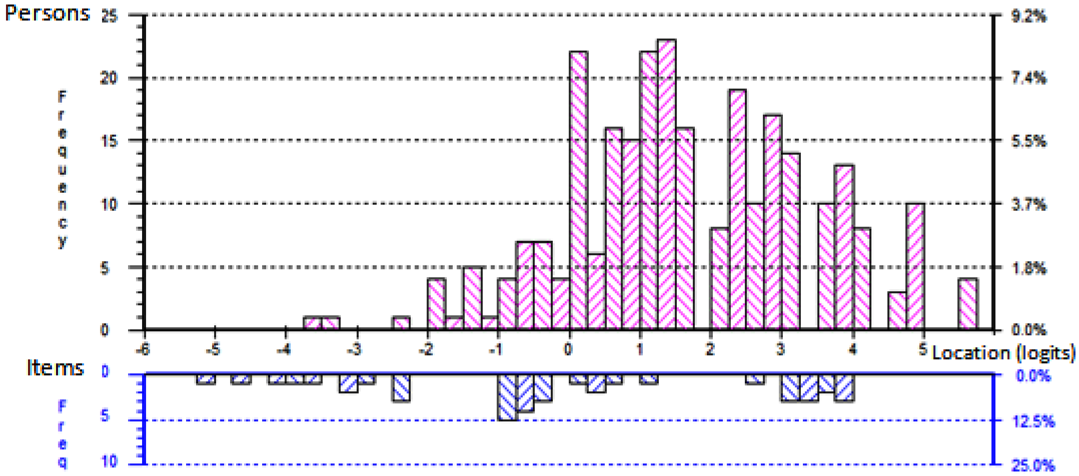


Figure 3.7 - Items-persons thresholds distribution for Lindor®.

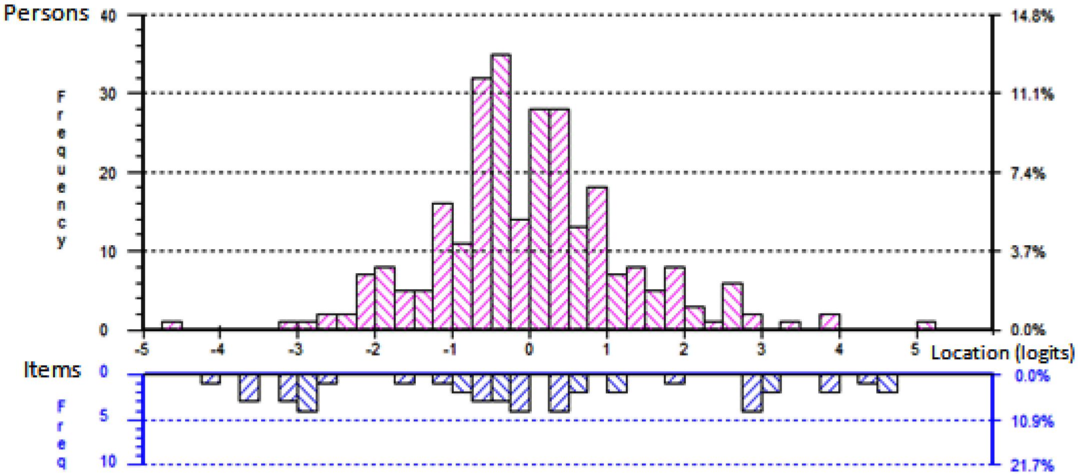


Figure 3.8 - Items-persons thresholds distribution for Caramel®.

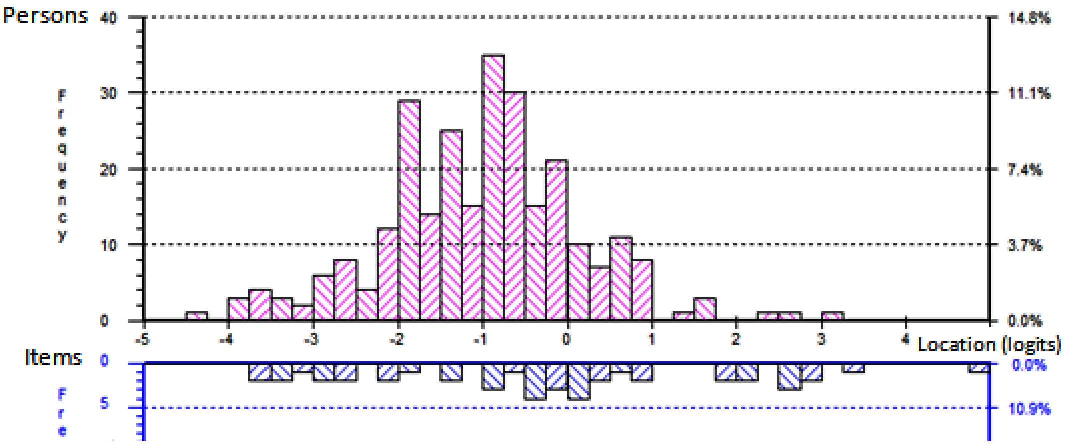


Figure 3.9 - Items-persons thresholds distribution for Milky Way®.

3.5 REMARKS

3.5.1 Limitations of the Empirical Approach

During calibration and co-calibration of the scales around 10% of the respondents were removed to fit the sample to the model. This was not unexpected because the sample was not from a narrowly defined demographic. Analysing such misfits might be a source of additional information. This has to do perhaps with the use context, products' brand or cultural factors. However, such issues demand further studies.

Twelve statements were discarded from the preliminary pool while co-calibrating scales. Test for identifying misfit to the model indicated different reasons to remove items (see Section 3.5.2). Nevertheless, those items could belong to a parallel scale of confectionery specialness or they could be elements of sub-scales for other different factors of specialness.

3.5.2 Outcomes from the Empirical Approach

Rasch analysis disclosed difficulties with the statements for characterising the consumers' experience that could mislead conclusions when using classical measurement approaches. One of the difficulties was evidenced by the misfits to the model when respondents misinterpret an item resulting in an unexpected response. For example, take the Item I24 *this chocolate is for children*. A respondent might have been unfavourable to this statement if one understood that *the chocolate is expensive, so it is more appropriate for grown-ups*. However, a respondent might have endorsed it if one believed that *the chocolate is wholesome for children*. In this case the statement held different meanings for those participants. Another difficulty was the redundancy of statements. Statements with very similar semantic meanings inflate the results in classical statistics. For example, respondents could have understood that the adjectives *special* and *premium* held the same meaning and therefore, those statements presented response dependency. Likewise, statements that hold a negative correlation can deflate the results.

The examination of the respondents' use of the response categories through analysis of the thresholds for each item indicated disordered transitions between adjacent response categories in the preliminary scales. This demanded to group adjacent thresholds for some items. This is possible because the RM allows the meaningful comparison of scores even when their levels differ in the scale (Embretson and Reise, 2000). Contrary to many scales that offer too many response options to obtain more information from a respondent, the analysis suggested that such an offer can puzzle respondents, rather than conveying more information or improving precision.

An important analysis of the outcomes is concerned with targeting. Targeting can indicate whether the statements in a measurement structure are appropriately related to the product or to a design feature. Poorly targeted items result in some respondents being above or below the range of measurement captured within the scale, i.e., floor or ceiling effects respectively. One of the characteristics of poorly targeted statements is that they convey little information. This was observed through the analysis of the removed items, such as Item I15 *this chocolate would be nice during a break from housework*. Participants in this case might not have done housework often enough to develop a taste for a treat during a break from it and thus, there was a negative response for the item for all of the stimuli.

Rasch analysis deals with data on an individual basis according to different ability levels. The RM has primarily been applied in education where individual grades are contrasted and in health sciences where a patient treatment has been assessed. By analogy, this allows an analyst in the field of product design to identify what affective reaction would be caused by an improvement of a design element on an individual basis.

3.5.3 Adapting the Model for Applications in Product Design

3.5.3.1 Limitations of the application

Even though the unidimensional measurement structures were co-calibrated, taking the same set of statements, each piece of confectionery obtained its own metric. One of the reasons is that the distinct response patterns yielded chocolate-specific distributions when comparing the chocolates across their individual scales. Take the Stimuli Ferrero Rocher® and Lindor®, for example. Responses on their individual scales indicated that respondents could not tell whether or not these chocolates appeared to be cheap (Item I18) in a consistent fashion. Furthermore, when respondents were directly asked whether these chocolates are special (Item I6), their responses were inconsistent. However, respondents consistently endorsed those chocolates to Item I1 *with this chocolate, you feel as though you are getting more than just a chocolate*, indicating some level of specialness. On the other hand, respondents consistently endorsed Item I18 *this chocolate is cheap* on the individual scale for Caramel® although they were not consistent to tell whether or not this chocolate would show that someone took the time to choose just the right chocolate for the occasion (Item I8), or whether it would make a romantic gift (Item I9). These singularities of responses revealed differences in response pattern for each chocolate. However, the co-calibrated, 12-item set adjusted such differences using consistent response patterns common for all chocolates. Thus, if a respondent endorsed Ferrero Rocher® and Lindor® to Item I10 *this is a premium chocolate*, he or she did not endorse this item for Milky Way®, for example, indicating objective inclinations regarding the chocolates' specialness.

However, those differences in pattern provoked different score systems across stimuli. For example, Item I1 obtained a maximum score of four for Milky Way® and Caramel® although just a maximum score of three for Ferrero Rocher® and two for Lindor®. Consequently, an item can have been plotted at different locations on the scales continuum taking into account each stimulus. Therefore, even though using a common set of items, the stimuli cannot be directly compared through these scales. The solution to solve this problem could be to adapt a derivation of the RM that includes the stimuli as an independent parameter (see Chapter 4).

3.5.3.2 Stability of the measurement structure

In this study the PSI indicated similar values in comparison with the Cronbach's α (i.e., the alpha coefficient of reliability) (Cronbach, 1951), which is commonly used for indicating the index of internal reliability of a scale when using classical psychometric approaches although their calculations are quite dissimilar (see Section 2.4.7.10). It is worth noting that the original scale presented acceptable value of Cronbach's α . However, the value of α is associated with the number of items on the construct (Cortina, 1993). It is possible to obtain a greater value of α when the structure contains a larger number of items and therefore, such a value should be taken cautiously. For this reason, one could interpret that the preliminary structure held enough internal reliability. However, according to Grayson

(2004), data sets with the same α can have very different structures. Additionally, Cortina (2003) demonstrated that for a structure including more than 12 items, and which holds some high correlations ($r > 0.50$), the internal consistency index can exceed the value of 0.70.

Therefore, to avoid assuming that there was modest improvement in the internal reliability of the structure when comparing the indices of the preliminary scales against those of the co-calibrated scales, it would be worth investigating the stability of the structure across different empirical circumstances (see Chapter 6).

3.5.3.3 Using the measurement structure for predicting outcomes

Another matter is concerned with the validity of the measure. Applying the RM has enabled a thorough examination of the validity of the measurement structure because Rasch theory provides procedures to examine how well the data fit together and cooperate to define the attribute being measured.

As far as the Rasch theory is concerned, in a considerably less elaborate prediction about the respondents' attitude in relation to the construct, the validity can be ascertained through the face-valid measure of endorsement of the chocolates' specialness. Clearly the manufacturers of Ferrero Rocher® and Lindor® have displayed their products in the stores shelves and promote their advertising campaigns such that a consumer can decide to purchase them for a special purpose (e.g., to say *thank you* or to choose *the right chocolate for the occasion*). Otherwise, the manufacturer of Celebrations® assortment (i.e., Milky Way® and Caramel®) has posited its product in the market as multi-purpose purchase. The results from participants responses pointed to such products' characteristics and one could take them as means of assessing the construct's validity.

The matter of face validity is controversial. There is consensus amongst researchers that, in terms of scientific measurement, face validity is not sufficient for validating a measurement and therefore, its demonstration is not necessary. Anastasi (1988) suggested face validity should not substitute other forms of validity. On the other hand, Kline (1986) stated participants might not co-operate if the items do not appear to have face validity.

In relation to affective-based experiments for design purposes, the measure's validity ought to reside in the concordance between the theoretical expectations and the empirical results. This is achieved if a construct predicts subsequent performance on some attribute (e.g., specialness). Thus, predictive validity could provide somewhat more useful data about measure's validity because it has greater fidelity to the real situation in which the measure will be used (see Chapter 7).

3.5.4 Concluding Comments

The empirical approach has demonstrated that the data fitted the model to some extent, enabling the conversion of categorical level scales to interval level measures. Nevertheless, respondents presented different response patterns for each confectionery, yielding different measurement structures. Therefore, the stimuli cannot be directly compared through these scales. The development of a same continuum for different stimuli requires adapting a more complex model from the Rasch family of models. One possible solution could be the faceted Rasch approach, which includes one more parameter to the RM (see Chapter 4).

Although the empirical approach demonstrated that the data from affective responses to confectioneries fitted the RM producing calibrated scales, the scales might not be a fully representation of the chocolates' specialness. The scales do not enable one to infer that the calibrated set of items covers all facets of the relevant attribute. However, this is not a drawback of the model, but a difficulty originated from the AE methodology. On the contrary, the model reasonably dealt with the problem when reducing its complexity to one dimension although it is not assumed in this research that there is just one dimension for specialness. The RM approach indicated that within the calibrated scale the items worked well collectively to measure the same attribute and this fact considerably reduces measurement complexity. One of the solutions for dealing with the lack of an extensive representation of the relevant attribute by the scales is to use the calibrated items found in this analysis as a mainstay and then to test new items (i.e., additional statements). Using a smaller sample size, it is possible to verify whether the new items violate the Rasch assumptions and conditions, having as the scale's core the calibrated items found previously (see Chapter 6).

CHAPTER 4

A rationale for eliciting responses to stimulus objects using the multi-facet Rasch model

In this chapter a derivation of the Rasch model, the many-facet Rasch model, is adapted to applications in the domain of product design. A rationale of a common metric for comparing stimulus objects is proposed. Furthermore, based on the theoretical principles of the faceted Rasch model, a set of particular verifications was developed and its applicability was demonstrated when comparing stimulus objects. Such verifications indicate whether the interpretation of a pairwise comparison between curves originated from a cumulative function presents some degree of bias. The practical implication of these verifications is concerned with the reliability to identify differences between stimuli taking into account the measurement errors. The data from the previous study of confectioneries were used to demonstrate the theoretical approach. The results indicated that the faceted Rasch model can overcome the difficulties of comparisons between stimulus objects as a consequence of differences in response pattern[21].

4.1 THE BASIS FOR THE THEORETICAL APPROACH

In the previous empirical approach, each stimulus obtained its own metric. One of the reasons is that the distinct response patterns yielded specific-stimulus distributions across their individual metrics (see Section 3.5.3). Even though using a common set of items, the stimuli cannot directly be compared through those metrics. Thus, the development of a common measurement scale for stimulus objects has required adapting a derivation of the RM. A common scale can help an analyst to know whether the differences between stimuli are consistent.

To verify consistency it is necessary to test whether responses to a stimulus present any bias, whether respondents present unusual profiles of ratings across stimuli and whether it is harder for respondents to decide on the endorsement of a stimulus when comparing against other ones. Furthermore, an analyst might not be able to calibrate items for all stimuli if they are poorly chosen. Even if it is possible to calibrate them all together, inadequate stimuli might provoke inconsistent responses.

[21] Publication based on this chapter can be found in Camargo and Henson (2012c).

A feasible solution is to adapt the many-facet Rasch model (MFRM) (Linacre, 1989). The many-facet approach has frequently been reported in applications where there exists analysis of assessor-mediated ratings, such as analysis of graders behaviour in proficiency exams (Lunz et al, 1990; Myford & Wolfe, 2003, 2004; Engelhard, 1994, 2011), clinical diagnostics and judges' severity in sport competitions. The aim is to provide fair judgements and meaningful measures without the drawbacks of statistical significance tests that do not help to decide what useful fit is. Although there are mentions of the model's usefulness regarding other applications than those for fairness assessment of judges (Linacre, 1989; Eckes, 2009; Wolf, 2009), most of the applications are concentrated on adjusting different degrees of severity (or lenience) of the judges.

The Rasch model has primarily been applied in education where individual responses to questions are contrasted and in health sciences where patient's responses to a treatment have been assessed. In both cases the respondents are assessed, rather than the test conditions. In the domain of product design the interest lies on the conditions, i.e., on the stimuli. The aim is to find the most appropriate design characteristic to manifest a particular affective response. This requires the development of a rationale that incorporates the particularities of the domain. One of these particularities is, for example, the situation where respondents fail to distinguish between stimuli with distinct design features and give ratings in a highly inconsistent fashion across them. This could be consequence of respondents' distinct levels of ability when perceiving physical differences.

4.2 THE MANY-FACET RASCH MODEL

Linacre (1989) introduced the concept of facets to the derivations of the RM for the polytomous case using the rating scale (Andrich, 1978) and using the partial credit model (Master, 1982) (see Section 2.4.5.3).

A facet can be defined as a component or variable of the measurement condition that is assumed to affect the scores in a systematic fashion (Linacre 2002c; Eckes, 2009). There are three facets in the particular case of affective responses to stimulus objects. One of the facets is concerned with persons' responses themselves. A second facet corresponds to the items. The many-facet approach extends the RM and allows including as many facets as necessary to the measurement condition, in this case one more facet called stimulus. Furthermore, the Rasch assumption of specific objectivity based on fundamental measurement (i.e., additivity, invariant comparisons and constant unit) are valid for the many-facet model (Linacre, 1994c). The framework for applying the many-facet Rasch model (MFRM) when there are a number of stimulus objects consists of one facet being replicated across the other facets (Figure 4.1), generating an item location (δ) and its associate standard error (SE) for each stimulus.

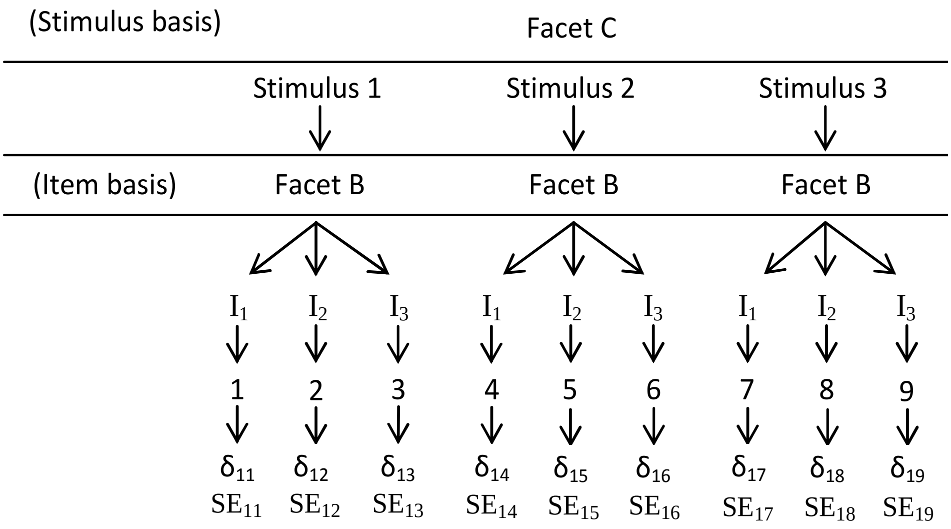


Figure 4.1 - The many-facet Rasch model framework.

The outcomes from the previous empirical approach have shown that every item yielded a particular score system and every stimulus yielded a different response pattern. That is, the items of each stimulus' measurement structure had different intervals between two consecutive categories. This suggests considering the thresholds parameter as a multiplicative term for partial credit (see Section 2.4.5.3). Thus Linacre's MFRM on the basis of partial credit acquires the exponential form of Equation 4.1 (Linacre, 1989).

$$\pi_{nirj} = \frac{\exp\left(\sum_{j=0}^{x}(\beta_n - \delta_i - \gamma_r) - \sum_{j=0}^{x}\tau_{ij}\right)}{\sum_{k=0}^{m}\exp\left(\sum_{j=0}^{k}(\beta_n - \delta_i - \gamma_r) - \sum_{j=0}^{k}\tau_{ij}\right)} \qquad (4.1)$$

where π_{nirj} is the probability of a respondent *n* give a rating of *j* (j=0,1,…,x) on item i for judge *r*; β_n is the ability of a person *n*; δ_i is the difficulty of item *i;* γ_r is the severity of judge *r* and τ_{ij} is the threshold parameter where ij represents the multiplicative term for partial credit.

The general many-facet model can be expressed as a function of the parameter values (Equation 4.2) (Linacre, 1989) such that

$$\Pr(x|\{\Theta\}) = \frac{\exp(F(x|\{\Theta\}))}{\sum_{k=0}^{K}\exp(F(K|\{\Theta\}))} \qquad (4.2)$$

where x is an observed datum, Θ is a parameter value, $\{\Theta\}$ is the set of all parameter values, $\Pr(x|\{\Theta\})$ is the probability of the datum given the parameter values, $F(\ )$ is a linear function which includes only the parameters that are combined to generate the observed datum x, K is the maximum possible value of each observed datum, in terms of which x is the corresponding value empirically observed and k is each of the possible value of the observed datum $(0,...,K)$.

4.3 THE FACETED RASCH APPROACH FOR COMPARING STIMULUS OBJECTS

Analysis of affective responses to stimulus objects is treated in this study as a special case of the MFRM (Linacre, 1989). Thus, an adaptation has been required to the participation of the facet denoted in this thesis as *stimulus fulfilment*. The faceted Rasch approach[22] takes into account that the independent parameters are additive, i.e., they share a linear continuum. The data ought therefore to support a structure with interval measures. Such a condition is examined by the fit of the data to the model for every parameter following the Rasch model's stochastic structure of analysis (see Section 2.4), where the values of the random error associated with the persons' ratings are separated from systematic error.

4.3.1 Items and Stimulus Objects on the Same Continuum

The facets of a measurement situation are represented by independent parameters following the concept of specific objectivity (see Section 2.4.3). When such parameters are combined, it is possible to obtain the probability of a person's endorsement to any item for any stimulus, providing a frame of reference for measurement. Thus, the relative performance level of an affective attribute for a stimulus is expressed by the probability that the persons will endorse the attribute.

Adapting from Linacre (1989), the endorsement level of stimuli S_a and S_b can be established when comparing their relative rating frequencies in any of the categories of a scale. Thus, let F_{kr} be the frequency of ratings in category k for S_a when S_b is rated in category r, and contrariwise for F_{rk}, for all $k > r$ with $r = (k\text{-}1)$. The relative observed

[22] The MFRM (Linacre, 1989) uses unconditional (or also called joint) maximum likelihood estimation method (UCON or JMLE) (Wright and Panchapakesan, 1969), which is an integral part of the software package FACETS® (Linacre, 2012). However, this research used the software package RUMM2030® (2012) with the embedded method of the weighted maximum likelihood estimation (WMLE) (see Section 2.4.6). For a matter of technical distinction between the estimation methods, in this study the approach using RUMM2030® is termed faceted Rasch model.

frequencies can be established by the ratio F_{kr}/F_{rk}. An approximation to the unobservable probability ratio of a rating *k* for S_a when S_b is rated in category *r* is given by

$$\frac{S_{a\,kr}}{S_{b\,rk}} = \frac{\Pr\{(X_a = k, X_b = r)\}}{\Pr\{(X_a = r, X_b = k)\}} \qquad (4.3)$$

which is defined as the ratio of the stimuli's level of endorsement. Where $\Pr\{(X_a = k, X_b = r)\}$ is the probability of a person *n* to give a rating of *k* on item *i* for stimulus *a* when stimulus *b* is given a rating of *r*, and $\Pr\{(X_a = r, X_b = k)\}$ is the probability of a person *n* to give a rating of *r* on item *i* for stimulus *a* when stimulus *b* is given a rating of *k*.

Following the Rasch assumption of response independence, the ratio of the stimulus' endorsement can be written as

$$\frac{S_{a\,kr}}{S_{b\,rk}} = \frac{P(X_{nai} = k) * P(X_{nbi} = r)}{P(X_{nai} = r) * P(X_{nbi} = k)} \qquad (4.4)$$

The concept of fundamental measurement (see Section 2.4.1) requires invariant levels of endorsement when comparing any pair of adjacent categories, given a structure in an ascending order (Linacre, 1989). Thus, the ratio in Equation 4.4 is equal to

$$\frac{S_{a\,kr}}{S_{b\,rk}} = \frac{S_{a\,k'r'}}{S_{b\,r'k'}} = \frac{P(X_{nai} = k') * P(X_{nbi} = r')}{P(X_{nai} = r') * P(X_{nbi} = k')} \qquad (4.5)$$

where *k' > r'* with *r' = (k'-1)*, which gives

$$\frac{P_{niak}}{P_{niar}} = \frac{P_{nia,k'}}{P_{nia,r'}} * \frac{P_{nibk}}{P_{nibr}} * \frac{P_{nib,r'}}{P_{nib,k'}} \qquad (4.6)$$

where P_{niak} denotes the probability of a person *n* to give a rating of category *k* on item *i* for stimulus *a* and P_{niar}, $P_{nia,k'}$, $P_{nia,r'}$, P_{nibk}, P_{nibr}, $P_{nib,k'}$, and $P_{nib,r'}$ are similarly defined.

Taking the level of endorsement to the stimulus *b* at the origin of the scale, denoting as $b_0=0$, item *i* with difficulty at the origin of the scale, denoting it as $i_0=0$, person *n* with a level of endorsement at the origin, denoting it as *n=0* and re-arranging the terms, then Equation 4.6 becomes

$$\frac{P_{00ak}}{P_{00ar}} = \left(\frac{P_{00a,k'}}{P_{00a,r'}} * \frac{P_{000,r'}}{P_{000,k'}} \right) * \frac{P_{000k}}{P_{000r}} \qquad (4.7)$$

in which each term expresses the relationship between a component of the facet stimulus and the arbitrary origin for a particular pair of categories (Linacre, 1989). This condition is independent of the structure of the scale and holds for any parameter. Thus, the comparison of the two stimuli, S_a and S_0, is defined by the ratio of their endorsement level of the relevant affective attribute and it is independent of the person and item locations on the continuum. Letting the term in brackets in Equation 4.7 be represented by P_{00S} , then

$$\zeta_s = \log(P_{00S}) \text{ for all } k \in (0,...,K) \tag{4.8}$$

P_{n00} and P_{0i0} can similarly be defined such that

$$\beta_n = \log(P_{n00}) \text{ and } \delta_i = \log(P_{0i0})$$

Adapting Linacre's Equation 4.1, the probability of a response in category *k* formulated in log-odds unit or logits is given by

$$\frac{P_{nisk}}{P_{nis0}} = \exp\left[k(\beta_n - \delta_i + \zeta_s) - \sum_{m=0}^{k} \tau_m\right] \tag{4.9}$$

where $\tau_0 \equiv 0$. The general exponential form to obtain the probability of a person *n* to give a rating of any category *k* on item *i* for stimulus *s* becomes

$$\pi_{nisk} = \frac{\exp\left[k(\beta_n + \zeta_s - \delta_i) - \sum_{m=0}^{k} \tau_m\right]}{\sum_{k=0}^{K} \exp\left[k(\beta_n + \zeta_s - \delta_i) - \sum_{m=0}^{k} \tau_m\right]} \tag{4.10}$$

which represents the faceted RM, taking the particular form expressed in terms of affective responses to stimulus objects, where π_{nisk} is the probability of a respondent *n* giving a rating of k , $k \in (0,...,K)$, on item *i* for stimulus *s*; β_n is the inclination of a person *n* to endorse the item *i* for stimulus *s*; ζ_s is the level of fulfilment of stimulus *s*; δ_i is the difficulty of endorsement of item *i* and τ_m is the threshold parameter given a rating *k* on item *i* for stimulus *s*. The denominator of this equation is a normalising factor based on the sum of numerators.

The stimulus facet ζ_s takes the positive signal rather than the usual negative signal for the parameter of severity of judgement used in other domains of application. This is justified for understanding that the higher the estimate, the more evident the characteristic of an attribute on a stimulus (Table 4.1).

Table 4.1 - Vector directions for the facets of affective responses.

Person endorsement (β)	Stimulus fulfilment (ζ)	Item difficulty (δ)
↑ Higher parameter estimate	↑ Higher parameter estimate	↑ Higher parameter estimate
↑ Higher level of Endorsement	↑ Higher level of attribute fulfilment	↓ Less persons endorsing
+ Vector	+ Vector	- Vector

4.3.2 Origin of the Logit Scale

One of the characteristics of the RM is that parameter estimates are always located on the scale continuum on the basis of an arbitrary zero of the measurement scale (see Section 2.4.7.3). Usually, the arbitrary zero is established as a default of the method applied in an analysis (Linacre, 1989). Typically, the default for the origin constrains the judge basis facet and the item basis facet at the centre of the logit scale. That is, both facets have a measurement mean of zero. Another constraint is that the sum of the category coefficients comes to zero. Therefore, the sole facet floating on the scale is the person facet. Similar conditions are adopted in the case of responses to stimulus objects. Thus, Equation 4.10 will have the conditions as follows

$$\sum \delta_i = 0; \sum \zeta_s = 0; \sum \tau_{ij} = 0 \tag{4.11}$$

4.4 THEORETICAL APPROACH FOR STIMULUS OBJECTS

4.4.1 Stimulus Fit

Stimulus fit means the degree to which the stimuli were used in a consistent manner. The statistics of stimuli fit are associated with unexpected ratings summarised over respondents and items. The degree of differences between observed ratings and expected ratings indicates the fit to the model. The examination of fit follows the same procedures described in the chocolates' specialness experiment (see Section 2.4.7). This implies investigating the scoring system (i.e., how consistent the responses are when using the available categories in a scale), testing for DIF (i.e., whether responses present any bias on items), testing for local dependency (i.e., whether items are correlated) and testing for unidimensionality (i.e., whether items correspond to a same and sole attribute).

4.4.2 Separation of Stimuli

Separation of stimuli is a particular test for this application of the faceted Rasch model, aiming to verify between-stimuli heterogeneity. Information from data concerning responses to a stimulus will be useful if it is compared against other stimuli. That is, if a design feature varies in degree or type, then the degree of endorsement should vary as well. For instance, Figure 4.2 presents curves, denoted in this study as *stimulus characteristics curves* (SCC), that represent the mean scores of the items for two stimuli used in the chocolates' specialness experiment (see Chapter 3). The lower curve represents the *less special* chocolate and upper curve represents the chocolate that holds more of the attribute specialness. It is noteworthy that the score correlated to a logit location will be higher for chocolates with *higher level of specialness* than a chocolate with *lower level of specialness* at the same logit location. Take in Figure 4.2 the logit location zero where approximately some of the responses are concentrated, for example. At this point on the scale score differences are graphically identified. For Milky Way® the score is approximately 12 and for Ferrero Rocher® is 25.

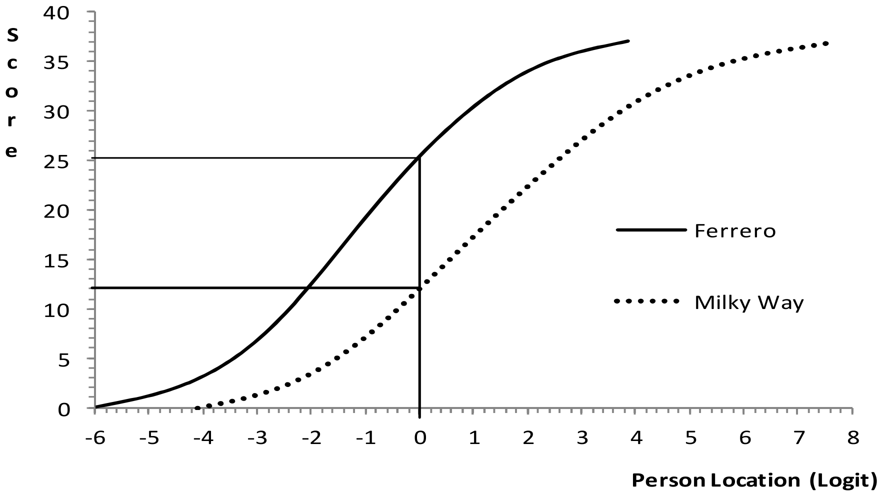


Figure 4.2 - SCCs representing two out of four stimuli used in the chocolates' specialness experiment.

4.4.2.1 Uniform Separation of Stimuli

There are two distinguishable situations with regard to the distinction between features of stimuli for different groups of person ability. The first situation is, in this thesis, called *uniform separation*. This means if the stimuli used in an experiment hold distinguishable features, then their curves will not cross over each other although they could have different slopes.

One of the methods to test such a separation is similar to the test of DIF used in FACETS by Linacre (Wright and Masters, 1982). The difference is, comparing with the test for DIF, to find statistically significant difference of scores at the same location using a pairwise comparison between curves. This takes the form of Equation 4.12

$$t_{j,k} = \sum_{l}^{l'} \left(\frac{\alpha_j - \alpha_k}{\left(SE_j^2 + SE_k^2\right)^{1/2}} \right) \tag{4.12}$$

where $t_{j,k}$ is the level of distinction between curves *j* and *k;* α_j is the score of the higher curve at location *l*; α_k is the score of the lower curve at location *l*; SE_j is the standard error associated with stimulus j measure; SE_k is the standard error associated with stimulus k measure.

The null hypothesis is that the variance between curves is statistically significant. This variance is welcome because it indicates that there is a difference between design features of stimuli and therefore, differences of endorsement are legitimate.

4.4.2.2 Non-uniform Separation of Stimuli

The second situation is when a curve crosses another one. This situation is in this thesis called *non-uniform separation*. A non-uniform separation is a pairwise comparison between curves, where some groups of ability are more inclined to endorse a stimulus in comparison against another stimulus and some groups of ability are less inclined to endorse that same stimulus comparing against the same other stimulus (Figure 4.3). For this reason an analyst cannot directly interpret whether or not a stimulus holds more or less design features for the considered attribute.

The test for uniform separation of stimuli or an ANOVA-like approach, if they are applied in the case of non-uniform separation, cannot provide specific information about which groups are affected. ANOVA can, for example, tell that the means are not equal. However, there are many ways in which the means can differ. If there is enough variance between Stimulus 1 represented by Curve 1 and Stimulus 2 represented by Curve 2 in Figure 4.3, for example, the result might be as follows: $Stimulus2 \succ Stimulus1$.

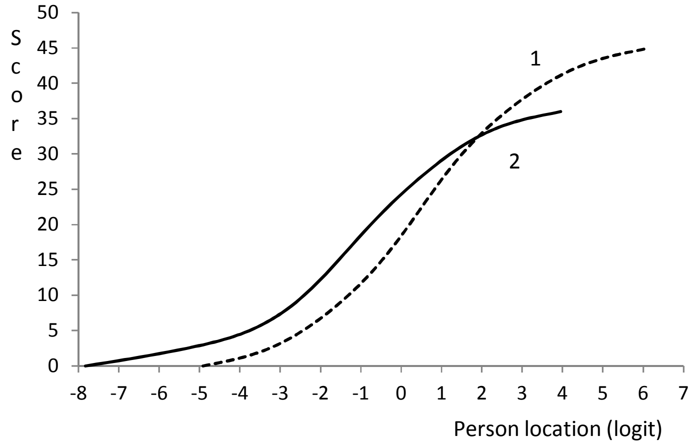


Figure 4.3 – Example of non-uniform separation.

This might not be true, though. For some of the ability groups this relationship corresponds to their affective responses and for other part of the ability groups this is not correct. That is, there is no agreement amongst ability groups even in different levels of endorsement inclination. Also, the comparison between those curves could present significant difference even if $Stimulus1 \equiv Stimulus2$. Thus, the interpretation of such a comparison between crossed curves could be biased.

4.4.3 Differential Stimuli Functioning

The term *differential stimulus functioning* (DSF) is used in this thesis to indicate stimulus bias following the Rasch terminology on the topic. Nevertheless, the approach is somewhat different.

Non-uniform separation of stimuli is indicated for crossed curves, which point to at least part of the information being compromised. Thus, the approach aims to retain as much information as possible, if so, for comparing two stimuli. This implies the need to verify whether after subtracting the areas between curves the difference between stimuli is still significant and for which stimulus the remaining information represents more inclination of endorsement. The area between curves is, therefore, an overall measure of cumulative responses and represents opposite inclinations of endorsement (Figure 4.4).

Take a situation where $\omega \cong \omega'$, for example. If the method indicates non-significant difference between curves, the difference between areas will be $\omega - \omega' \cong 0$. This condition points to the impossibility of assuming that the stimuli, represented by curves, are distinct. Another situation is when $\omega \neq \omega'$. In this case the difference would be $\omega - \omega' > 0$ or $\omega' - \omega > 0$, indicating which stimulus would have preference of endorsement if the difference between them is significant.

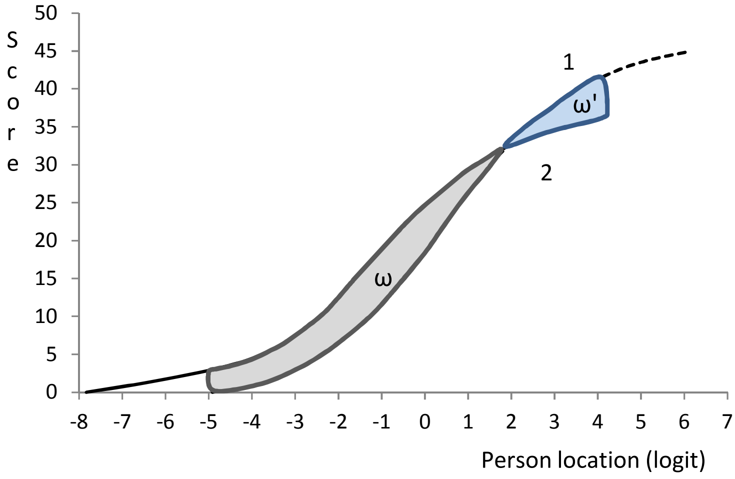


Figure 4.4 – Area between curves of the Stimuli 1 and 2.

Use of areas enclosed between SCCs for different items in a measurement structure as indices of item bias was described by many authors who compared it with other methods of item bias detection (Ironson and Subkoviak, 1979; Rudner et al., 1980; Linn et al., 1981; Shepard et al., 1981). Those authors estimated the area by integrating the appropriate function that originated the curves or by adding successive rectangles between two finite points. Raju (1988) proposed a method for item bias calculating the exact area defined by appropriate integrals taking an infinite interval.

However, there are many limitations for the use of areas as indices of item bias. For example, Raju's approach assumes that the item parameters for the two groups are on a common metric. But item bias is a source of multidimensionality and, consequently, it is difficult to obtain stable estimates of the item parameters without firstly identifying the sources of bias. Furthermore, areas will vary when estimates of item parameters are obtained from small samples as a consequence of sampling error. Also, although the difference between two SCCs is more relevant at the region which most of respondents are in, for item bias detection the region is arbitrarily established. This seems as though there is more information at the ability level than without the fixed bounds.

Although those limitations could yield misleading interpretations of item bias, the use of areas could be precise enough for verifying if two stimuli are different. When analysing such a difference, parameters have already been estimated, i.e., the measurement structure contains calibrated items that shall yield stable results. This eliminates many of the limitations concerning item bias. Furthermore, when comparing stimuli, any item bias has already been identified as well as ability groups. This opens the possibility to establishing the boundaries on the logit scale, such that it includes all the ability groups.

The rationale to obtain the area between curves that represent the stimuli and the area for comparison is obtained as follows:

Let Curve O_1 and Curve O_2 represent Stimulus S_1 and Stimulus S_2 respectively and the probabilities $\Pr$ and $\Pr'$ be obtained from a cumulative function based on a set of parameter values $\{\theta\}$.

The set of parameters $\{\theta\}$ is established according to the Rasch model used during analysis. The set of parameters of the dichotomous case for the conditional probability of x is given by $\Pr\{x|\beta,\delta\}$ (see Section 2.4.5.1). The linear function through the dichotomous RM in terms of a logit model can be expressed by

$$\ln\left(\frac{P_{ni}}{1-P_{ni}}\right)=\beta_n-\delta_i \tag{4.13}$$

where $\ln\left[P_{ni}(1-P_{ni})^{-1}\right]$ is denoted as log-odds unit or logits, β_n is the parameter of person n and δ_i is the parameter of item i. For the polytomous case the probability of x is established by $\Pr\{x|\beta,\delta,\underline{\tau}\}$ (see Section 2.4.5.3). The logit form of the polytomous RM is given by

$$\ln\left(\frac{P_{ni,k}}{P_{ni,k-1}}\right)=\beta_n-\delta_i-\tau_k \qquad (4.14)$$

where the term $\ln\left[P_{ni,k}\left(P_{ni,k-1}\right)^{-1}\right]$ is denoted as log-odds unit or logits and τ_k is the threshold parameter for category *k*. For the particular case of the probability of a person's endorsement to items and physical elements of products the linear function is given by $\Pr\{x|\beta,\delta,\zeta,\underline{\tau}\}$ (see Section 4.3.1). Thus, in logit terms, Equation 4.10 can be expressed by

$$\ln\left(\frac{P_{nis,k}}{P_{nis,k-1}}\right)=\beta_n-\delta_i+\zeta_s-\tau_k \qquad (4.15)$$

where $\ln\left[P_{nis,k}\left(P_{nis,k-1}\right)^{-1}\right]$ is denoted as log-odds unit or logits and ζ_s is the parameter associate with stimulus *s*.

From the general Equation 4.2, expressed as a function of the parameter values, let $F(\ )$ be a linear function which includes only the parameters that are combined to generate the observed datum x for Curve 1, and x' for Curve 2, both on item *n* such that

$$\forall\ x: x=F\left(x|\{\theta\}\right) \text{ and } \forall\ x': x'=F'\left(x|\{\theta\}\right) \qquad (4.16)$$

where the locations x and x' are obtained from the average of the set of calibrated items expressed as

$$x=\frac{1}{n}\sum_{n=1}^{n}x_n \text{ and } x'=\frac{1}{n}\sum_{n=1}^{n}x'_n \qquad (4.17)$$

given the conditions specified in Equation 4.11. Since both stimuli are on the same continuum with a set of calibrated items *n*, then $x=x'$. Thus, the comparison between stimuli at a location x is solely given by the difference between the associated probability of endorsement P represented by the two curves, such that

$$\Pr=\Pr\left(x|\{\theta\}\right)=\frac{\exp\left(F\left(x|\{\theta\}\right)\right)}{\sum_{k=0}^{K}\exp\left(F\left(K|\{\theta\}\right)\right)} \qquad (4.18)$$

and

$$\Pr'=\Pr'\left(x|\{\theta\}\right)=\frac{\exp\left(F'\left(x|\{\theta\}\right)\right)}{\sum_{k=0}^{K}\exp\left(F'\left(K|\{\theta\}\right)\right)} \qquad (4.19)$$

thus,

$$P=\left|\Pr-\Pr'\right| \ . \qquad (4.20)$$

Note that the cumulative function of P is monotonic, right-continuous, non-decreasing as characteristics of the RM, where

$$\lim_{x\to-\infty} F(x) = 0 \text{ and } \lim_{x\to\infty} F(x) = 1 \tag{4.21}$$

thus,

$$\Lambda = \frac{1}{n}\sum_{n=1}^{n}\int_{-\infty}^{\infty} P(x|\{\theta\})dx\,. \tag{4.22}$$

However, the location of interest for comparison is in the range of the higher frequency of responses, which is defined as x and x', limited to the minimum and the maximum location common for both curves. Thus, through Equation 4.16 it is found that

$$\Lambda = \frac{1}{n}\sum_{n=1}^{n}\int_{l}^{l'} P(x|\{\theta\})dx \tag{4.23}$$

where $P(x|\{\theta\})$ is a linear function which includes only the parameters which are combined to generate the probability associated with a point x, θ is a parameter value, $\{\theta\}$ is the set of all parameter values, x represents an observed datum, n is the number of items in the scale, l is the minimum value of common location between two curves, l' is the maximum value of common location between two curves.

The area between curves can further be expressed in terms of persons' scores associated with locations on the continuum. Let Curve O_1 and Curve O_2 represent Stimulus S_1 and Stimulus S_2 respectively. Let areas ω and ω' be obtained from a cumulative function based on a set of parameter values $\{\theta\}$.

If the stimuli are well spread on the same scale continuum, i.e., there are few responses concentrated on the extremes of the scale, and since the scores represent responses to the same set of calibrated items for all stimuli, then there is a high correlation between the scores and the Rasch measure.

An approximate association[23] between the total person score z for a stimulus object s and the logit measures can be obtained by[24]

$$z = \sum_{i=1}^{I_S}\frac{e^{\beta_n - \delta_{ij}}}{1+e^{\beta_n - \delta_{ij}}} \tag{4.24}$$

[23] The association between raw scores and Rasch measures in logits can directly be obtained from the software package RUMM2030®.

[24] An alternative method for conversion between raw scores and Rasch measures can be found in Linacre (2002c).

where I_s is the number of calibrated items related to the stimulus s, β_n is the person location equivalent to stimulus total score and δ_{ij} the location estimate of the threshold j of item i for the stimulus s. Similarly, the association between the total score z' and the linear measures can be obtained for a second stimulus object s'

$$z' = \sum_{i=1}^{I_{s'}} \frac{e^{\beta_n - \delta'_{ij}}}{1 + e^{\beta_n - \delta'_{ij}}} \tag{4.25}$$

where $I_{s'}$ is the number of calibrated items related to the stimulus s', β_n is the person location equivalent to total score of the stimulus and δ'_{ij} is the location estimate of the threshold j of item i for stimulus s'.

The value of β_n is obtained by the substitution of δ_{ij} estimates produced from the calibration of items into Equation 4.24 and establishing a value of the total score z with integers from 1 to the maximum score value minus 1. The score values are then plotted against the values of β_n to establish the coordinates of the curve for stimulus s. Similarly, resolving Equation 4.25 for β_n by the substitution of δ'_{ij} and z', the curve representing stimulus s' can be constructed.

Taking z as the score for Curve 1 and z' for Curve 2 at location x, then the difference between scores at the location x is given by

$$Z = |z - z'| \tag{4.26}$$

The cumulative function expressed as the area between curves is then

$$\lambda = \int_x^{x'} |\omega(Z|\{\theta\})| \; dx \tag{4.27}$$

and taking the average of the set of calibrated items, then the difference between those curves is given by the area enclosed between them, such that

$$\Lambda = \frac{1}{n} \sum_{n=1}^{n} \int_x^{x'} |\omega \; (Z|\{\theta\})| \; dx \tag{4.28}$$

where $\omega(Z|\{\theta\})$ is a linear function which includes only the parameters which combined to generate the difference between curves, θ is a parameter value, $\{\theta\}$ is the set of all parameter values, Z represents each difference of score values between curves, n is the number of items in the scale, x is the minimum value of location between two curves, x' is the maximum value of location between two curves.

When the situation of non-uniform separation of stimuli[25] arises (i.e., when a curve crosses another one), Equation 4.28 is extended. Let Curve 1, which represents Stimulus 1 and Curve 2, which represents Stimulus 2, intersect at point x. Let areas ω and ω' be obtained from a cumulative function of raw scores based on a set of parameter values $\{\theta\}$ and be opposite in direction. The difference between those curves is given by the difference of two areas enclosed between them, such that

$$\Lambda = \frac{1}{n}\sum_{n_i=1}^{n}\left(\int_{x_1}^{x} \omega \ \left(Z\middle|\{\theta\}\right)dx - \int_{x}^{x_2} \omega'\left(Z\middle|\{\theta\}\right)dx\right) \tag{4.29}$$

where x is the point of intersection of two curves projected on the logit scale, x_1 is the minimum value of location common to two curves and x_2 is the maximum value of location common to two curves.

The areas ω and ω' are obtained when integrating the difference of the polynomials that define the curves. The reference point is the intersection of two curves projected on the logit location axis. The intersection point x is found when the two polynomials are equalled, such that $y_1 = y_2$.

It is noteworthy that if $\Lambda > 0$, then the remaining area enclosed between curves represents upper curve $\succ$ lower curve. Consequently, if $\Lambda < 0$ then the condition of lower curve $\succ$ upper curve is established. Both of the cases indicate relative preference in a pairwise comparison.

4.4.3.1 Imparity criterion

Separation of stimuli depends on the characteristics of the items that establish the measurement of an attribute and of the context in which the stimuli will be compared. Some cases will require large separation between stimuli if a clear effect on the affective responses is sought. Thus, a cut-off criterion based on statistical significance might not indicate whether stimuli are sufficiently different to distinguish design characteristics. In this study the differentiation between SCCs, which represents two independent stimuli, was obtained through of a comparison between the area computed between those curves and an imparity criterion[26].

[25] The term non-uniform stimuli separation is defined for this research and has solely been applied in it (see Section 4.4.2.2).

[26] Although the method for finding statistically significant difference of scores at the same location using a pairwise comparison between curves could be applied to test uniform stimuli separation, the method

Let $\omega(\varepsilon)$ represent the area enclosed between the curve derived from a cumulative function based on a set of parameter values $\{\theta\}$ and the curve obtained from the same cumulative function plus its standard error for the person location. Given a pair of curves at a relative upper position and a second pair of curves at a relative lower position, based on the person location on the continuum, their areas will subsequently be denoted as $\omega(\varepsilon)_{upper}$ and $\omega(\varepsilon)_{lower}$, respectively (Figure 4.5).

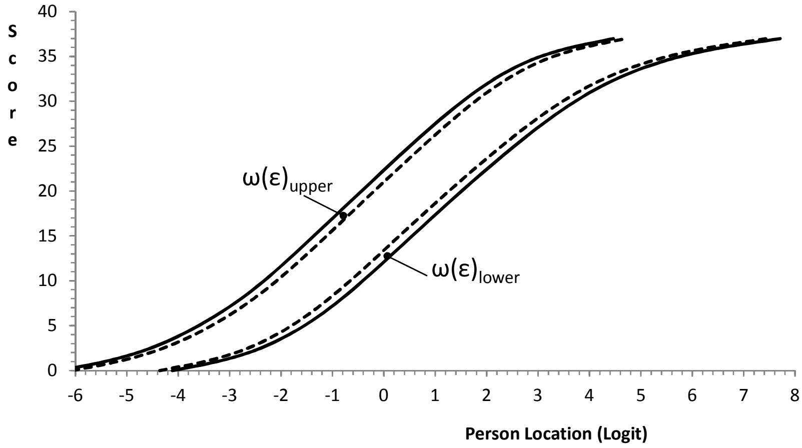


Figure 4.5 – Upper and lower areas between the curves representing the cumulative function for a stimulus and curve including the standard error.

Assuming that

$$\forall\ x : x = F(x|\{\theta\}) \text{ and } x(\varepsilon) = x + \varepsilon \tag{4.30}$$

$$\omega(\varepsilon)_{upper} = \frac{1}{n}\sum_{n=1}^{n}\left(\int_{x_1}^{x_2}\left(y(\varepsilon)_{upper} - y_{upper}\right)dx\right) \tag{4.31}$$

and

$$\omega(\varepsilon)_{lower} = \frac{1}{n}\sum_{n=1}^{n}\left(\int_{x_1}^{x_2}\left(y(\varepsilon)_{lower} - y_{lower}\right)dx\right) \tag{4.32}$$

then the imparity criterion is given by the sum of those two areas, such that

$$\omega(\varepsilon) = \omega(\varepsilon)_{upper} + \omega(\varepsilon)_{lower} \tag{4.33}$$

proposed to deal with uniform and non-uniform separation in this study is the comparison of areas enclosed between curves against the imparity criterion.

4.4.4 Applying the Faceted Rasch Approach

Data from the prior experiment concerning specialness of wrapped confectionery (see Chapter 3) was used to test the hypothesis of applicability of the faceted Rasch approach in the domain (see Section 1.6.1). To use the approach through the software package RUMM2030®, licensed version (2011), was necessary to ensure that the individual scales of the four chocolates had previously been calibrated. Thus, the steps for items calibration, the tests for detecting DIF, and the tests for the Rasch assumptions of local independence and unidimensionality are not herein presented (see Chapter 3). The pool of statements used in the facet analysis was the co-calibrated 12-item set as well as the common score system across stimuli. Thus, the faceted approach replicated the calibrated set, generating a 48-item set (see Section 4.2).

The summary of the facet locations is presented in Table 4.2. An arbitrary zero was established as the default of the method applied in the analysis. The default for the origin constrained the stimuli facet (See Section 4.5), the items facet and the sum of the category coefficients at the centre of the logit scale (see Section 4.3.1). The fit statistics for the calibrated scale[27] indicated invariance across the measurement structure with p ≥0.05. Furthermore, the scale presented PSI of 0.88, considered as indication of reliability when differentiating three groups of ability (Fisher, 1992).

The facets map (Figure 4.6) is the representation of the relative locations of all facets on the same logit scale. Person locations are plotted on the scale represented in the first column. Participant locations that indicate more inclination to endorse the attribute specialness of the pieces of wrapped confectionery are plotted on the top of the scale and those less inclined to endorse at the bottom. The top of the second column of the facets map indicates items that are more difficult to endorse, i.e., items that obtained less consensus amongst participants to endorse them. The location of stimuli on the continuum demonstrated that Milky Way® was posited at the bottom of the scale in relation to the confectioneries with higher endorsement to specialness. The facets map also identified that there was shrinkage of the persons spread on the logit scale when the common metric was applied for all of the stimuli. However, Figure 4.7 indicates that the threshold distribution is widely spread, revealing that the respondents are well targeted to the set of calibrated items.

[27] The locations of stimuli obtained through RUMM2030® were multiplied by minus one. The reason was that the RUMM2030® software package version 2011 has not allowed other configuration, such as in Equation 4.10, but that usually used for fairness of judgement. Thus, the stimulus locations are presented in a proper magnitude although they are placed at reversed locations on the continuum.

Table 4.2 – Fit statistics of Facet approach

Stimulus facet				Item facet				Metric			
Stimulus	Mean location	*SE*	Fit-Res	Items	Mean location	*SE*	Fit-Res	Chi-Square	df	p	PSI
Ferrero®	1.08	0.10	0.58	I1	0.15	0.10	0.19	199.54	192	0.34	0.88
Lindor®	0,78	0.10	0.12	I8	0.23	0.10	-0.07				
Caramel®	-0.50	0.10	-0.19	I10	0.35	0.10	0.03				
Milky Way®	-1.37	0.10	0.28	I11	0.15	0.08	0.76				
				I12	-0.10	0.09	0.28				
				I13	-0.31	0.11	0.14				
				I14	-0.28	0.10	-0.15				
				I17	-0.16	0.10	-0.05				
				I19	-0.38	0.10	-0.32				
				I20	0.10	0.10	-0.11				
				I22	0.03	0.10	1.14				
				I23	0.20	0.09	0.45				

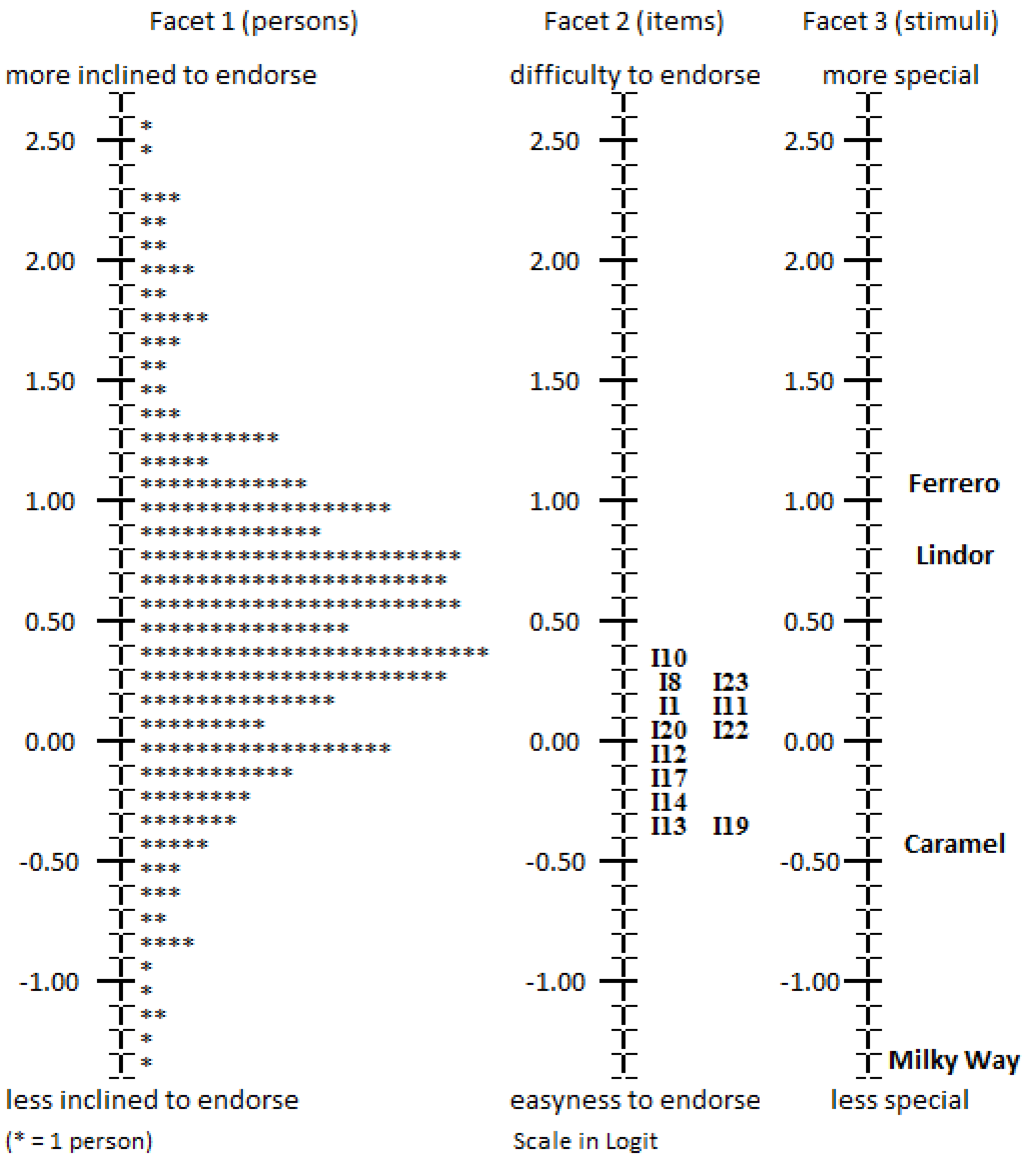


Figure 4.6 – Facets map for the specialness of four wrapped confectioneries.

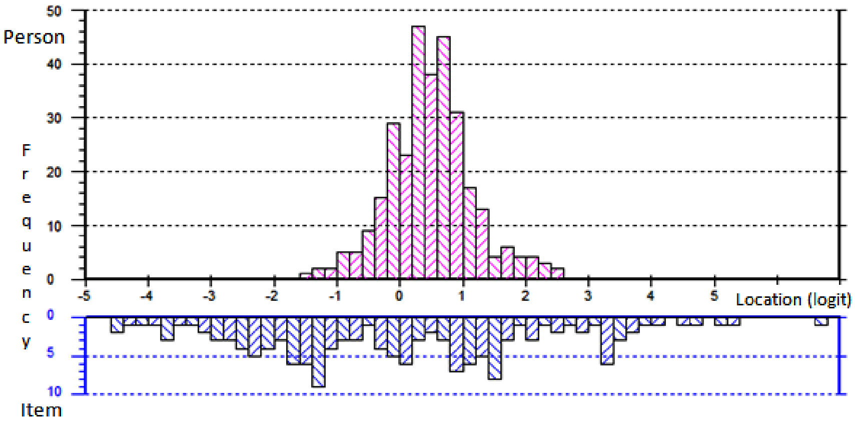


Figure 4.7 - Person-item threshold distribution using faceted Rasch model.

4.4.4.1 Differential stimuli functioning

The SCCs were obtained from a cumulative function of raw scores based on the set of parameter values computed by RUMM2030® (see Section 4.4). The areas were obtained by integrating the difference of the polynomials that defined the SCCs. For practical purposes in this study polynomials of 5th and 6th order were considered approximate mathematical representations of the curves (see Section 4.4.3).

Thus, let the curve denoted Ferrero® in Figure 4.8 be represented by Equation 4.34, let the Curve Lindor® be represented by Equation 4.35, let the Curve Caramel® be represented by Equation 4.36 and let the Curve Milky Way® be represented by Equation 4.37 such that

$$y_1 = -0.0003x^6 + 0.0009x^5 + 0.0239x^4 - 0.1089x^3 - 0.6281x^2 + 5.7679x + 25.355 \qquad (4.34)$$

$$y_2 = 0.0009x^5 + 0.0019x^4 - 0.0973x^3 - 0.1671x^2 + 5.4289x + 22.397 \qquad (4.35)$$

$$y_3 = -0.0001x^6 + 0.0004x^5 + 0.0074x^4 - 0.0604x^3 - 0.0747x^2 + 4.699x + 16.783 \qquad (4.36)$$

$$y_4 = -0.00006x^6 + 0.0013x^5 - 0.0017x^4 - 0.099x^3 + 0.2595x^2 + 5.1119x + 12.034 \qquad (4.37)$$

Because the SCCs for Ferrero® and Lindor® crossed over each other, the reference point was the intersection of those two curves projected on the logit location axis. Thus, the intersection point *x* between Curve Ferrero® and Curve Lindor® is given by equalling y_1 and y_2, such that $y_1 - y_2 = 0$; thus, *x* = -2.67

Let area ω and area ω' be represented by Equation 4.38 and Equation 4.39, respectively, both equations derived from Equation 4.33, such that

$$\omega' = \frac{1}{n}\sum_{n=1}^{n}\int_{x}^{x_2}(y_1 - y_2)dx \qquad (4.38)$$

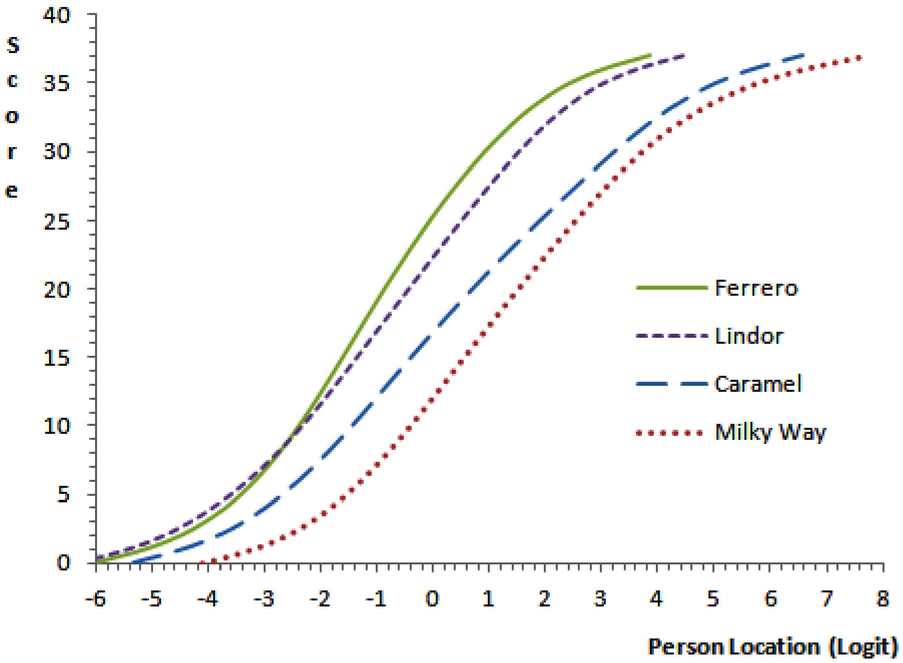


Figure 4.8 - Graphic representation of the four stimuli used in the previous empirical study on the logit continuum. The stimuli's SCCs were originated from a set of parameter values $\{\theta\}$ and their respective polynomial equations. The area bounds are defined as the minimum location and the maximum location common to all curves.

and

$$\omega = \frac{1}{n}\sum_{n=1}^{n}\int_{x_1}^{x_2}\left(y_{upper} - y_{lower}\right)dx \tag{4.39}$$

where y_{upper} and y_{lower} are the polynomial expressions that represent the upper curve and the lower curve respectively. Letting $x_1 = -4$ and $x_2 = 4$ as the approximate minimum and the maximum location values common to all curves, the approximate values for ω and ω' are obtained by integrating Equations 4.38 and 4.39 (Table 4.3). A similar calculation was used to obtain the areas between the other pairs of stimuli, taking $\omega' = 0$ because no intersection is observed in Figure 4.8 for any other pair of curves (Table 4.3).

To calculate the imparity criterion established by Equation 4.33, the area enclosed between two curves was computed through the person locations and the person locations along with their respective standard errors[28] (i.e, $\bar{\varepsilon} = 0.25$). The pairwise comparisons through the enclosed area (ω) were greater than the imparity criterion [*ω(ε)*], indicating dissimilarity between stimuli except for the comparison between Ferrero Rocher® and Lindor®, indicated in bold (Table 4.3). In this case the information conveyed by the

[28] SE obtained by the mean of standard errors for the person locations.

affective responses using the metric cannot distinguish the attribute specialness between those two stimuli.

Table 4.3 – Comparison of areas between SCCs and areas between SCCs along with errors

	Ferrero Rocher®		Lindor®		Caramel®		Milky Way®	
	A	Ae	A	Ae	A	Ae	A	Ae
Ferrero®	x	x	x	x	x	x	x	x
Lindor®	**2.67**	**5.83**	x	x	x	x	x	x
Caramel®	50.79	16.29	39.19	15.99	x	x	x	x
Milky®	78.06	16.37	66.46	16.01	27.27	15.45	x	x

Looking at the facets map (Figure 4.6) and the comparisons of areas (Table 4.3) the specialness of the stimuli could be measured such that $Ferrero \approx Lindor \succ Caramel \succ MilkyWay$.

4.5 REMARKS

A different solution to solve the problem of comparing stimuli on the same continuum was previously endeavoured. The approach established the group of stimuli as a factor and analysing it through an ANOVA and regression methods. Using this solution, one could find mathematically precise results based on flawed assumptions, though. The reason is that in affective experiments is not obvious to decide on what items are better fitted to a stimulus. The emphasis on finding the *best-of-fit* items that appropriately characterise a given set of data for a stimulus could not simultaneously suit the *best-of-fit* items for another stimulus. That is, some items that work for a stimulus could not work for other ones in the same scale. This happened when analysing data of different chocolates as a factor through an ANOVA in the experiment on chocolate specialness and obtaining a sole scale for all of the stimuli. As a collateral effect 20 out of 24 items had to be removed from the measurement structure, representing 6,120 responses taken as misfits.

The faceted Rasch model enlarged the frame of reference when using the facet denoted in this study as *stimulus fulfilment*. Within the frame of reference the four pieces of confectionery fitted the model. Although the individual co-calibrations did not indicate local dependence (see Chapter 3), the enlarged frame of reference produced new correlations between items, indicating some degree of response dependence. This anomaly in the new relationship between variables can have an important impact on the measure interpretation and therefore, it ought to be investigated (see Chapter 5).

4.5.1 Concluding Comments

The research reported in this chapter has led to a novel contribution in the domain of product design. The difficulty of comparing stimuli that hold different design features on the same continuum configures a particular situation which required a more complex solution other than merely using statistical procedures.

The applications of the faceted Rasch model along with the verifications developed in this research have allowed the identification of objective differences between persons, between statements and between stimuli. Evidence that the hypothesis presented in Section 4.4.4 on the applicability of the faceted Rasch approach in the domain was given in the empirical study because the model parameterises the person facet, the item facet and the stimulus facet independently. Even though the measurement structure could be applied to a different sample of respondents, the location of items and stimuli would nearly be the same. That is, the measurement structure would be stable (see Chapter 6).

CHAPTER 5

Response dependence in a multi-conditional frame of reference

In this chapter the use of an enlarged frame of reference obtained from calibration of items through the faceted Rasch model is analysed in typical applications of affective engineering. The research is concerned with anomalies that have potential impact on measure interpretation. Data obtained in the previous empirical study using four pieces of confectionery fitted the model. However, local dependence was identified in the enlarged frame of reference. This dependence could be a consequence of the framework of the faceted approach and not a characteristic of the data. An alternative approach is postulated to form a structure of subtests when using items replicated across different conditions. In the approach, similar items of each condition are grouped to produce an enlarged item. To meet the assumptions of a quantitative structure, the persons' locations ought to be invariant over such subtests. Thus, the subtests were not formed by the facet conditions but by originating from individual items. On the other hand, the approach does not obscure signals of local dependence if they do exist in the data[29, 30].

5.1 FRAME OF REFERENCE IN THE CONTEXT OF PRODUCT DESIGN

Results from the application of the RM to develop a measurement structure for four pieces of confectionery demonstrated that the scales with the calibrated, 12-item set have fitted the model (see Section 3.4.4); however, they are in different frames of reference such as it is shown by the generic representation in Table 5.1. Those different structures were then re-calibrated through the multi-faceted approach (see Chapter 4) presenting an enlarged frame of reference, which embraced all four confectioneries, as generically represented by Table 5.2, where items are replicated to obtain persons' responses to different conditions (e.g., variations in the design feature) (see Section 5.6). The metric developed from the faceted approach ought to present the propriety of invariant comparisons. As a consequence, the scale would be thought additive and therefore, it would serve for the purpose of measurement.

[29] Part of the research reported in this chapter was undertaken at the Graduate School of Education of the University of Western Australia supported by a grant from the University of Leeds associated with the Research Mobility Programme (RMP) of the Worldwide Universities Network (WUN).

[30] The research reported in this chapter was presented and discussed in the 2013 Rasch working group meeting (Camargo, 2013).

Table 5.1 - Frame of reference for condition 1 and frame of reference for condition *S*.

	Condition 1			
Person	Item I_1	Item I_2	...	Item I_i
P_1	x_{111}	x_{121}	...	x_{1i1}
P_2	x_{211}	x_{121}	...	x_{2i1}
⋮	⋮	⋮		⋮
P_n	x_{n11}	x_{n21}	...	x_{ni1}

	Condition S			
Person	Item I_1	Item I_2	...	Item I_i
P_1	x_{11S}	x_{12S}	...	x_{1iS}
P_2	x_{21S}	x_{22S}	...	x_{2iS}
⋮	⋮	⋮		⋮
P_n	x_{n1S}	x_{n2S}	...	x_{niS}

Table 5.2 - Frame of reference for conditions 1 and *S*.

	Condition 1				*Condition S*			
Person	Item I_{11}	Item I_{21}	...	Item I_{i1}	Item I_{1S}	Item I_{2S}	...	Item I_{iS}
P_1	x_{111}	x_{121}	...	x_{1i1}	x_{11S}	x_{12S}	...	x_{1iS}
P_2	x_{211}	x_{121}	...	x_{2i1}	x_{21S}	x_{22S}	...	x_{2iS}
⋮	⋮	⋮		⋮	⋮	⋮		⋮
P_n	x_{n11}	x_{n21}	...	x_{ni1}	x_{n1S}	x_{n2S}	...	x_{niS}

Rasch (1961) employed the term *frame of reference* to formalise the relationship between a class of individuals and a class of items in a well-defined context established by the relevant attribute being measured. Thus, if the data fit the model, the total raw score will be sufficient statistic to estimate the model's parameters (see Section 2.4.5.2). That is, the comparison between any two persons will be independent of the items and the comparison between any two items will independently be obtained with regard to the locations of persons on the continuum within a frame of reference[31] (see Section 2.4.4). Furthermore, it has been demonstrated that when the same latent trait (i.e., the same attribute) is measured across different frames of reference with data fitting the model, the property of sufficiency is conserved (Humphry and Andrich, 2008). Thus, the dichotomous case can be expressed such that

$$\Pr\{(1,0);(1,0)|(r_{sn}=1,r_{tn}=1)\}=\frac{e^{-\delta_{si}-\delta_{ti}}}{\left(e^{-\delta_{si}-\delta_{ti}}+e^{-\delta_{si}-\delta_{tj}}+e^{-\delta_{sj}-\delta_{ti}}+e^{-\delta_{sj}-\delta_{tj}}\right)} \tag{5.1}$$

[31] In terms of Rasch modelling, each frame of reference holds a unit. However, Humphry (2011) demonstrated that the existence of different units does not directly represent a violation of the Rasch model, taking into account the same relevant attribute and context.

where $x_{sni} \in \{0,1\}$, $x_{tni} \in \{0,1\}$, $r_{sn} = \sum_{i \in s} x_{sni}$ and $r_{tn} = \sum_{i \in t} x_{tni}$ represent the raw score of a person in frame *s* and in frame *t*, respectively, and δ is the item location, which is independent of the persons parameter (Humphry and Andrich, 2008).

5.2 RESPONSE DEPENDENCE ACROSS ITEMS

Dependence across items is concerned with the RM assumption of statistical independence of person's responses. The violation of this assumption is identified when the probability of a person's response to an item is governed by his or her response to a previous item. The implications of response dependence in an analysis have frequently been discussed in the literature (Andrich, 1985b; Smith, 2005; Marais and Andrich, 2008a, 2008b).

Furthermore, response dependence can be a consequence of a halo effect (Engellard, 1994; Myford and Wolf, 2004; Eckes, 2005). Marais and Andrich (2011) has suggested a halo effect exists when a sole judge or a rater assessing some performance on multiple items, common activity in the domain of education and in psychology, would present a greater association amongst the items than if different judges or raters assessed the different items independently.

The interpretation of a halo effect from the perspective of the application in Section 4.4.4 has suggested that some responses to a stimulus could have interfered with responses to a second stimulus. That is, a person could have rated stimulus *s* based on the comparison with a stimulus *t* previously presented to him or her, indicating a greater association amongst items than if he or she had taken into account stimulus *r* and *t* independently. Nevertheless, some degree of the halo effect is not unexpected as a consequence of the design of affective engineering experiments themselves.

Marais and Andrich (2008b) have addressed the dependence process for dichotomous items using a different approach, rather than using correlation coefficients. They have calculated the magnitude of the violation of the assumption as an alteration in the location of thresholds on an item caused by the response dependence on another item. In their formulation the statistical independence in a dichotomous RM is represented by Equation 5.2, such that

$$\Pr\{X_{nj} = x_j | X_{ni} = x_i\} = \Pr\{X_{nj} = x_j\} \tag{5.2}$$

To violate this condition they introduced a dependence effect *d*, such that

$$\Pr\{X_{nj} = 1 | X_{ni} = 1\} = \exp(\beta_n - (\delta_j - d)) / [1 + \exp(\beta_n - (\delta_j - d))] \tag{5.3}$$

and

$$\Pr\{X_{nj} = 1 | X_{ni} = 0\} = \exp(\beta_n - (\delta_j + d)) / [1 + \exp(\beta_n - (\delta_j + d))] \tag{5.4}$$

where $d \neq 0$ is the magnitude of the dependence. Based on the above formulation Andrich and Kreiner (2010) demonstrated that the dependence effect *d* could be estimated and its deviation tested for statistical significance.

Andrich et al. (2012) extended the approach for polytomous items. They formulated the case based on the interpretation that response dependence induces to an alteration in the range of the continuum. Thus, letting $\delta_{kj}^{*} = \delta_{kj} - d$ for all $k \leq x_i$ and $\delta_{kj}^{*} = \delta_{kj} + d$ for all $k > x_i$, then

$$\Pr\{X_{nj} | X_{ni} = x_i\} = \left[\exp\left(\psi_{xj}^{*} + x_j \beta_n\right)\right] / \phi_{nj} \tag{5.5}$$

where $\psi_{xj}^{*} = -\sum_{k=0}^{x_i} \delta_{kj}^{*}$, with $k = 0,1,2,...,m_i; \delta_{0j}^{*} \equiv 0$ and $\phi_{nj} = \sum_{x=0}^{m_j} \exp\left(\psi_{xj} + x_j \beta_n\right)$.

An estimate of *d* is obtained from each of the m_j thresholds. Thus, the mean of the m_j estimates $\hat{d}_k, k = 1,2,...,m_j$ is formulated such that

$$\hat{d}_k = \frac{\sum_{k=1}^{m_j} \left[\hat{\delta}_{ji(k=x)(x_i-1)} - \hat{\delta}_{ji(k=x)(x_i)}\right] / 2}{m_j} = \frac{\sum_{k=1}^{m_j} \hat{d}_{jk}}{m_j} \tag{5.6}$$

According to Andrich et al. (2012), if $\hat{d} < 0$, a response x_{ni} in the independent item *i* would imply a smaller probability of a response x_{ni} in the dependent item *j*. The variance of the mean of the m_j estimates is given by Equation 5.7 as follows

$$\hat{\sigma}^2 = \frac{\sum_{k=1}^{m_j} \sum_{x=k-1}^{k} \hat{\sigma}_{ji(k)(x)}^2}{4m_j^2} \tag{5.7}$$

The method briefly described above supports decisions on suspected violations of response independence that hold a statistically significant impact in fitting data to the model. However, this method focuses on investigating response dependence in the base-level items and could perhaps not be appropriate in many cases that use a multi-conditional frame of reference.

5.3 INTERPRETATION OF RESPONSE DEPENDENCE IN A MULTI-CONDITIONAL FRAME OF REFERENCE

A frequent problem occurs when a measurement structure is composed of items that have the same or similar stimulus objects for subsets of items or if different subsets of items share other features (Wilson, 1988). Evidence of this problem is given by the residual correlation found in the case of confectioneries when using the enlarged frame of reference. Analysis of the residuals indicated different correlation indices compared with the residual correlation index found in the individual scale of each stimulus (see Section 3.4.4). When extending the frame of reference, a new sort of relationship was formed, pointing to response dependence as a consequence of a higher index of person-item residual correlation for some items than when taking the individual scales.

Two main reasons can induce such a correlation in a multi-conditional frame of reference. Firstly, in the multi-faceted framework the items are replicated across conditions. Thus, when estimating parameters, the model will consider a larger number of items. In the confectionery case, for example, 12 items were replicated across four different conditions (i.e., four chocolates) making 48 items. Such a procedure will derive different degrees of dependence between items. Thus, if one includes more conditions in that frame of reference, then there will be a higher degree of response dependence. Secondly, because in most of the cases when different conditions are compared, some degree of similarities will be associated amongst them; for example, the comparison of the effect of colours on the user's impression of a same sport car model and also the comparison between different materials although in the view of the same product's shape and colour. In each case an analyst is interested in the unidimensional latent attribute that underlies the unique measurement structure. However, the multi-faceted procedure forces the analyst to deal with a multidimensional structure.

Therefore, response dependence in a multi-conditional frame of reference should be tested from a different perspective although using the multi-facet model and following the principles of Rasch analysis. In the confectionery case the multi-faceted approach considered a sole persons' rating of an item and therefore, for the sake of computational simplicity, the procedure represented the responses of a person on the items replicated as though there were four different persons, providing unique identity numbers. Likewise, the procedure counted for 48 items even though they are just replications of a sole set of 12 items. However, under an actual perspective each person should have a unique parameter estimate and similarly, each item should have a unique estimate, without taking into account the number of replications.

5.3.1 Hypothesis

Drawing the argument from Section 5.3 into the discussion, it is worth investigating whether the scale obtained from the calibration in the confectionery case, which reasonably satisfied most of the procedures for testing fit, can be considered additive.

Therefore, it is hypothesized that the empirical conditions encompassing the four confectionery stimuli, which represent a finite set of members establishing a class of objects, would belong to the same frame of reference generated by the calibration of items using the multi-faceted Rasch model presented in Section 4.4.4.

5.3.2 An Alternative Perspective for Testing Response Dependence of a Multi-conditional Frame of Reference

To test the hypothesis established in Section 5.3.1 it is necessary to investigate the condition of response independence from different perspectives. One of the RM's assumptions is that a person location on the continuum is invariant over sub-structures (i.e., subtests) of the total measurement structure. However, if an item is replicated across Conditions *1* and *2*, say I_{11} and I_{12}, the person's location β_{11} associated with the I_{11}, and β_{12} associated with I_{12}, will not be the same if the conditions were distinctively different. By this means, a test for response dependence based on this one-response vector can give rise to interpretation of a lack of invariance.

A different perspective is to consider such cases in the view of a set of subtests each composed of a set of base-level items. Thus, if the shared features within items and across conditions have a significant impact on the validity of the local independence assumption amongst the base-level items, analysis at the subtest level may reduce that impact to non-significance.

Thus, the input of the raw scores (x_{ni}) of a person *n* on item *i* follow the same multi-faceted procedure, replicating items across conditions. After calibrating the structure, the original items replicated by the multi-faceted framework, originating one sub-item for each condition, are grouped in its primary form similar to subtests and such subtests are then re-parameterised (Table 5.3).

Table 5.3 - Frame of reference for subtests I_1 to I_i formed by conditions *1* to *S* under the alternative perspective.

Sub-items	I_{11}	I_{12}	…	I_{1S}	I_{21}	I_{22}	…	I_{2S}	I_{i1}	I_{i2}	…	I_{iS}
Item	I_1				I_2				I_i			
P_1	x_{111}	x_{112}	…	x_{11S}	x_{121}	x_{122}	…	x_{12S}	x_{1i1}	x_{1i2}	…	x_{1iS}
P_2	x_{211}	x_{212}	…	x_{21S}	x_{221}	x_{222}	…	x_{22S}	x_{2i1}	x_{2i2}	…	x_{2iS}
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$		$\vdots$
P_n	x_{n11}	x_{n12}	…	x_{n1S}	x_{n21}	x_{n22}	…	x_{n2S}	x_{ni1}	x_{ni2}	…	x_{niS}

Note: I_{11} represents item I_1 for condition *1* and I_{12} represents item I_1 for condition *2*, following similar representations for all items.

Item parameters and person parameters are in that case re-estimated because one of the algorithm's properties is that its statistic is a function of the frequencies of all response categories for estimating each threshold, rather than only a function of the frequency of the corresponding category (see Section 2.4.6). Accordingly, the estimate of the subtest will differ from the estimate using sub-items although its interpretation is similar. Letting the number of categories for the latter be *h*, then the number of thresholds will be *h-1*. For the estimate of the subtest, however, the number of thresholds will be equal to *k(h-1)*, where *k* is the number of conditions within the frame of reference. The location of an item on the continuum δ_i^* is obtained by the mean of the set of its uncentralised thresholds (RUMM2030, 2012) (see Note 16), such that

$$\delta_i^* = \frac{1}{(KM)}\sum_{k=1}^{KM}\tau_k \tag{5.8}$$

where *k* represents a condition within a frame of reference, such that $k = 1,2...,K$, *m* represents the thresholds of a sub-item, where $m = 1,2...,M$, and *i* represents the items in their subtest form, such that $i = 1,2,...,I$. Similar approaches for items associated with common structures and common content have been referred to in the literature as subtests (Andrich, 1985b), testlets (Wang et al., 2002), super-item (Cureton, 1965) or item bundles (Rosenbaum, 1988; Wilson and Adams, 1995). An overview of re-parameterisation to obtain the subtest locations expressed by Equation 5.8 can be found in Appendix A.

For example, let S_1, S_2, S_3 and S_4 represent the facet called stimuli (i.e., conditions) of a measurement structure. Let I_{11}, I_{12}, I_{13} and I_{14} be sub-items of the item I_1 for the stimuli S_1, S_2, S_3 and S_4, respectively. Let I_{11}, I_{12}, I_{13} and I_{14} be given ratings through a Likert-style four-category scale (e.g., strongly disagree (0), disagree (1), agree (2) and strongly agree (3)). To illustrate this example, take the estimate presented in Table 5.4 for the uncentralised item thresholds, represented graphically in Figure 5.1. The parameter for Item 1 is then obtained through Equation 5.8, where *M=(number of categories – 1)* and *K = 4*, such that

$$\delta_1^* = \frac{1}{(12)}\sum_{k=1}^{12}\tau_k$$

The condition facet parameter is computed by the mean of locations of the items associated with each condition, such that

$$\gamma_S = \frac{1}{I}\sum_{i_S=1}^{I_S}\delta_{i(S)}\text{, for } i = 1,2,\ldots,I \tag{5.9}$$

where *S* represents one condition enclosed in the frame of reference, *I* is the number of items and δ_i is an item location.

Table 5.4 – Illustrative computation of $\hat{\delta}^*$ through uncentralised thresholds.

Uncentralised thresholds (τ)		$\hat{\delta}_1^*$
τ_1	-5.57	-0.62
τ_2	-4.56	
τ_3	-3.65	
τ_4	-2.82	
τ_5	-2.03	
τ_6	-1.24	
τ_7	-0.44	
τ_8	0.42	
τ_9	1.36	
τ_{10}	2.42	
τ_{11}	3.63	
τ_{12}	5.01	

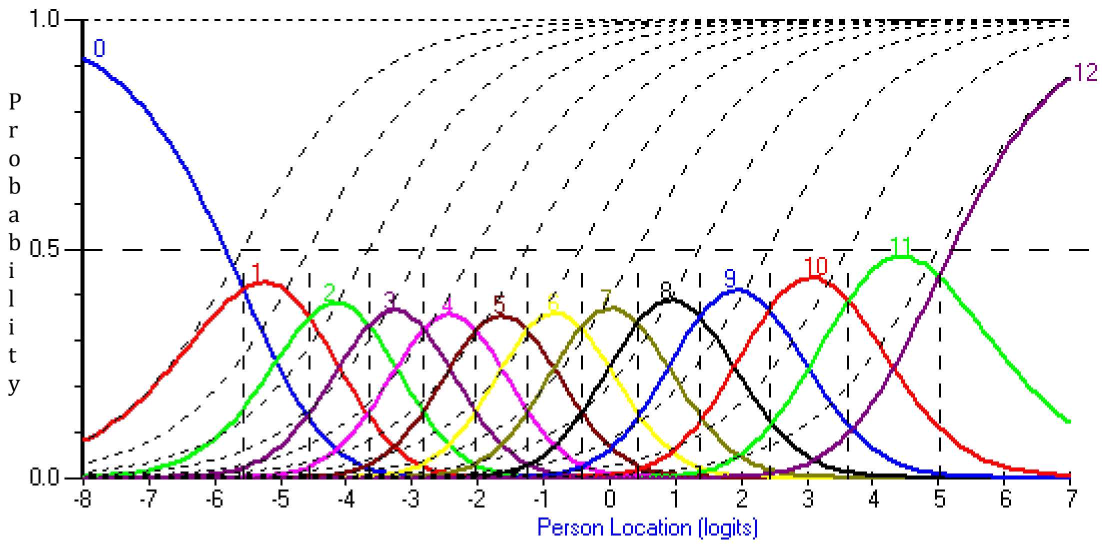


Figure 5.1 – Example of category characteristic curves (CCC) and threshold locations for a subtest formed by sub-items with responses to four conditions, belonging to an individual item with three thresholds.

5.4 EXAMINING THE RESIDUAL CORRELATIONS FROM THE PREVIOUS EMPIRICAL STUDY

5.4.1 Method

5.4.1.1 Testing the estimates of replications of items using the dependence effect d for statistical significance

The dependence effect d was obtained according to the procedure proposed by Andrich et al. (2012) (see Section 5.2) based on the person-item residual correlation matrix generated from the estimation of parameters of empirical study on confectioneries (see Section 4.4.4). The computation of the dependence effect d indicated whether there was an increase in probability of the response $X_{nj} = x_j$ of dependent item j as a result of the response of $X_{ni} = x_i$ on independent item i. Thus, the procedure focused on identifying and quantifying response dependence in terms of an increase in distance between thresholds. The procedure demanded the dependent item be split in as many sub-items as its categories. For example, in the confectionery case a category system was established with four options of responses (i.e., five categories collapsed into four after calibration); therefore, there were four sub-items split from the dependent Item j. However, those responses are statistically dependent on Item i. Thus, both Items j and i were deleted from the responses matrix, restructuring it with the four sub-items. The mean of the successive differences of each of the threshold pairs provided an estimate of d.

5.4.1.2 Test of the correlations between stimulus scores

A second approach for response dependence tested the correlation of the whole set of items for each stimulus (i.e., for each condition). Items were grouped in sub-tests. The term sub-test is used to represent a set of clustered items (Andrich, 1985b), which was in this case organised by stimulus (i.e., a condition representing a confectionery stimulus). The correlation was obtained through pairwise comparisons between persons' locations for the two subsets formed on the basis of two different stimuli.

Furthermore, reliability indices for subsets constituted of scores for each condition were compared against the reliability indices for the framework with items replicated across conditions.

5.4.1.3 Simulating applications using a multi-conditional frame of reference

To examine the effects of the shared features amongst items and amongst conditions a procedure was used to simulate applications of the multi-faceted approach. The aim was to determine whether the multi-facet framework would generate some kind of response dependence not only for the particular case of confectioneries used in Chapter 4 but also in most of the cases of the multi-faceted approach. This can demonstrate that response dependence is not a problem of the particular case of confectionery data, rather it is a characteristic associated with the structure of data in the multi-faceted approach.

To simulate such effects two conditions were established, which could, for example, be taken as different characteristics of a design component of a product. Different persons' mean locations were established for the conditions. The computer program RUMMss (Marais and Andrich, 2012) was used for simulating data. In the procedure a hypothetical sample of 500 persons was established with scores generated on the basis of five categories on 10 items with maximum score of four. Two person location means were established such that the persons mean location for Condition 1 was 0 logit and for Condition 2 was 1 logit with σ = 2 for both conditions. Trait dependence (i.e., multidimensionality) was induced in the structure by establishing four values of correlation (*r*) between the person locations for the two conditions. The *r* values of 1.00, 0.70, 0.40 and 0.20 were analysed independently although following identical procedures. For example, the correlation value of 1.00 assumed that the person locations between the two conditions were invariant. No response dependence between items was forced in the simulation. Therefore, neither a change of thresholds nor a change of the item difficulty was previously established for increasing or decreasing the probability of identical responses. Locations for all items were based on a specification of uniform distribution with random increments with minimum and maximum values of -4 and 4 respectively for each condition (i.e., Items 1 to 10 for condition 1 and Items 11 to 20 for condition 2).

After obtaining the simulated data, the person and items' parameters were estimated through the RUMM2030 computer program (Andrich et al., 2012). Based on those estimates a matrix of correlation for person-item residual was generated and examined.

5.4.1.4 Test of the residual correlations from the previous empirical study using the alternative framework

Test for violations of the assumption of response dependence when using a multi-conditional frame of reference at the item level may not address the behaviour of data satisfactorily (see Section 5.3). Thus, the data from the estimation of parameters of the study on confectioneries (see Chapter 4) were tested for response dependence using subtests formed according to Section 5.3.2. The person-item residual correlation matrix was examined to identify correlation indices ≥0.30 (see Section 2.4.7.9). Furthermore, reliability indices for subsets constituting of scores for each condition were compared against the reliability indices for the framework with items replicated across conditions.

The impact of a residual factor on the measurement system was indicated by the R-square of two subsets of items. Those two subsets were established by the groups of items with opposite signals obtained from PCA. The person locations on each subset of items were cross-plotted against the original locations taking the whole set of items (Linacre, 1998). Furthermore, the reliability index for subsets was compared against the overall reliability index for the framework with items replicated across conditions.

5.5 RESULTS

5.5.1 Approach Using the Dependence Effect *d*

A pair of items was chosen for each stimulus which held the highest index or one of the higher correlation indices of each stimulus in the correlation matrix of residuals. The items in the brackets in Table 5.5 (Column *item pair*) represent the pair chosen for each stimulus. The first number indicates the code of the dependent item on the second item, indicated by the second code. The Column *correlation* presents the correlation index found in the enlarged frame of reference, d is the mean value of the dependence effect for each threshold, $\hat{\sigma}_d$ is the square root of the sum of squares of the standard errors in each threshold (see Section 2).

The combination of items obtained from the correlation matrix of residuals of the enlarged frame of reference for the confectioneries indicated statistically significant response dependence at 5% level of the CI ($z > 1.96$) for all pairs.

Table 5.5 - Estimates of d from resolved item j associated with item i for four pairs of items obtained from the previous confectionery study with enlarged frame of reference.

Conditions	Item pair	Correlation	$\hat{d}_1$	$\hat{d}_2$	$\hat{d}_3$	$\hat{d}$	$\hat{\sigma}_d$	z
Stimulus 1	(3,8)	0.39	0.32	0.64	1.60	0.85	0.06	13.28
Stimulus 2	(22,21)	0.41	1.17	0.66	1.68	1.17	0.08	14.27
Stimulus 3	(29,25)	0.32	-3.30	0.51	0.73	-0.69	0.06	-12.55
Stimulus 4	(47,46)	0.37	-0.49	0.48	0.60	0.20	0.05	3.85

5.5.2 Pairwise Comparisons of the Correlation between Stimulus Scores

Pairwise comparisons were established by the correlation indices between persons' locations for the subsets formed by two different stimuli. Subtest *1* was formed by items I_{1i}, Subtest *2* by items I_{2i}, Subtest *3* by items I_{3i} and Subtest *4* by items I_{4i} where $i = 1,2,...,12$. The results indicated low correlation indices of person locations for all combinations (Table 5.6). The multidimensionality of the trait when adopting the general approach for identifying violations of the Rasch assumptions is further depicted in Figure 5.2. The dots represent the person locations in logits. The R-square values indicate the degree of correlation between person locations for two conditions, i.e., a pair of stimuli. Those low correlations imply that the person locations are not invariant over subtests. This perspective of analysis was established from the violation of the RM assumption of unidimensionality.

Table 5.6 – Pairwise comparisons of the correlations of person locations between stimuli used in the previous confectionery study.

	Stimulus 1	Stimulus 2	Stimulus 3	Stimulus 4
Stimulus 1	1			
Stimulus 2	0.47	1		
Stimulus 3	0.15	0.09	1	
Stimulus 4	0.03	-0.08	0.20	1

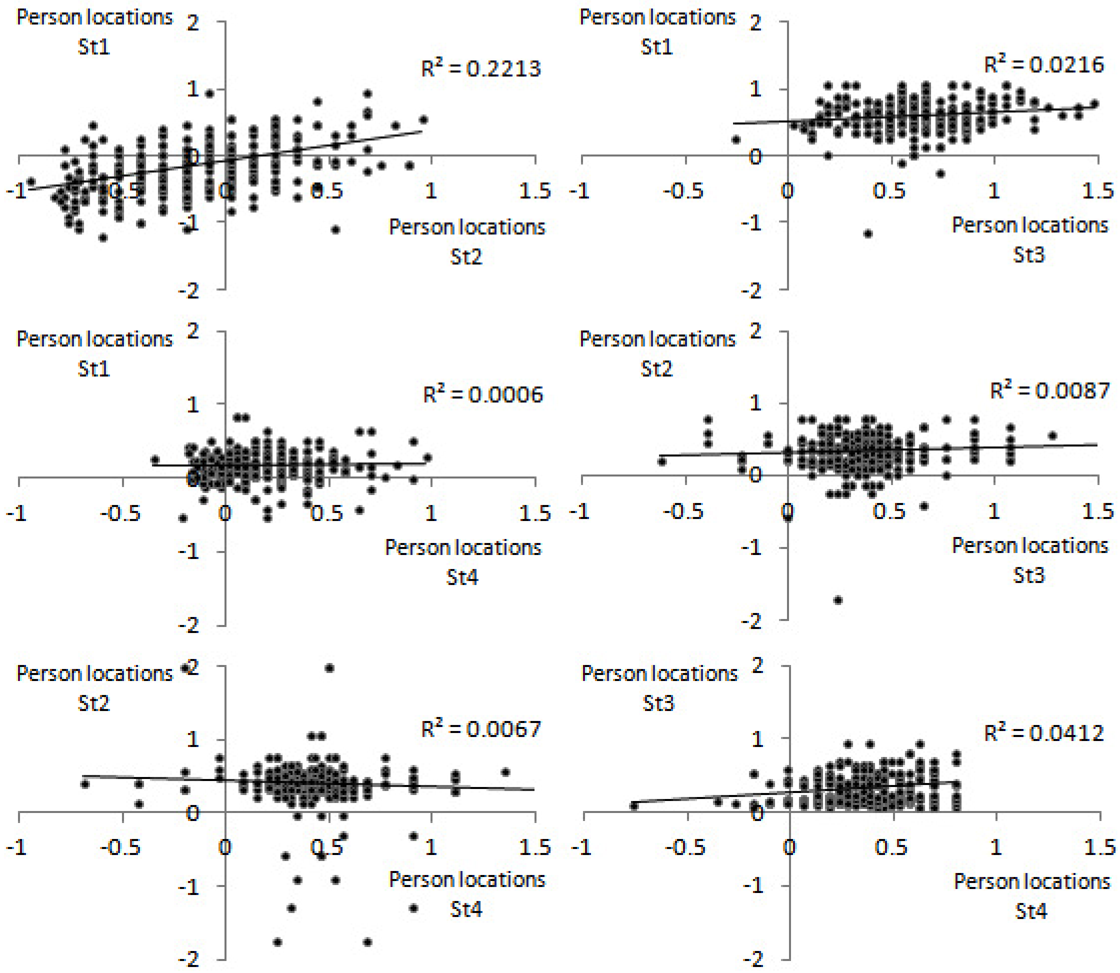


Figure 5.2 – Graphic representation of the trait dependence of the structure obtained from the previous confectionery study using pairwise comparisons of the correlations of person locations between stimuli.

The reliability indices held low values indicating a PSI of 0.34 and a coefficient α of 0.32 when comparing with the PSI of 0.87 and α of 0.87 for the framework without subtests. From the perspective of forming a framework of analysis using subsets constituted of items replicated for each stimulus, those indices suggested an important effect associated with dependence of responses of persons and with trait dependence.

5.5.3 Applications with Simulated Data

Results from the four simulations are presented in Table 5.7. The simulated value of the correlation coefficient *r* indicates the degree of multidimensionality introduced in the data. Figure 5.3 presents the cross-plots of the person locations regarding Component 1 and Component 2 for each simulation. The low values of *R*-square in Simulations 2, 3 and 4 (Figure 5.3 b, c and d, respectively) are similar to what was found in the confectioneries data set.

Table 5.7 – Simulation design and resultant *R*-square

Simulation	Persons mean	Persons mean	Simulated *r*	Resultant R^2
1	0	1	1.00	0.8349
2	0	1	0.70	0.3700
3	0	1	0.40	0.1320
4	0	1	0.20	0.0006

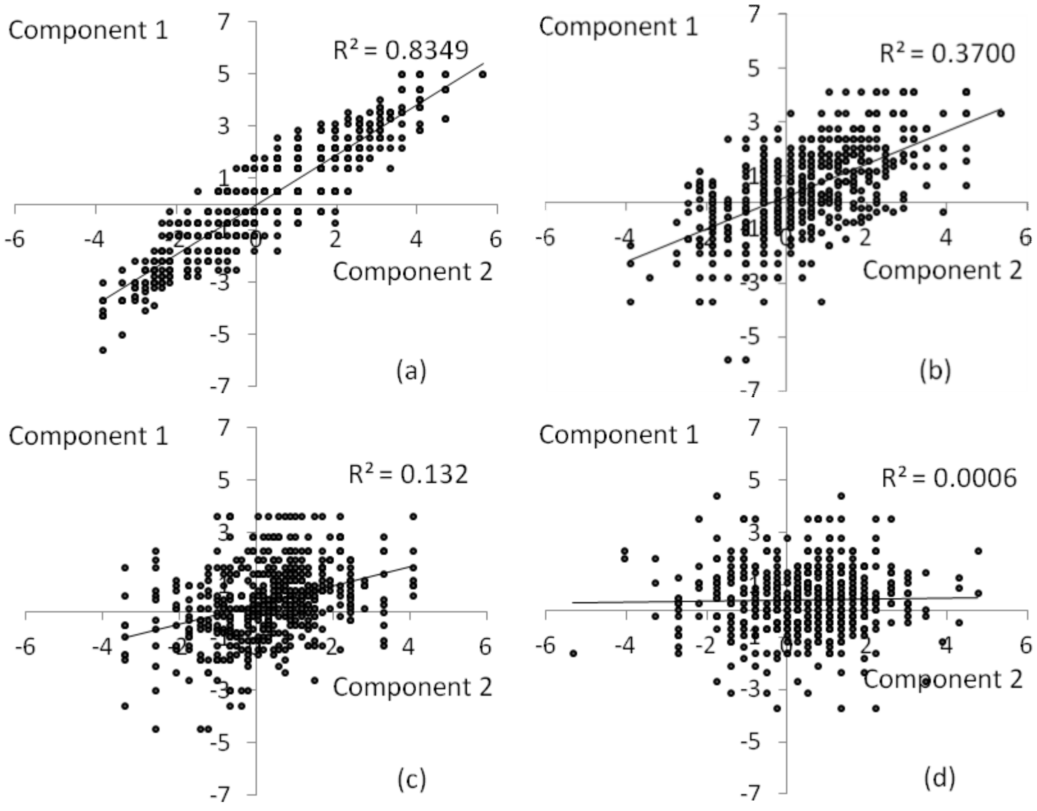


Figure 5.3 – Cross-plot of the person locations for the four simulated values of correlation, indicating (a) *r*=1.00, (b) *r*=0.70, (c) *r*=0.40 and (d) *r*=0.20.

However, the simulations showed that some degree of multidimensionality is expected in a data set with two components. This can be seen when establishing *r*=1.00 for the two components in the Simulation 1. This simulation aimed to indicate no statistical difference of the person locations between the two components. Nevertheless, the resultant value of *R*-square indicated some degree of multidimensionality although the effect of multidimensionality at this degree does not affect the fit-to-the-model significantly.

5.5.4 Approach Using the Alternative Framework

The 12 subtests formed by the combination of the sub-items of each item in the structure indicated correlations of person-item residuals lower than ±0.30 (Table 5.8). The highest residual correlation was found between ST05, formed by Item 5 for Conditions 1, 2, 3 and 4 and ST10, formed by Item 10 for Conditions 1, 2, 3 and 4 (Table 5.8, highlighted in italic).

Table 5.8 - Person-item residual correlations matrix under the alternative framework.

Subtest	ST01	ST02	ST03	ST04	ST05	ST06	ST07	ST08	ST09	ST10	ST11	ST12
ST01	1.00											
ST02	-0.11	1.00										
ST03	-0.06	-0.05	1.00									
ST04	-0.01	-0.12	-0.10	1.00								
ST05	-0.09	-0.15	-0.11	-0.10	1.00							
ST06	-0.15	-0.11	-0.21	-0.13	0.06	1.00						
ST07	-0.11	-0.18	-0.14	-0.21	-0.05	-0.05	1.00					
ST08	-0.05	-0.02	0.12	0.03	-0.17	-0.22	-0.09	1.00				
ST09	-0.13	-0.08	-0.10	-0.22	-0.14	-0.13	0.00	-0.01	1.00			
ST10	-0.13	0.13	-0.03	-0.12	*-0.23*	-0.14	-0.13	-0.06	**0.18**	1.00		
ST11	0.02	-0.20	0.00	0.00	-0.03	-0.07	-0.07	-0.17	-0.19	-0.18	1.00	
ST12	-0.19	-0.09	-0.17	0.00	-0.13	0.01	-0.03	-0.14	-0.06	-0.17	-0.16	1.00

However, the expected correlation among the residuals is negative. It is particularly important to identify positive correlations that are unusually high. The highest residual positive correlation was found between ST9, formed by Item 9 for Conditions 1, 2, 3 and 4 and ST10 (Table 5.8, highlighted in bold). Those low values for the correlation of person-item residual suggest the data did not violate the Rasch assumption of response dependence.

Only anomalies that can affect the empirical meaning or use of the measures are of concern. This was empirically investigated by cross-plotting the person locations based on a group of items with positive loadings (Figure 5.4) and another group of items with negative loadings (Figure 5.5), both obtained from PCA of the residuals against the original person locations taking the overall set of items (Linacre, 1998). The R-square for the positive and for the negative loadings subsets of approximately 0.88 and 0.87 respectively when comparing with the overall set of calibrated items suggest the differences between person locations could originate from measurement error and not from trait dependence.

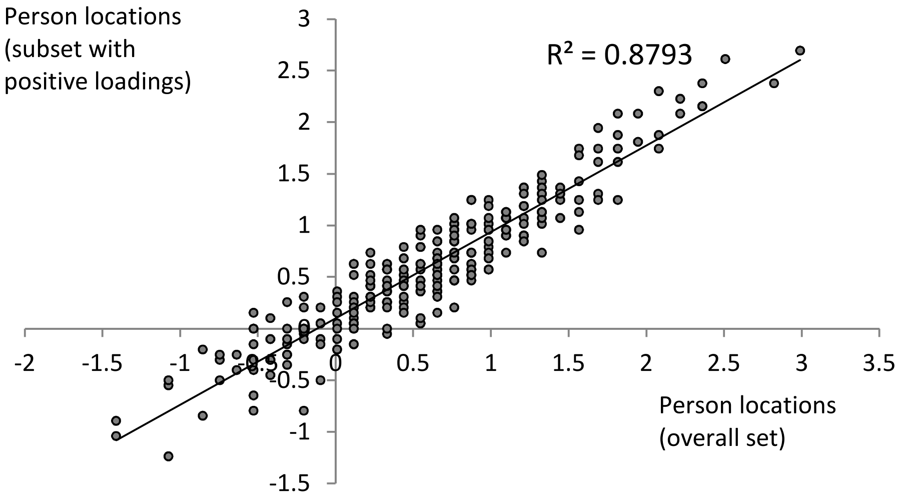


Figure 5.4 - Cross-plot of the person locations based on a group of items with positive loadings in comparison with the original person locations taking the overall set of items.

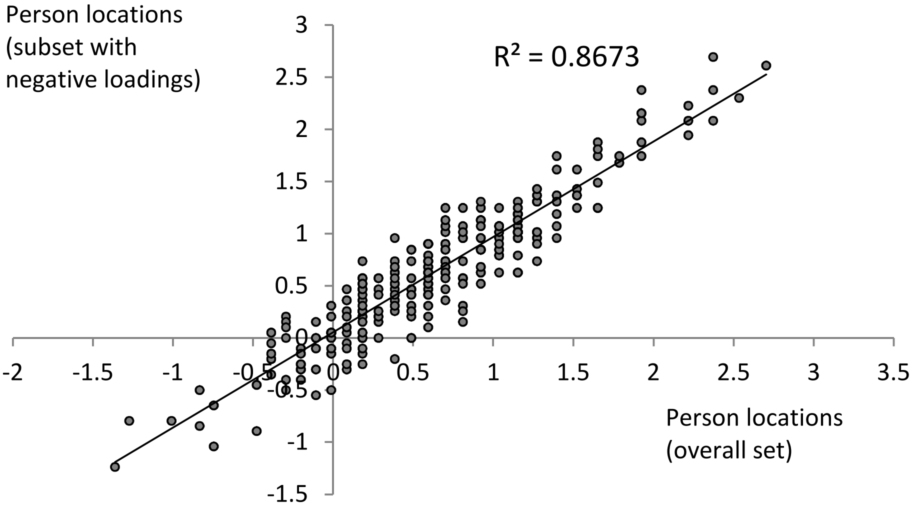


Figure 5.5 - Cross-plot of the person locations based on a group of items with negative loadings in comparison with the original person locations taking the overall set of items.

The PSI and coefficient α for subsets constituted of items replicated for each condition of 0.88 for both indices were slightly higher when comparing against the PSI and α of 0.87 for both indices for the framework without subtests.

The quality of fit was also examined through the fit statistics for the subtests under the alternative framework. Table 5.9 shows that the fit of the specific subtests are adequate (p >0.05). Although Subtest ST11 and ST12 presented just marginal fit (p =0.05 and p =0.08 respectively), their residual was lower than ±2.50 (see Column *Fit residual*).

5.6 REMARKS

The term *level* of a facet has been used by many authors to identify different degrees of a latent variable, such as the severity of judgement. In this thesis, however, the term *level*

Table 5.9 – Subtest fit statistics for the previous confectionery study under the alternative framework.

Subtest	Location	*SE*	Fit residual	χ^2	Degree of freedom (*df*)	*p*
ST05	-0.49	0.04	0.59	2.49	4	0.65
ST07	-0.43	0.04	0.36	4.68	4	0.32
ST09	-0.42	0.05	-0.57	2.92	4	0.57
ST11	-0.37	0.05	2.11	9.43	4	0.05
ST12	-0.06	0.05	1.31	8.33	4	0.08
ST08	-0.02	0.05	-1.53	6.63	4	0.16
ST06	-0.02	0.05	1.11	1.62	4	0.81
ST01	0.20	0.05	0.28	1.37	4	0.85
ST04	0.25	0.05	0.54	5.62	4	0.23
ST10	0.33	0.05	-0.86	4.96	4	0.29
ST03	0.44	0.05	-0.83	4.22	4	0.38
ST02	0.58	0.05	-0.25	4.36	4	0.36
Overall fit				53.63	48	0.18

is substituted for the term *condition*, denoting not only degrees of the latent variable but also different characteristics of the facet related to the design elements associated with the attribute being measured. The term *stimulus*, which has hitherto been used in this thesis, represents a particular condition, which does not necessarily represent degrees of a characteristic.

In Figure 5.1 an example of a subtest is given with a format in which the categories are intended to reflect order. However, when using subtests a different structure is formed without requiring ordered thresholds. The interpretation of the threshold estimates is different from those associated with a typical polytomous item composed of ordered categories. Andrich (2006) has stated that the more local dependence is accounted for with the subtest, the more the thresholds will be disordered. When subtests are formed from a set of dependent items, the thresholds might be disordered. This effect follows because the dependent items within a subtest will yield more extreme scores (i.e., closer to zero and the maximum on the subtest), for any person location. Andrich (1985b) has stressed that the difference in item locations of the subtest will trade off with their local dependence.

5.6.1 Concluding Comments

In many cases that use the faceted model, the overall fit can cover signals of response dependence in the data. Such an item response analysis tends to overestimate the precision of measures obtained from subtests and yields biased estimation for item difficulty and discrimination parameters. Such a problem can be solved when disclosing misfit using a subtest protocol and an investigation of the item-person residual correlation.

To meet the assumptions of a quantitative structure, the persons’ locations ought to be invariant over subtests. However, variance of persons’ locations might not be a characteristic of the data. In a number of cases the faceted framework might force a multi-dimensional structure and response dependence. Therefore, it seems, the most realistic approach to identify response dependence when using the faceted model is to group an item replicated across different conditions in a subtest structure. The subtests are not formed by the facet conditions but by the sub-items originated from individual items.

In this chapter empirical evidence has been provided for a novel contribution that can support the test of the assumption of local dependence in a multi-conditional frame of reference. This statement is based on the confirmation of the hypothesis that the four confectionery stimuli belong to the same frame of reference generated by the calibration of items using the multi-faceted Rasch model. The representation of items as subtests formed by the responses to their sub-items is associated with each condition in an enlarged frame of reference. Nevertheless, the novel approach does not modify the original structure of the data. This means the approach does not obscure signals of local dependence if they do exist in the data (see Appendix B).

CHAPTER 6

Stability of a Rasch-based metric for affective responses to product features

Quantifying affective response to products demands that measurement stability be examined. This chapter is concerned with an empirical approach to evaluate a scale calibrated on the fundamentals of Rasch measurement theory for affective responses to product containers. Respondents, statements and stimulus products were parameterised independently using the faceted Rasch model for establishing a measurement structure. Affective responses were compared with the physical properties related to compliance of the containers. Empirical consistency of the measurement was assessed by the replication of a calibrated pool of statements across two different samples. Furthermore, a cross-validation strategy compared calibrations using different groups of respondents. The results indicated that the differences between samples and between calibrations are statistically non-significant. As an implication of the metric's stability it is possible to add statements to the calibrated core of measurement without loss of comparability. Therefore, the results support the proposed role of Rasch theory to refine and to generalise measurement structures of affective features of products[32].

6.1 ASSESSING STABILITY OF A MEASUREMENT STRUCTURE

Typically, reliability has been associated with variance of persons' responses and to variance on account of persons and items interaction. That is, reliability is associated with stability or consistency of scores over time or across individuals. Additionally, the internal consistency of scales is usually indicated by reliability indices. The internal consistency is an estimate of the degree to which sets of scores between items are associated. Rasch analysis provides two indices of reliability, the person separation index (PSI) and Cronbach's α [33] (see Section 2.4.7.10).

However, it has been demonstrated that a construct could obtain high internal consistency value even when the items measure several dimensions, i.e., unrelated attributes (Cortina, 1993; Green et al., 1977; Cronbach, 1951). Cronbach's α could also be influenced by the number of items in a structure (Cortina, 2003). In addition, computation

[32] Publications based on this chapter can be found in Camargo and Henson (2012d).

[33] Cronbach's alpha is provided if there is no missing data.

of the reliability index using classical theory is influenced by the characteristics of the sample (Traub, 1994). A formal analysis based on more recent psychometric theory has been considered an effective solution (Wright and Stone, 1979; Embretson and Reise, 2000). Different approaches have referenced item analysis as a mechanism to achieve consistency in a measurement instrument. This analysis resides in computing person ability and item difficulty parameters independently.

Accordingly, this study is concerned with providing evidence of stability given by the empirical consistency of a scale for measuring the affective attribute. Empirical consistency is a fundamental examination for any concept of validity of measurement (Messik, 1989; Traub and Rowley, 1991). Additionally, a cross-validation strategy compared different items calibrations using different groups of respondents. This demonstrated whether the calibrated item sets originated from different calibrations measured the same characteristic of the stimulus products.

6.2 HYPOTHESES OF THE EMPIRICAL APPROACH

Taking into account that one of the benefits of the model is its independence with regard to sample distribution, it is hypothesized that measurement of affective responses to design elements through calibrated structures using the RM does not vary within a same context even if different groups of people are used (see Section 1.6.1). If this hypothesis is demonstrated to be true, then the measurement structure of affective responses could be generalised.

It is also hypothesized that if the calibrated set of items can be kept as a mainstay of the measurement structure, then further items could be calibrated and accommodated into the structure to convey further information about the affective attribute (see Section 1.6.1).

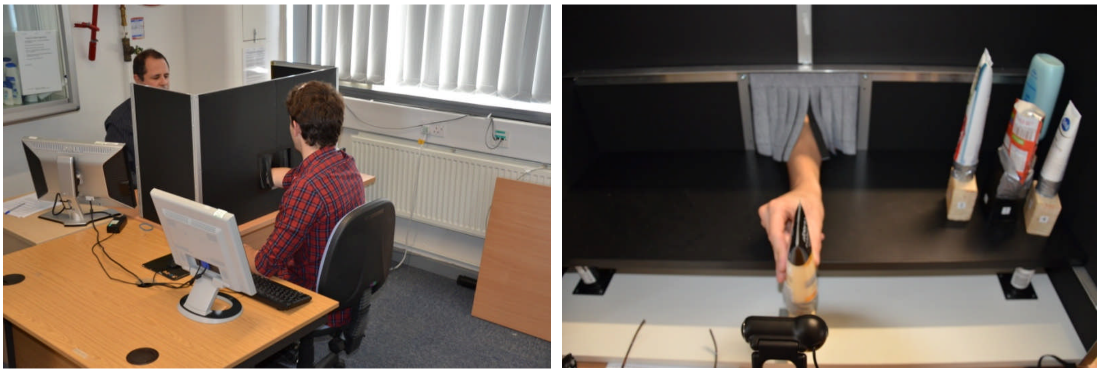
6.3 AIMS

The empirical approach aimed to establish a common metric for comparing affective responses to different packaging characteristics through tactile impression when squeezing containers of everyday products. An empirical investigation was designed to examine whether after distinct calibration using the RM the scale would yield stable results.

Additionally, the study aimed to collect sensory information from the tactile interaction when squeezing the containers. The reason was that the outcomes from this study might indicate a further investigation to determine whether there is a correlation of the calibrated metric for affective responses with the force applied by persons on containers that hold different characteristics of compliance.

6.4 METHOD

The scope of the experiment was to measure the relative importance of the packaging material for obtaining an intuitive impression of a face moisturizer cream as a product feature, denoted *perceptiveness* henceforth. The term *perceptiveness* was, in this study, used to name the affective dimension being measured and attributed to persons' feelings about the products when squeezing their containers. Participants were not, therefore, required to understand the meaning of term. Five products with different characteristics related to compliance of their containers were used to provide a variety of values of physical properties that were of interest to the study. Participants were neither able to see the containers nor able to make contact with the product inside them. After squeezing each container (Figure 6.1) participants rated their endorsement on a five-point Likert-style scale to statements related to *perceptiveness* of the product using computer-base self-report questionnaires. Additionally, participants were asked to squeeze the containers once more wearing tactile sensors on their fingers to measure the force applied on the containers.



(a) (b)

Figure 6.1 - Experiment lay-out (a) and participants touching a stimulus through a visual barrier (b).

One hundred and ninety two participants took part in the whole experiment, which was split into three stages. The reason for this strategy was to facilitate the test of the hypotheses. The first stage was concerned with collecting words and statements. The second stage aimed to obtain affective and sensory responses to a preliminary set of items and subsequently, to calibrate the measurement structure. Finally, the third stage provided data from affective responses to the stimuli based on a same metric with calibrated set of items using a different group of respondents. Results were then compared against those from the second stage. Ethical approval for this empirical study was obtained from the University of Leeds Research Ethics Committee (ethics reference number MEEC 10-032).

6.4.1 Statements Collection

The first step was to establish a preliminary pool of statements to capture participants' responses to a set of stimuli related to the context of everyday products packaging. These statements, such as *I feel the product in this packaging could be sticky*, emerged from words and statements collected by qualitative research using a focus group, which documented verbatim statements to express affective requirements (Henson, 2006; Barnes and Lillford, 2009). Other statements originated from publicly available online consumers' reviews, manufacturers' catalogues and advertisements to some everyday goods such as personal care, food, healthcare, cosmetic and household products.

6.4.1.1 Focus group

A video-recorded one-hour session was carried out with six volunteers, two females and four males, age from 25 to 36 years old. Participants received £5.00 as a compensation for taking part in the study.

After handling product containers, participants expressed their impression of the product inside the containers and of the containers themselves. They did not make any contact with the products inside the containers. To keep the session on track while allowing participants to talk freely and spontaneously, the investigator used a discussion protocol containing four timed exercises as a guide to stimulate and inspire participants to cover as much as possible the facets related to products' containers while touching them.

In the first exercise, the investigator split the group into subgroups of two. He presented a set of nine everyday product containers for each subgroup. Participants were asked to handle the containers, feel them and picture them in their everyday activities. After a while the investigator asked participants to say how the containers feel and how they would describe the containers in their own words.

In the second exercise, participants were asked to choose a favourite product's container out of all the items. They were asked to consider the container regardless of its functionality but based on the containers' feel as much as possible. Participants were told to write down a list of adjectives that would better describe what the container feels like.

The third exercise consisted of giving three containers to each participant. They were asked to pair containers that they thought to be similar and to leave the other one out. Then, the investigator asked them to say why they organised the containers in that way.

In the last exercise, the investigator presented five adjectives written on a piece of paper and displayed on the table. The adjectives were established as follows: *perceptible*, *boring*, *dense*, *confident* and *runny*. These adjectives were chosen because they have frequently arisen in customer online product reviews although they do not seem to have a clear definition amongst consumers. Participants were then asked to indicate which container would fit best to the adjectives given to them and tell why they had chosen that match. When different opinions arose, the investigator prompted a debate between participants.

6.4.1.2 Preliminary pool of statements

Analysis of the transcription of statements and words originating from the focus group as well as from the other sources previously mentioned, consisted of clustering them according to their characteristics. These characteristics were elected as those of functionality and usability, graphics and form, and tactile perception.

Sixteen clustered statements that represented affective requirements related to or based on characteristics of tactile perception were included in the original pool of items (Table 6.1).

Table 6.1 – Preliminary pool of items and Items codification.

Code	Items
I1	The product in this container would give me a heavy, greasy film on my skin.
I2	The product in this container is likely to look and smell delightful.
I3	I might get a bit watery product in this container.
I4	I feel the product in this container would hydrate my skin.
I5	The product in this packaging might be pricey.
I6	The container feels only half filled when squeezing it.
I7	The container makes me feel like I would be buying a great product.
I8	The product inside the container would spread easily.
I9	There is a lightweight cream in this container.
I10	It is easy to know how much is left in the packaging.
I11	The product inside this container could be sticky.
I12	The product in this packaging is likely to flow easily.
I13	The product in this packaging might seem more medicinal than anything else.
I14	It is quite hard to explain the product when touching its packaging.
I15	The product in this container could give me a refreshing sensation.
I16	The product in this packaging could be a bit boring.

6.4.2 Sampling

6.4.2.1 Stimuli

Five everyday products with different characteristics related to compliance of their containers were used as stimuli and presented to the participants (Figure 6.2). Products commercially available were selected according to capacity, dimensions, proportions, packaging material and characteristics of the content (Table 6.2). The containers were classified as: cylinder tube, oval bottle, downward taper tube and gusseted pouch. The stimuli were mounted with their closure firmly fixed on wooden blocks of dimensions

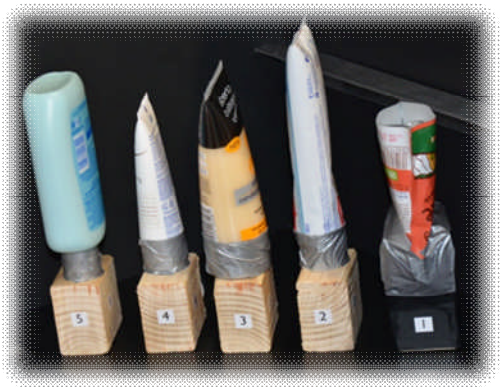


Figure 6.2 - Products used as stimuli in the empirical study.

Table 6.2 – Material and compliance of stimulus objects.

Packaging	Code	Product	Material	Compliance (3N)
Stimulus 1	ST1	Baby food	Polypropylene/aluminium/polyethylene laminated, gusseted, squeeze pouch with a polypropylene screw-top closure, capacity 130 ml.	6.08 mm
Stimulus 2	ST2	Toothpaste	Polyethylene/aluminium/polyethylene laminated tube with a polypropylene flip-top closure, capacity 150 ml.	5.70 mm
Stimulus 3	ST3	Hair conditioner	Low density polyethylene squeeze tube with a polypropylene flip-top closure, capacity 200ml.	4.74 mm
Stimulus 4	ST4	Moisturizer	Multi-layer low density polyethylene and ethyl-vinyl-alcohol squeeze tube with a polypropylene flip-top closure, capacity 75ml.	4.11 mm
Stimulus 5	ST5	Baby bath lotion	Oval, flat based, multi-layer high density polyethylene and ethyl-vinyl-alcohol bottle with a polypropylene flip top closure, capacity 300ml.	1.02 mm

100mm x 100mm x 40mm with a magnet attached to one of the sides for easy switching. The blocks were designed such that participants could easily find the container and settle their thumb, index finger and middle finger on it, and to ensure uniform positioning of the fingers for every participant. Participants were also instructed to touch the stimuli in the way that was most natural to them when squeezing ordinary containers.

For the compliance measurement a testing system was used that consisted of a force platform (MiniDyn, multi-component dynamometer Type 9256C2, Kistler), an X–Z motion table (Series 1000 Cross Roller, Motion link), a steel ball of radius 10mm, a controller and a computer. The containers were positioned between the steel ball and the force platform. The ball was pressed against the surface of each stimulus and the ball's displacement D_y with increasing load F_y was recorded. The measure of compliance was empirically taken to be the value of D_y (mm) when F_y was 3N (Chen et al., 2009; Shao et al., 2010). Measurements were repeated four times for every stimulus, taking their average as the final value of D_y.

6.4.2.2 *Sample size*

The size of the sample for Rasch analysis was considered upon two aspects. The first aspect took into account the number of participants for calibrating the scale. The second aspect was that in the third stage of the experiment a smaller sample was required than that of the previous stage because the experiment used the calibrated measurement structure.

Despite the Rasch characteristic of being independent of sample distribution, slightly different results were expected. However, these differences should be smaller as the sample size increases. Thus, a question arose about what would be similar enough in this study for a cross-validation.

According to Linacre (1994b), in Rasch analysis the stability of items calibration is related to its modelled standard error (SE) and item calibration with random deviations up to 0.50 logit are "*for all practical purposes free from bias*." (Linacre, 1994b). For constructs with few items (less than 30 items), the sample size is estimated using Equation 6.1 (Wright and Stone, 1979),

$$4/SE^2 \leq n \leq 9/SE^2 \tag{6.1}$$

where n is the sample size and SE is the confidence interval of errors. In this study a two-tailed 95% confidence interval of ±2.00 and for ± 0.50 logit interval is assumed. This gives a minimum sample size in the range of

$$\frac{4}{\left(\frac{0.50}{2.00}\right)^2} \le n \le \frac{9}{\left(\frac{0.50}{2.00}\right)^2}$$

thus, $64 \le n \le 144$

Taking into account that in the second stage of the experiment the scale had not been calibrated, for a sample with moderately off-target observations and within of 95% confidence that items are less than 0.50 logit from their stable value, a sample between 100 and 144 participants would be enough for stable results. For the third stage of the experiment the items that have already been calibrated were taken into account. In this case, participants had been reasonably targeted and a sample of 64 participants would be considered sufficient to produce the expected *SE* (Linacre, 1994b).

In the third phase of the experiment, the analysis based on classical theory was slightly different. The analysis was concerned with the difference between items location of two groups (from Stage 2 and Stage 3). The sample size calculation is based on Cohen's Case 1 for comparison between means with sample of different sizes ($n_2 \neq n_3$). Taking the sample of the second stage as n_2 = 120, it is possible to estimate n_3 based on the Equation 6.2 (Cohen, 1988, p.59),

$$n_u = \frac{n_f(n)}{2(n_f) - n} \tag{6.2}$$

where n_u is the estimated sample, $n_f = n_2$ and n is a basis for the number of sampling units. Let the effect size d be 0.40 and take Equation 6.3 (Cohen, 1988, p.62)

$$d = \frac{|m_2 - m_3|}{\sigma} \tag{6.3}$$

where $|m_2 - m_3|$ is the maximum difference between mean locations of two independent items in logits, which was estimated equal to 0.20 and $\sigma = SE = 0.50$. Taking α = 0.05 (non-directional) and power of 0.70, then n = 78 obtained from Table 2.4.1 in Cohen (1988, p.55). Thus,

$$n_u = \frac{120(78)}{2(120) - 78} \approx 58$$

Errors due to imperfections of the measurement methods themselves are smaller as the sample size increases. However, errors also vary according to the method chosen for a determined analysis and its expected effect size. Taking into account three different methods used in the analyses (i.e., Rasch analysis, analysis of variance and difference between means), the size of the sample was adopted as the highest value found in the estimates.

Nevertheless, the decision on the sample size remained quite arbitrary although within of parameters. This study adopted a minimum sample size of 120 participants for the second stage of the experiment and a minimum sample size of 64 participants for the third stage, in which addresses an effect size of 0.40 $\left(|m_2 - m_3| = 0.2\right)$ regarding differences between locations of the same items for two different samples. Both of the sample size estimates were large enough for classical approach as well as for Rasch analysis.

6.4.3 Data Collection from Persons' Affective Responses

One hundred and twenty volunteers took part in the second stage of the experiment for calibrating the scale, 65% male and 35% female, 51.6% between the age of 18 and 25, 31.7% between the age of 26 and 35, and 16.7% over 35. Participants received £5.00 as a compensation for taking part in the study. This value has been a typical compensation in previous similar experiments. The sessions took place in the Engineering Systems Laboratory at the fifth floor of the School of Mechanical Engineering, University of Leeds, Leeds, UK, in September and October 2011.

Data from participants' affective responses were obtained from ratings collected by computer-based self-report questionnaires. The questionnaires were designed on software using Microsoft® Visual Basic® (Figure 6.3).

The order in which participants were required to consider the containers was determined using a counterbalanced design. The order of the statements on the questionnaires was randomised. The software control panel is exclusively accessed by password. Responses code, list of participants and list of items were also protected by password. Participants' responses in the experiments were recorded against a randomly allocated participant number so that, although the researcher administering the experiment knew individuals' responses, others are not be able to match results with individuals' identities and personal data.

Before commencing the study, participants were informed about how long the experiment would last and how results would be used. Participants were also informed that they could withdraw from the activity any time without giving a reason and that their anonymity would be protected.

Contact details for the purposes of coordinating the attendance of participants at the study were stored in the University, on the University's computer systems, under password protection. The contact details of the participants were destroyed immediately after the study, unless participants gave consent for the details to be retained for the purposes of inviting them to participate in future studies. These data handling procedures are in accordance with the Data Protection Act 1998.

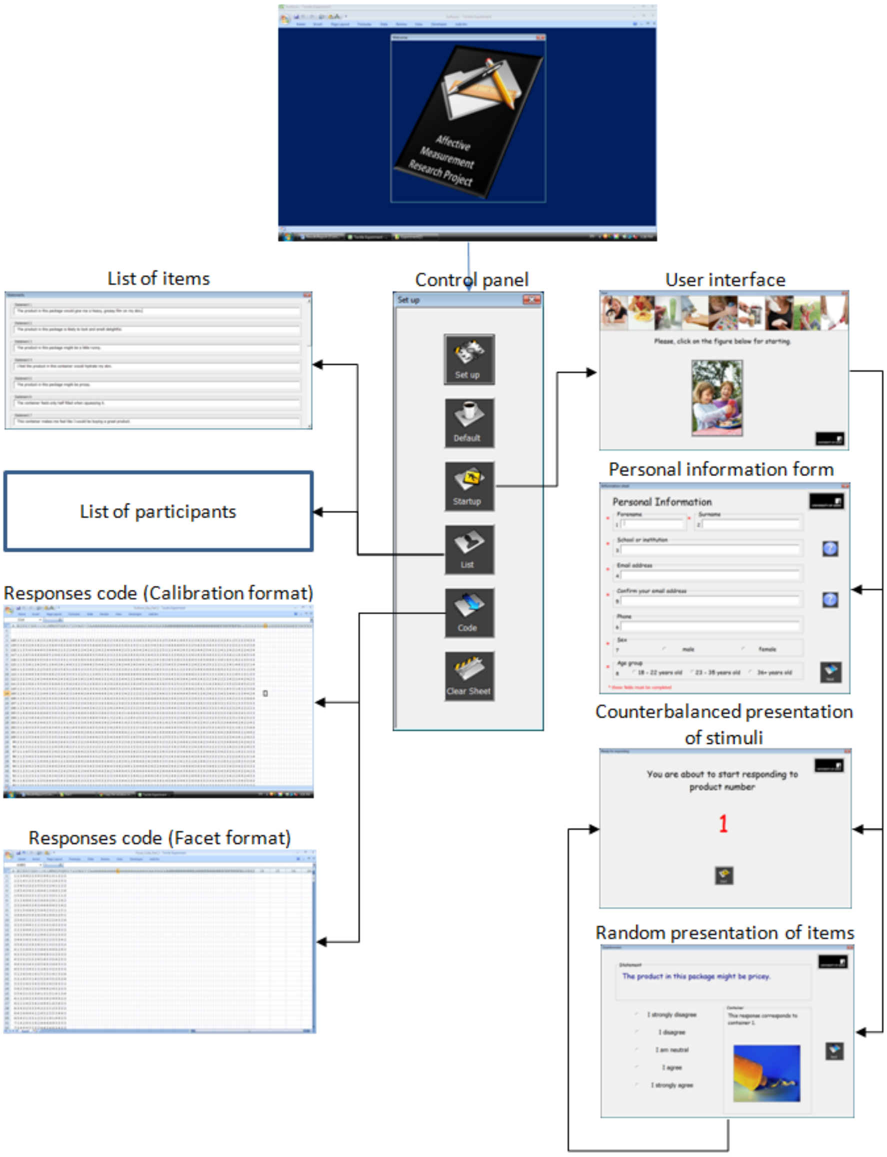

Affective Measurement Research Project
List of items
Control panel
User interface
Set up
Default
Startup
List
Code
Clear Sheet
Please, click on the figure below for starting
List of participants
Personal information form
Personal Information
Forename
Surname
School or institution
Email address
Confirm your email address
Phone
Sex
male
female
Age group
18 - 22 years old
23 - 28 years old
36+ years old
Responses code (Calibration format)
Counterbalanced presentation of stimuli
You are about to start responding to product number
1
Responses code (Facet format)
Random presentation of items
The product in this package might be pricey.
I strongly disagree
I disagree
I am neutral
I agree
I strongly agree

Figure 6.3 – Computer-based questionnaires framework.

6.4.4 Rasch Analysis

Rasch analysis was undertaken using the software package RUMM2030®, licensed version (2011). Items were calibrated for each stimulus independently. Subsequently, items and stimuli were compared on the same continuum through the faceted Rasch approach.

The calibration of items as well as the cut-off points adopted in this empirical study followed the procedures in a similar fashion of those established for the first study (see

Section 2.4.7). These included a verification of the scores system, tests of fit, analysis of the item-person interaction, analysis of the persons' response pattern and test for differential item functioning. Also, the data were tested to determine whether they had met the assumptions of response independence and unidimensionality.

6.4.5 Faceted Rasch Approach

6.4.5.1 Facet input

After individual calibration has been carried out for each stimulus, items were re-scored and adjusted for a set of co-calibrated items for all of the stimuli. The individual data sets that derived the co-calibrated scales were then used as input for the facet approach (see Chapter 4).

6.4.5.2 Stimuli separation

Stimuli separation is a particular test to verify between-stimuli heterogeneity (see Section 4.4). Two stimulus conditions were identified from pairwise comparisons. The functional stimuli condition identified whether all of the stimuli were working well together and whether the whole set of information could be used.

6.4.5.3 Differential stimuli functioning

DSF was tested through pairwise comparisons between SCCs (see Section 4.4.2). The first case indicated by the test was whether the comparison between two stimuli presented uniform separation, i.e., whether the stimuli used in the experiment hold distinguishable features. The second case indicated by the test was that of non-uniform separation between stimuli, i.e., when a SCC crosses another one.

Both of the cases used the approach for detecting DSF, in which the area between SCCs is computed. The area was obtained when integrating the difference of the polynomial equations that defined two SCCs. The reference points for uniform separation were taken as the minimum and maximum person location amongst all of the SCCs. For the case of non-uniform separation, the intermediate point was the intersection of two SCCs projected on the logit location axis.

The cut-off criterion of differentiation involving two SCCs was obtained through comparison of the area computed between curves and an imparity criterion. The value of the imparity criterion for each pairwise comparison was calculated by the sum of areas enclosed between two SCCs and the SCCs obtained from the same cumulative functions plus its standard errors for the ability levels.

6.4.6 Intra-class Consistency and Cross-validation of the Calibrated Metric

In order to evaluate the measurement structure, data was obtained from affective responses to the same set of stimuli used for calibrating the scale. A different group of 66 participants took part in this third stage of the study, 28.8% females and 71.2% males, 78.7% with age in the range of 18 to 25, 15.2% with age in the range of 26 to 35, and 6.1% over 35.

The empirical consistency of the scale was verified through an intra-class correlation within the framework of ANOVA for each item and for every stimulus using two different samples. The first sample was used to calibrate the measurement structure and the second sample, from the second data collection, was used to verify the scale's consistency when compared with the first group. An independent *t*-test assessed the stability of the scores comparing locations on the scale continuum between samples.

A stratified *k*-fold cross-validation consisted of splitting the whole sample of 186 participants into four groups, making k = 4, arranged in two groups of 47 individuals and two groups of 46 individuals. The sample was stratified by sex and age-group. Thus, groups had similar proportions of those for the previous calibration of items, called *datum calibration*. According to the stratification, respondents were randomly chosen for each group. The statements were established as the same of the datum calibration; however, without anchoring item locations. The item set was then calibrated by selecting three groups forming a new sample for each run, leaving one group out, denoted *n*-calibration, where n is the identification of the group combination used for calibration. Differences between calibrations were calculated using a *t*-test, which indicated how many of the *t*-tests were significant at 5% level of observations.

6.4.7 Incorporating Further Statements to the Calibrated Items Set

In the second data collection five items were incorporated to the scale of *perceptiveness* (Table 6.3). If a scale is calibrated, then it should be possible to incorporate a new set of non-calibrated items and test them to determine whether they will work well. Thus, the calibrated items originating from the first administration were used as the core of a second calibration with the incorporated items. Those additional statements were obtained from publicly available online consumers' reviews and elected as characteristics of tactile perception. However, the incorporated items were not taken into account when determining the empirical consistency of the scale and its cross-validation.

Responses to the stimuli were analysed based on data including the five additional items in the calibrated scale obtained from the prior stage. Unidimensionality, response dependency, items and persons misfit, DIF and disordered thresholds were once more tested against the same indices used in the previous stage.

Table 6.3 – Items incorporated to the calibrated metric and tested according to Rasch analysis.

Code	Items
I17	I could get just the right amount of the product when I squeeze its container.
I18	I've got a pleasant touch with this container.
I19	I feel this container as a skin care product.
I20	It's too soft for a creamy product.
I21	I could find no consistency in the product inside this container.

6.4.8 Sensory Data Collection

After squeezing all of the products and responding to the questionnaires, participants were required to squeeze the containers once again while wearing tactile sensors on their fingers. The reason for wearing sensors on fingers was to capture the force applied on the containers when squeezing the products for each individual. The objective was achieved through a technical tactile device and its software (FingerTPS^{TM} - Wireless Tactile Force Measurement System). The system provided and recorded time-series, force average and maximum force. Wireless Bluetooth connectivity was used between an interface module and a computer. The procedure consisted of fitting the cable harness and the sensors of the system on participants' wrist and fingers respectively (Figure 6.4). Subsequently, the system was calibrated for every participant using a reference sensor at a force of 13.35N (Figure 6.5a). Because this experiment is related with tactile perceptions, the same visual barrier when collecting affective responses was placed between the participant and the stimulus to avoid visual contact (Figure 6.5b). Participants were told to force the containers

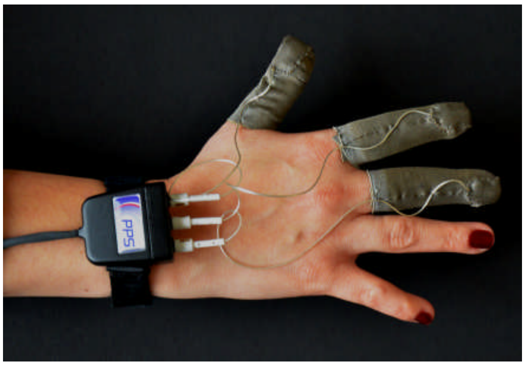
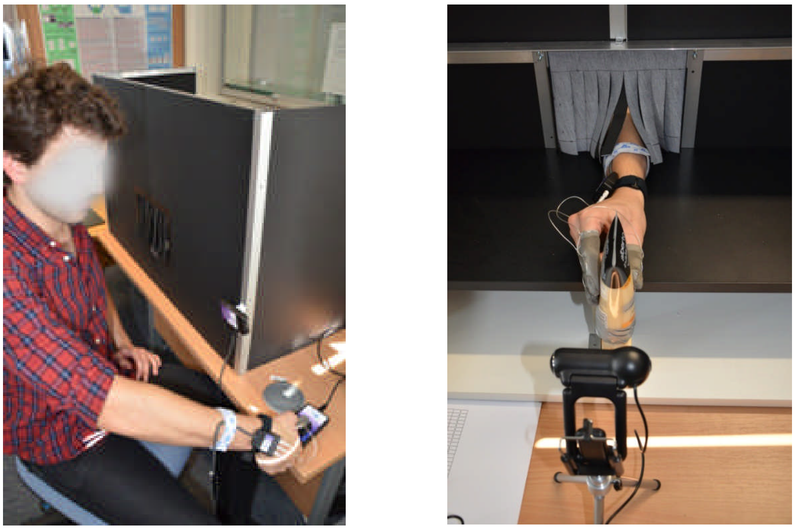


Figure 6.4 - Tactile sensors used in the experiment.

as though they were to squeeze the product out just enough for obtaining a feeling about it. Participants squeezed each container twice in an interval of around two seconds between each touch.

The containers were presented in a counterbalanced design. After completing the whole sequence of containers a second sequence was presented to the participants. Thus, four indications were obtained from every participant for each container. The peak value of each touch was taken as that of interest to the study. The median value of the four individual indications was computed and the standard error of all median values was calculated.



(a) (b)

Figure 6.5 - (a) sensors calibration using a reference load cell and (b) squeezing a container wearing sensors on three fingers.

6.4.8.1 Outliers and influential cases

Certain cases were assessed for identifying whether they exerted excessive influence on the trend of the data using the software SPSS®, version 19.0. For this purpose, results were firstly standardised into a Fisher's z-distribution. After standardising, potential outliers were identified using a confidence interval of ±1.96 (95% CI). Finally, influential cases were identified and removed from the regression model.

An observation was considered to be an influential case if the error variance of the predicted values changed largely when that observation was deleted from the linear regression model. Influential cases were flagged using residual statistics. However, different diagnostic mechanisms identified particular elements of what makes a case exert significant influence on the error variance. For this reason, an observation was taken as an influential case when at least two out of three diagnostic indices indicated that the regression model was biased by that observation. Thus, the effect of a single case on the whole regression model was indicated by Cook's distance (D) (Cook, 1979; Cook and Weisberg, 1982), such that the expression $D_i \geq 4/n$ flagged potential influential observations, where n is the sample size and i is the observed case. The difference between the predicted value when using an observation and the predicted value without using that observation was indicated by the statistical index DFFit (Belsey et al, 1980), such that $DFFit_i \geq 2\sqrt{1/n}$ indicated an influential case. Finally, the covariance ratio (CVR) indicated the influence of an observation in the confidence interval (Belsey et al, 1980). Influential cases were flagged by CVR when $CVR_i \leq 1-(6/n)$.

6.5 RESULTS

6.5.1 Preliminary Scales

The derivation of the RM used to analyse the data was the partial credit model. The model was adopted after a likelihood-ratio test, which presented a significant outcome indicated by p <0.05. This is an indication of that a nearly equal distance between thresholds across items is unexpected by the model.

A preliminary analysis using the software package RUMM2030® identified significant item-trait interaction, giving evidence of misfit to the model (Table 6.4). The σ of item-fit residuals for Stimulus 1 and the σ of person-fit residuals for Stimuli 2, 3, 4 and 5 were higher than the value of σ =1.40 established as a reference. Such values along with p <0.05, which pointed to lack of the invariance across the trait, indicated misfit to the model.

Table 6.4 – Preliminary analysis.

Stimulus	Item-fit res		Person-fit res		Chi-square	df	p	n	PSI
	Mean	σ	Mean	σ					
ST1	0.19	1.58	-0.14	1.37	84.46	32	<0.05	120	0.41
ST2	0.64	0.89	-0.19	1.60	101.57	32	<0.05	120	0.48
ST3	0.65	1.34	-0.13	1.50	97.30	32	<0.05	120	0.69
ST4	0.10	0.87	-0.36	1.44	55.40	32	<0.05	120	0.67
ST5	0.53	0.93	-0.18	1.51	73.81	32	<0.05	120	0.40

6.5.2 Calibration of Items

The score system was recoded by applying reversed order for Items I1, I11, I13, I14 and I16 for all of the individual scales. In addition, analyses of individual stimulus scales indicated inconsistent response pattern for some items. A response pattern was identified for each item and every stimulus, in which indicated disordered thresholds. Thus, some items were collapsed to four categories (Table 6.5).

Table 6.5 - Reversed items and items collapsed to four categories.

Item	Reversed	Re-scored				
		St1	St2	St3	St4	St5
I1	Y	Y	Y	Y	Y	Y
I2						
I3		Y	Y	Y	Y	Y
I4		Y	Y	Y		
I5				Y		
I6		Y	Y	Y	Y	Y
I7		Y				
I8		Y	Y	Y	Y	Y
I9		Y	Y	Y	Y	Y
I10		Y	Y	Y	Y	Y
I11	Y	Y	Y	Y	Y	Y
I12		Y	Y	Y	Y	Y
I13	Y		Y	Y	Y	Y
I14	Y	Y	Y	Y	Y	Y
I15		Y	Y	Y		
I16	Y	Y				

Examination of the pattern of residuals through a person-item correlation matrix identified items with associations greater than or equal to the absolute value of 0.30, which were taken as evidence of local dependency. The person-item correlation analysis combined with the individual item-fit analysis evidenced items with potential misfit to the model. The procedure removed Items I3, I12, I13 and I15 from the original set for Stimulus 1, Items I6, I10 and I13 for Stimulus 2, Items I2, I7 and I13 for Stimulus 3, Items I6, I9 and I16 for Stimulus 4, and finally Items I12 and I13 for Stimulus 5.

DIF was tested with Bonferroni adjustment for sex and age after re-scoring and removing items, indicating non-significance. Analysis of the individual persons-fit indicated that the pattern of responses from nine participants for Stimulus 1 presented high residuals, 10 participants for Stimulus 2, five for Stimulus 3, two participants for Stimulus 4 and six participants for Stimulus 5.

The fit of data to the model was examined from the set of the remaining items and the sample distributed into three groups of ability for each and every stimulus. Rasch analysis identified a non-significant item-trait interaction, which was deemed to be evidence that the data fit the model (Table 6.6, Column p).

The fit statistics for the calibrated scales indicated invariance across the measurement structure with p >0.05 for every stimulus. In addition, the σ of item-fit residuals and the σ of person-fit residuals obtained values between 0.70 and 1.29 and, thus, are within the acceptable limits in this study. The model's assumption of unidimensionality was met through a binomial test, which indicated that less than or equal to 5% of observations were expected to fall outside of the t-range of ±1.96 for the confidence interval for every stimulus (Column 95% CI).

Table 6.6 – Fit statistics for the calibrated scales.

Stimulus	Item-fit res		Person-fit res		Chi-square	df	*p*	*n*	Items	PSI	<95%CI
	Mean	σ	Mean	σ							
ST1	0.36	0.73	-0.20	1.19	35.67	24	0.06	111	12	0.62	0.036
ST2	0.27	0.92	-0.24	1.24	35.01	26	0.11	110	13	0.76	0.036
ST3	0.20	1.85	-0.26	1.25	33.93	26	0.14	115	13	0.70	0.039
ST4	0.18	0.82	-0.29	1.29	25.97	26	0.46	118	13	0.65	0.028
ST5	0.31	0.70	-0.23	1.25	39.83	28	0.07	114	14	0.56	0.048

6.5.3 Co-calibration of the Items Set

The calibration of a unified set of items aimed to reveal prevailing item characteristics for all stimuli. Thus, a further calibration based on the individual scales removed or added items from each stimulus to seek a well-adjusted set for a common scale. Eleven items were found to be the most balanced solution (Table 6.7). In addition, the scores were recoded to establish a common system for all of the stimuli. Items I1, I11 and I4 and I16 were kept with reversed scores.

Table 6.7 - Remaining 11-item set for the co-calibrated scales.

Code	Items
I1	The product in this container would give me a heavy, greasy film on my skin.
I2	The product in this container is likely to look and smell delightful.
I4	I feel the product in this container would hydrate my skin.
I5	The product in this packaging might be pricey.
I7	The container makes me feel like I would be buying a great product.
I8	The product inside the container would spread easily.
I11	The product inside this container could be sticky.
I12	The product in this packaging is likely to flow easily.
I14	It is quite hard to explain the product when touching its packaging.
I15	The product in this container could give me a refreshing sensation.
I16	The product in this packaging could be a bit boring.

Independent *t*-tests determined whether the scales met the model's assumption of unidimensionality. The outcomes indicated that the proportion of *t*-tests that fell outside of the *t*-range of ±1.96 for the confidence interval was less than 5% for every stimulus deemed acceptable to satisfy the assumption. The fit-statistics for the co-calibrated scales are summarised in Table 6.8. Although DIF had been tested with Bonferroni adjustment for sex and age after removing and adding items, no significant item bias was identified. The fit statistics for the calibrated scales indicated invariance across the measurement structure with $p \geq 0.05$ for every stimulus.

Table 6.8 - Fit statistics for the co-calibrated scales.

Stimulus	Item-fit res		Person-fit res		Chi-square	df	*p*	N	Items	PSI	<95%CI
	Mean	σ	Mean	σ							
ST1	0.21	0.77	-0.28	1.21	24.6	22	0.32	107	11	0.73	0.027
ST2	0.22	0.86	-0.25	1.15	29.0	22	0.14	108	11	0.71	0.036
ST3	0.01	0.80	-0.36	1.22	22.6	22	0.42	111	11	0.77	0.048
ST4	0.06	0.86	-0.40	1.20	27.7	22	0.18	113	11	0.73	0.037
ST5	0.07	0.96	-0.34	1.15	30.7	22	0.10	117	11	0.66	0.046

The person-item distribution for each threshold across all items is shown in Figures 6.6, 6.7, 6.8, 6.9 and 6.10. The groups of respondents and their ability levels are on the upper part. The location of item thresholds and their distribution are on the lower part of the graphs. The graphs indicate that the threshold distribution holds a good spread, revealing that the respondents are well targeted to the set of calibrated items even though there are gaps between some item-thresholds.

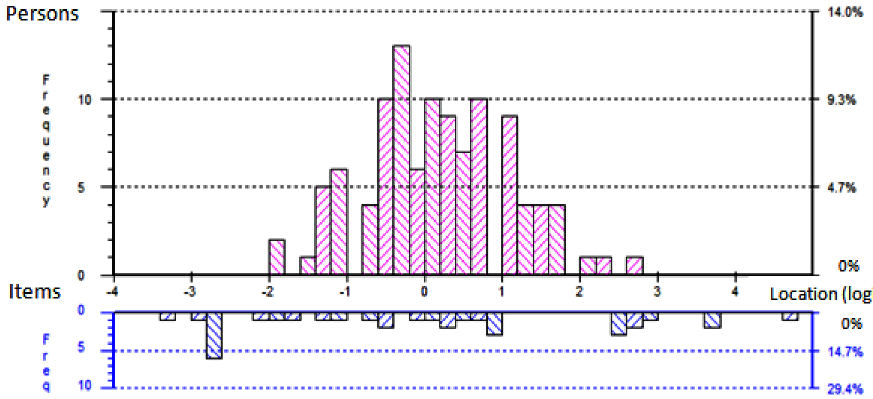


Figure 6.6 - Persons-item threshold distribution for Stimulus 1.

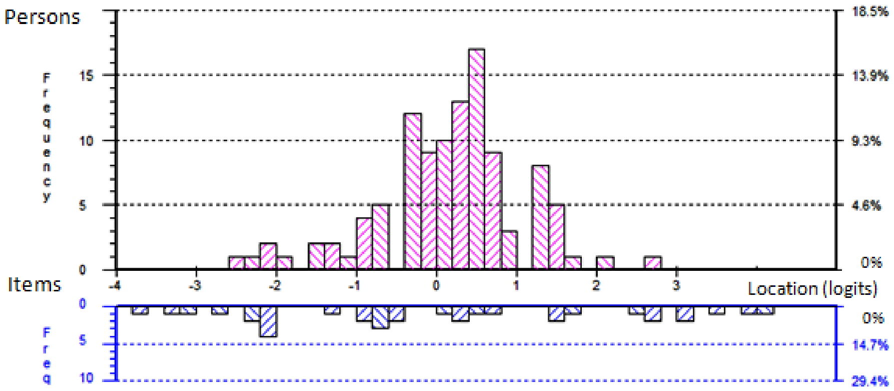


Figure 6.7 - Persons-item threshold distribution for Stimulus 2.

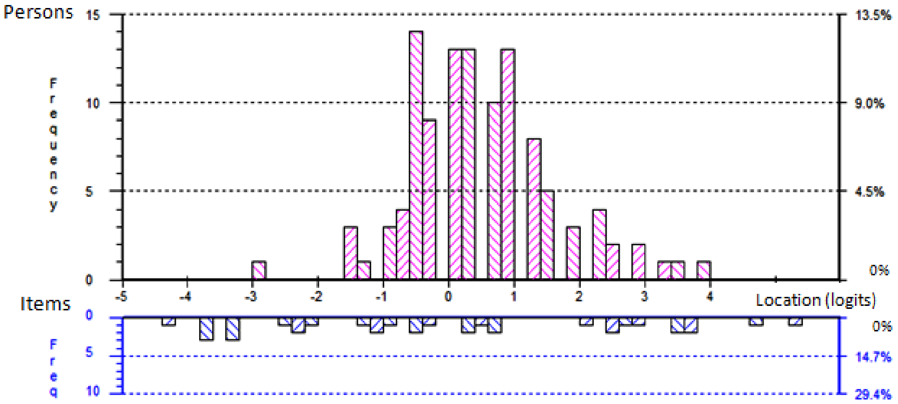


Figure 6.8 - Persons-item threshold distribution for Stimulus 3.

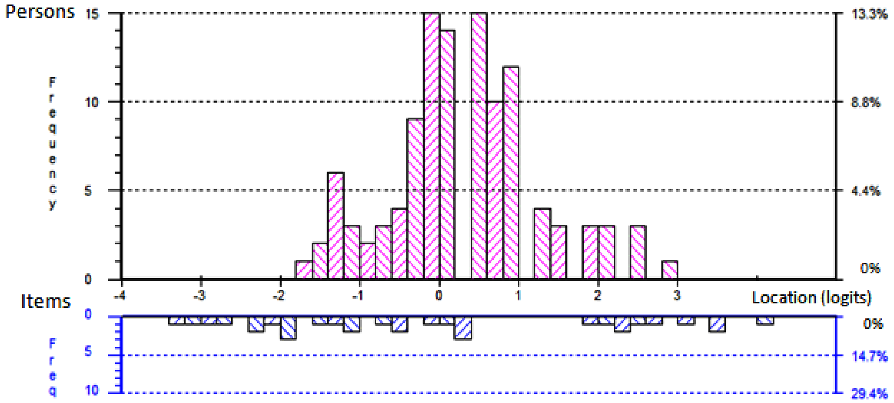


Figure 6.9 - Persons-item threshold distribution for Stimulus 4.

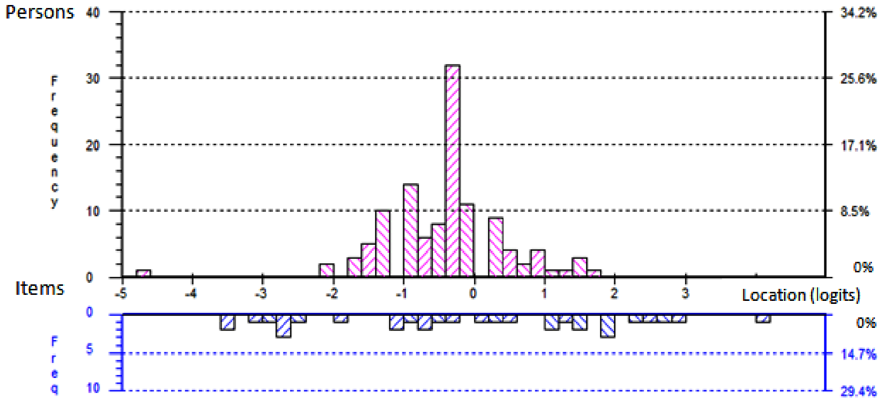


Figure 6.10 - Persons-item threshold distribution for Stimulus 5.

6.5.4 Faceted Rasch Analysis

The faceted Rasch approach allowed comparisons and interpretations of results on a sole frame of reference for all stimuli. The effect of different characteristics when comparing containers is demonstrated by the prevailing tendency amongst individuals of endorsing the affective attribute *perceptiveness*. The degree of endorsement was associated with the group of stimuli with higher measure of compliance than that with lower compliance although without any relationship regarding the product inside the container whatsoever.

The pool of statements used in the facet analysis was the co-calibrated 11-item set as well as the common score system across stimuli. Thus, the faceted approach replicated the calibrated set, generating a 55-item set.

The summary of the facet locations is presented in Table 6.9. The origin of zero was established as the default of the method applied in the analysis. The default for the origin constrained the stimuli facet and the items facet at the centre of the logit scale. That is, both facets had measurement a mean of zero. Also, the sum of the category coefficients was constrained to zero. Thus, solely the person facet floated on the continuum.

However, the locations of stimuli obtained through RUMM2030® were multiplied by minus one. The reason was that the software package did not allow any other configuration for the facets signal but that usually used for fairness of judgement. For this reason, the stimulus locations are presented in a proper magnitude although they are placed at the reversed side on the continuum.

The fit statistics for the calibrated scale indicated invariance across the measurement structure with p ≥0.05. However, the scale presented a PSI of 0.65, considered as indication of poor reliability when differentiating groups (Fisher, 1992).

Table 6.9 – Fit statistics of the facet approach.

Stimulus	Location x(-1)	SE	Fit-residual	Items	Mean location	SE	Fit-residual	ChiSquare	Df	p	PSI
St1	0.03	0.14	0.04	I1	-0.04	0.14	0.37	121.64	110	0.21	0.65
St2	-0.07	0.13	0.07	I2	-0.25	0.15	-0.08				
St3	0.34	0.15	-0.04	I4	0.00	0.13	-0.06				
St4	0.28	0.15	0.09	I5	-0.22	0.15	-0.03				
St5	-0.58	0.14	0.27	I7	0.04	0.15	0.02				
				I8	-0.50	0.13	0.04				
				I11	0.43	0.14	0.31				
				I12	-0.31	0.14	0.12				
				I14	0.27	0.13	0.44				
				I15	-0.02	0.14	-0.05				
				I16	0.60	0.16	-0.06				

The facets map (Figure 6.11) is the representation of the relative locations of all facets on the same logit scale. Person locations are plotted on the scale represented in the first column. Participant locations that indicate more inclination to endorse the attribute *perceptiveness* of a delicate moisturizer cream are plotted on the top of the scale those less inclined to endorse at the bottom. The second column of the facets map on the top

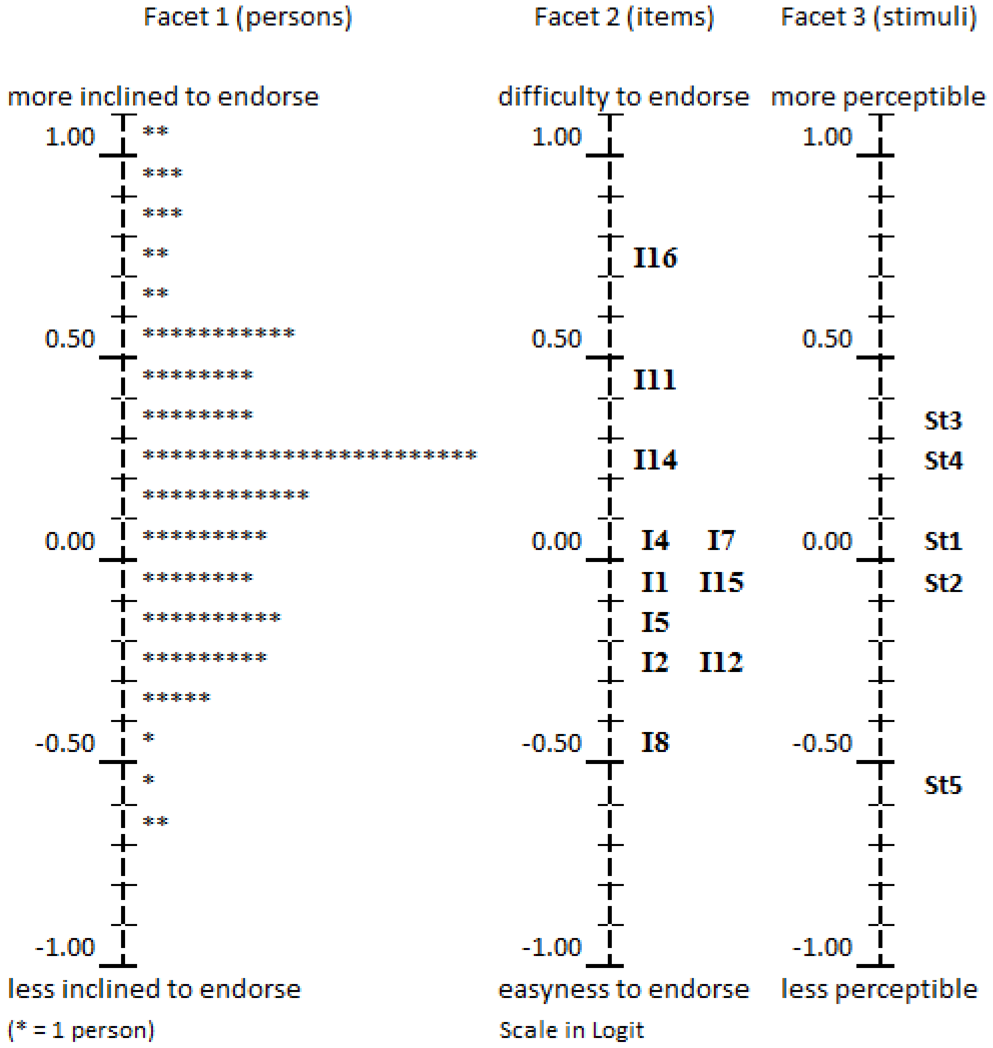


Figure 6.11 – Facet map for the latent attribute of the everyday product containers.

indicates items more difficult to endorse, i.e., items that obtained less consensus amongst participants to endorse them. The location of stimuli on the continuum demonstrated that the container with the lowest compliance was posited at the bottom of the scale in relation to the stimulus with higher compliance, indicating lower degree of endorsement to *perceptiveness* for the latter. On the other hand, the map indicates that according to participants' perception, there is an intermediate range of compliance subject to be more inclined to relate a container to a delicate cream product. Nevertheless, that range does not follow the order of the physical measurement for the stimuli compliance.

The facets map also identifies that there was shrinkage of the spread of persons on the logit scale when applied the common metric for all of the stimuli. However, Figure 6.12 indicates that the threshold distribution is widely spread, revealing that the respondents are well targeted to the set of calibrated items although there is low statistical power to differentiate two groups of persons.

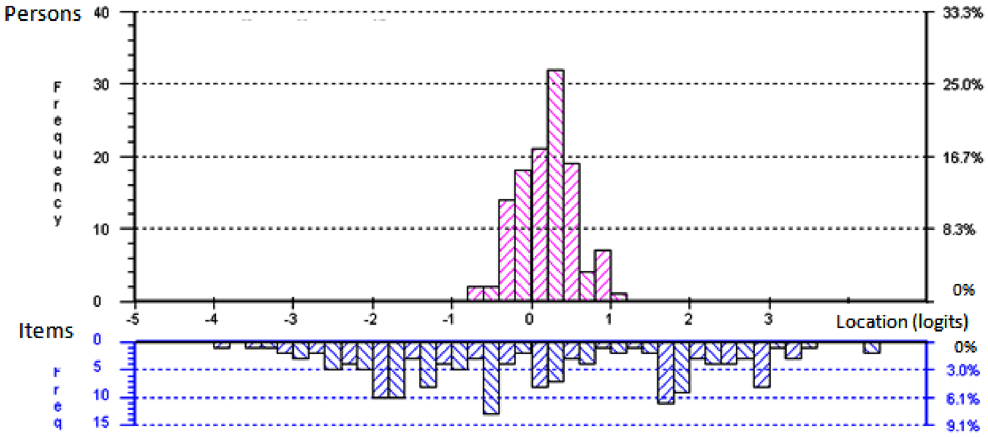


Figure 6.12 - Person-item threshold distribution using faceted Rasch model.

6.5.4.1 Differential stimuli functioning

The difference between SCCs was given by comparing the enclosed area and the imparity criterion. The SCCs were obtained from a cumulative function of raw scores based on the set of parameter values computed by RUMM2030® (Figure 6.13). The areas were obtained by integrating the difference of the polynomials that defined the SCCs. If two SCCs crossed over each other, then the reference point was the intersection of those two curves projected on the logit location axis (Table 6.10).

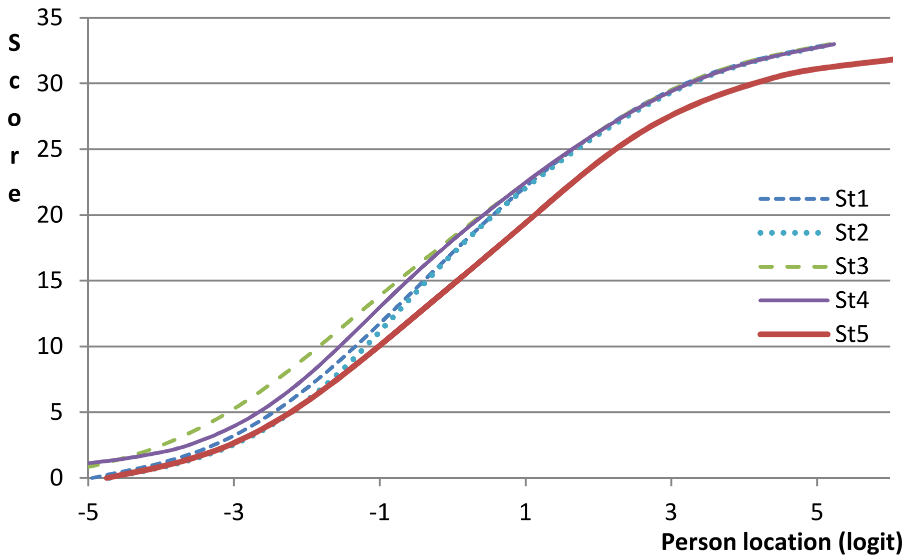


Figure 6.13 - Stimulus characteristics curves (SCCs).

It is noteworthy that if the difference was greater than zero, then the remaining area enclosed between SCCs represented upper SCC $\succ$ lower SCC. On the other hand, if the difference was less than zero then lower SCC $\succ$ upper SCC. Both of the cases indicated relative preference in a pairwise comparison.

Table 6.10 – Intersection points of the SCCs.

Stimulus	St1	St2	St3	St4	St5
St1	x	x	x	x	x
St2	3.935	x	x	x	x
St3	-	-	x	x	x
St4	2.013	3.186	4.865	x	x
St5	-	-3.444	-	-	x

The pairwise comparisons between stimuli through the enclosed area (A) contrasting with the enclosed area that included the measurement error (Ae) indicated that some of the stimuli did not present significant difference (Table 6.11). However, Stimulus 5 presented difference against Stimuli 1, 3 and 4, indicated in bold, and a marginal difference against Stimulus 2, indicated in italic. Another marginal difference can be observed when comparing Stimulus 2 with Stimulus 3.

Table 6.11 – Comparison of areas between SCCs and areas between SCCs along with errors.

	St1		St2		St3		St4		St5	
	A	Ae	A	Ae	A	Ae	A	Ae	A	Ae
St1	x	x	x	x	x	x	x	x	x	x
St2	2.84	15.11	x	x	x	x	x	x	x	x
St3	9.67	14.94	*12.51*	*13.41*	x	x	x	x	x	x
St4	3.52	13.65	6.36	12.12	5.75	11.95	x	x	x	x
St5	**16.60**	**14.81**	*12.70*	*13.27*	**26.27**	**13.11**	**20.12**	**11.82**	x	x

Based on those comparisons between areas it is possible to draw a conclusion that there are differences amongst containers with regard to the attribute *perceptiveness* when taking their compliance. Looking at the facets map (Figure 6.11) and the comparisons of areas (Table 6.11) the perceptiveness of the stimuli could be measured such that $St3 \approx St4 \approx St2 \approx St1 \succ St5$.

6.5.5 Assessment of the Metric's Stability

6.5.5.1 Fit statistics of the scale for the second sample

The fit statistics from the second sample were compared with those from the preliminary scale before calibration and those from the co-calibrated 11-item set. After calibration the scale presented a non-significant item-trait interaction, which was deemed evidence that the data fit the model. DIF was tested with Bonferroni adjustment for sex and age indicating non-significance.

Fit statistics indicated invariance across the measurement structure with p>0.05 for every stimulus. Similarly, the second sample also indicated invariance across the measurement structure with p ≥0.05 (Table 6.12, Column p). In addition, the σ of item-fit residuals and σ of person-fit residual for the second sample obtained values ≤1.21 and therefore, are within the acceptable limit in this study.

Furthermore, the model's assumption of unidimensionality was met through a binomial test, which indicated that less than or equal to 5% of observations were expected to fall outside of the t-range of ±1.96 confidence interval for every stimulus, pointing to an acceptable amount of deviating results, according to Table 6.12 (Column 95%CI). The PSI indicated power to distinguish between at least two groups (PSI ≥ 0.70), except for Stimulus 5. Column n shows the resultant sample after removing individuals with discrepant responses.

Table 6.12 - Fit statistics for the co-calibrated scales from the second sample.

Stimulus	Item-fit residual		Person-fit residual		Chi-square	df	*p*	*n*	PSI	95%CI
	Mean	σ	Mean	σ						
ST1	0.24	0.61	-0.12	0.96	18.07	22	0.70	59	0.71	0.023
ST2	0.23	0.47	-0.17	1.11	12.14	22	0.95	57	0.70	0.014
ST3	0.03	0.78	-0.25	1.06	16.56	22	0.79	60	0.70	0.022
ST4	0.30	0.58	-0.16	1.10	17.57	22	0.73	63	0.71	0.031
ST5	0.28	0.71	-0.20	1.21	32.06	22	0.08	50	0.58	0.043

6.5.5.2 Empirical consistency of the metric

Test for DIF using the two different samples and the common metric, indicating whether there was empirical consistency, presented non-significant variations in the expected values obtained from the first sample and those obtained from the second sample. That is, the expected values for both samples co-varied. The results of a two-way ANOVA pointed to p-values greater than a Bonferroni probability adjustment of 0.002 for all items of every stimulus (Table 6.13), suggesting that the difference between expected values of the first sample and of the second sample is not different from zero.

An independent t-test presented t-value (df = 327) =0.031 with two-tailed p =0.974, indicating that the differences between locations of scores for the first sample and for the second sample were statistically non-significant. Therefore, the analyses suggest the scale presents empirical consistency.

Table 6.13 - Results from a two-way ANOVA comparing the residuals of the two independent samples.

Item	Stimulus1			Stimulus2			Stimulus3			Stimulus4			Stimulus5		
	F	df	*p*	*F*	df	*p*	*F*	df	*p*	*F*	df	*p*	*F*	df	*p*
1	0.040	1	0.841	0.053	1	0.818	3.669	1	0.057	0.960	1	0.329	1.298	1	0.256
2	0.203	1	0.653	0.030	1	0.864	0.277	1	0.599	0.705	1	0.402	0.011	1	0.915
4	0.100	1	0.752	0.588	1	0.444	0.175	1	0.680	1.466	1	0.228	0.148	1	0.701
5	0.329	1	0.567	7.640	1	0.006	1.166	1	0.282	0.153	1	0.696	7.636	1	0.006
7	0.014	1	0.907	2.123	1	0.139	0.007	1	0.934	3.235	1	0.074	0.782	1	0.378
8	0.018	1	0.895	0.122	1	0.728	0.011	1	0.918	0.033	1	0.856	0.005	1	0.946
11	0.061	1	0.806	0.293	1	0.589	0.152	1	0.697	0.188	1	0.665	0.195	1	0.660
12	0.350	1	0.555	0.498	1	0.482	1.488	1	0.224	3.768	1	0.054	2.558	1	0.112
14	2.869	1	0.092	0.128	1	0.721	2.012	1	0.158	1.083	1	0.299	3.149	1	0.078
15	3.800	1	0.053	5.425	1	0.021	4.965	1	0.027	1.269	1	0.262	0.000	1	0.990
16	0.043	1	0.836	0.011	1	0.916	0.386	1	0.535	0.744	1	0.389	4.903	1	0.028

6.5.5.3 Cross-validation of the metric

Cross-validation was accomplished by the comparison between the calibration of the 11-item set obtained previously (Table 6.7) and located on the common metric (Figure 6.11), used as a benchmark, and calibrations based on the same items set using three out of the four groups of respondents. Item-by-item comparisons were carried out within each pair, i.e., the *datum* and an *n*-calibration after parameterisation through the facet model, taking into account the mean location of the item by stimulus (i.e., 11 items split in 55 sub-items). Thus, the two-sided test compared 11 pairs, each pair with five sub-items, taking the difference between estimates based on their standard errors. The results pointed to a non-significant difference between the datum and the *n*-calibrations (i.e., p >0.05) (Table 6.14).

Table 6.14 - Results from *t*-tests comparing mean location of items between the datum calibration and an n-calibration.

Item	Calibration 1			Calibration 2			Calibration 3			Calibration 4		
	t-stat	df	*p*	*t*-stat	df	*p*	*t*-stat	df	*p*	*t*-stat	df	*p*
1	0.015	8	0.988	0.292	8	0.777	0.427	8	0.680	0.644	8	0.537
2	0.613	8	0.567	0.024	8	0.982	0.246	8	0.812	0.383	8	0.712
4	0.278	8	0.788	0.323	8	0.755	0.026	8	0.980	0.047	8	0.964
5	1.134	8	0.290	0.525	8	0.614	0.743	8	0.478	0.212	8	0.837
7	0.310	8	0.764	0.174	8	0.866	0.999	8	0.347	0.238	8	0.818
8	0.479	8	0.645	0.166	8	0.873	0.440	8	0.672	0.033	8	0.975
11	0.635	8	0.543	1.057	8	0.321	0.339	8	0.743	0.599	8	0.566
12	0.015	8	0.988	0.118	8	0.909	0.037	8	0.971	0.418	8	0.687
14	0.216	8	0.834	0.457	8	0.660	0.490	8	0.637	0.393	8	0.705
15	0.128	8	0.902	0.046	8	0.964	0.364	8	0.725	0.748	8	0.476
16	0.103	8	0.921	0.048	8	0.963	0.079	8	0.939	0.045	8	0.965

Further evidence of the consistency of the calibration using cross-validation strategy was obtained by the differences of scores between two ability locations based on the common item equating. Figure 6.14 illustrates the comparison between the estimated scores associated with the ability locations obtained from the datum calibration and the estimated scores from an n-calibration. The diagonal line with a 45-degree slope through the origin defines perfect consistency of a calibration pair. Thus, the plotted points would lie on the identity line if they were perfectly matched.

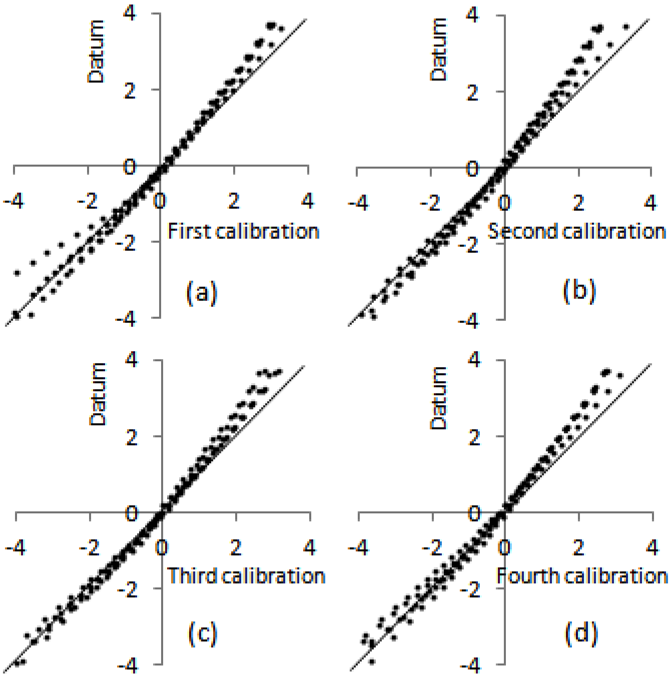


Figure 6.14 - Plots of score estimates on the datum calibration against score estimates on the n-calibrations. Plots of score estimates are associated with their equivalent ability locations (axes *X* and *Y*). Scale in logits.

6.5.6 Sensory Information

The force applied by participants on each container was transformed into a z-distribution with a mean equal to zero and σ equal to one. This standardization demonstrated that a few influential cases skewed the distribution of results, clustering at the lower forces and tailing toward the higher forces (Figure 6.15). Consequently, those observations presented non-negligible residual correlations.

In some cases the error associated with an observation can significantly affect the estimates of the regression model although without being detected when residuals are verified (Davies and Hutton, 1975)[34]. For this reason, three different mechanisms for

[34] The presence of influential observations can corrupt the purpose of making generalizations beyond the sample studied. For example, a few observations with exceedingly high values make the mean a less useful statistical mechanism in describing the central tendency of a distribution.

identifying influential cases were used to analyse the residuals of the linear regression model for sensory data.

The impact of the i^{th} observation on the whole regression coefficients was obtained by Cook's distance (D), taking values of D_i ≥0.03 as a potential influential case for the first sample (n=120) and D_i ≥0.06 for the second sample (n=66). The CVR indicated the influence of an observation on the error variance of the regression coefficients, considering values of CVR ≤0.95 highly influential for the first sample and values of CVR ≤0.91 for the second sample.

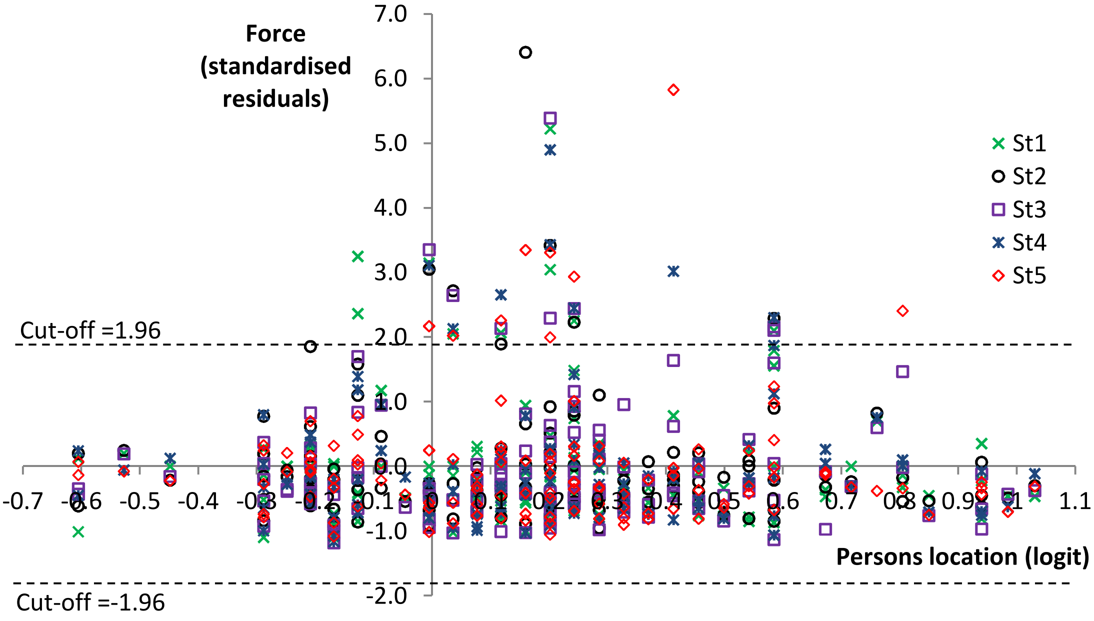


Figure 6.15 - Graphic representation of the force applied by participants of the first sample after a z transformation contrasting with the person location in logit.

Standardised DFFit indicated the effect of the change in the predicted value for an observation after removing it from the regression model. Highly influential cases were taken when DFFit ≥0.18 for the first sample and DFFit ≥0.25 for the second sample. Influential cases are displayed in Table 6.15 for the first sample and Table 6.16 for the second sample. In both Tables are only presented observations flagged as influential cases in at least two out of three influential diagnostic indices.

Removing the influential cases flagged by the diagnostic mechanisms demonstrated a significant effect in the error variance. σ was largely affected by deletion of 16 cases for the first sample and five for the second sample (Table 6.17).

6.5.6.1 Similarity of means between the first and the second sample

The individual means of the force applied on the containers from the first sample were compared with the individual means of the second sample using an independent *t*-test for determining whether they differed significantly.

Table 6.15 – Influential cases of the forces applied from the first sample (n=120) flagged through diagnostic tools.

Person ID (i)	Stimulus 1			Stimulus 2			Stimulus 3			Stimulus 4			Stimulus 5		
	D	CVR	DFFit	D	CVR	DFFit	D	CVR	DFFit	D	CVR	DFFit	D	CVR	DFFit
5	-	-	-	0.03	0.98	0.26	-	-	-	-	-	-	-	-	-
6	0.02	0.95	0.21	0.04	0.91	0.28	0.04	0.91	0.28	0.02	0.95	0.22	0.02	0.95	0.21
8	0.05	0.96	0.31	0.06	0.94	0.34	0.05	0.96	0.30	0.05	0.95	0.33	-	-	-
11	0.03	0.98	0.26	-	-	-	0.04	0.96	0.29	0.03	0.98	0.26	-	-	-
13	0.04	0.94	0.30	-	-	-	-	-	-	-	-	-	-	-	-
17	-	-	-	-	-	-	-	-	-	-	-	-	0.04	0.88	0.28
19	-	-	-	-	-	-	0.04	1.02	0.27	-	-	-	0.11	0.95	0.47
43	-	-	-	0.17	0.44	0.73	-	-	-	-	-	-	0.05	0.84	0.32
51	0.06	0.87	0.35	0.52	0.33	0.88	0.07	0.84	0.38	0.06	0.87	0.35	0.03	0.95	0.24
55	0.02	0.95	0.20	0.02	0.95	0.20	0.02	0.95	0.20	0.03	0.91	0.26	0.02	0.94	0.22
72	0.02	0.94	0.21	0.22	0.94	0.21	0.03	0.93	0.23	0.03	0.93	0.23			
83	-	-	-	-	-	-	-	-	-	0.05	0.88	0.34	0.20	0.52	0.75
84	0.04	0.87	0.29	-	-	-	-	-	-	0.05	0.83	0.33	-	-	-
93	-	-	-	-	-	-	0.02	0.94	0.21	-	-	-	-	-	-
101	0.12	0.61	0.55	0.05	0.83	0.33	0.12	0.58	0.57	0.10	0.65	0.50	0.05	0.84	0.32
113	0.08	0.86	0.42				0.02	0.98	0.22	-	-	- -	-	-	-

Table 6.16 - Influential cases from force applied by participants of the second sample (n=66) flagged through diagnostic indices.

Person ID (i)	Stimulus 1			Stimulus 2			Stimulus 3			Stimulus 4			Stimulus 5		
	D	CVR	DFFit	D	CVR	DFFit	D	CVR	DFFit	D	CVR	DFFit	D	CVR	DFFit
122	1.04	0.23	2.11	0.30	0.78	0.84	0.16	0.92	0.58	0.08	1.01	0.40	0.22	0.86	0.70
127	-	-	-	0.10	0.66	0.50	0.31	0.15	1.27	0.15	0.52	0.63	0.19	0.39	0.79
133	-	-	-	-	-	-	-	-	-	-	-	-	0.09	1.05	0.42
151	-	-	-	-	-	-	-	-	-	-	-	-	0.04	0.90	0.28
182	0.06	0.87	0.42	0.08	0.87	0.41	-	-	-	0.14	0.52	0.63	0.19	0.39	0.79

Table 6.17 – Mean, standard deviation and standard error of the force applied on the containers with and without influential cases for the first sample and for the second sample.

Stimulus	First sample (n=120)						Second sample (n=66)					
	Applied force with influential cases (N)			Applied force without influential cases (N)			Applied force with influential cases (N)			Applied force without influential cases (N)		
	Mean	σ	SE	Mean	σ	SE	Mean	σ	SE	Mean	σ	SE
St1	16.9	13.1	1.2	13.7	6.7	0.6	13.6	9.5	1.2	12.4	5.9	0.7
St2	19.9	16.8	1.5	16.3	8.3	0.8	15.4	8.6	1.1	14.2	6.5	0.8
St3	21.3	15.5	1.4	17.4	8.5	0.8	18.0	13.4	1.7	16.2	7.3	0.9
St4	20.4	15.1	1.4	16.9	8.2	0.8	16.3	8.8	1.1	15.1	6.8	0.9
St5	36.4	29.5	2.7	29.4	13.9	1.3	31.0	19.3	2.4	27.1	11.7	1.5

To compare the means data sets were similarly taken without influential cases. The results from the *t*-tests have suggested the differences between mean values of the first sample and of the second sample are not significantly different from zero at 5% level of the confidence interval (Table 6.18) although Stimulus 2 presented just marginal non-significance.

Table 6.18 – Results of *t*-tests for differences of mean values of the applied forces on containers from the first sample and from the second sample.

Stimulus	*t* Stat	df	*p*
St1	1.257	172	0.211
St2	1.772	173	0.078
St3	0.919	171	0.359
St4	1.487	172	0.139
St5	1.109	170	0.269

6.5.7 IMPLICATIONS FROM THE METRIC'S STABILITY

6.5.7.1 Co-calibration of the items incorporated into the metric

The locations of item facet and the stimuli fulfilment facet provided a stable frame of reference for further analysis. Thus, the logit values of item location obtained from the co-calibrated 11-item set were used as reference values for determining the calibrations of five additional statements (Section 6.5.7).

Those reference values were established by the thresholds obtained from the calibrated metric for each scale using the partial credit model. Accordingly, the parameter values of the additional item estimates were associated with the core metric.

Preliminary analysis using the software package RUMM2030® identified significant item-trait interaction, evidencing some misfit to the model (Table 6.19). The chi-square probability for all of the stimuli was <0.05 (Column *p*), pointing to a lack of the invariance across the trait. Binomial tests indicated statistical significance of deviations from a theoretically expected proportion of observations (Section 2.4.7.9). The expected proportion of paired *t*-tests that fell outside of the *t*-range of ±1.96 for the confidence interval was greater than 5% for all of the stimuli, indicating the scales presented some degree of multidimensionality.

To co-calibrate the scales, the score system was recoded by applying reversed order for Items I20 and I21. In addition, analyses of individual stimulus scales indicated inconsistent response patterns for some items. Thus, Items I17, I18, I20 and I21 were collapsed to four categories. The person-item correlation analysis combined with the individual item-fit analysis evidenced items with potential misfit to the model.

The procedure removed Items I18 and I20 from the preliminary scale for all of the stimuli. Table 6.20 presents the additional items that fitted the model along with the co-calibrated 11-item set.

Table 6.19 - Fit statistics from co-calibration using a non-anchored, 11-item set along with the additional items.

Stimulus	Scale	Item-fit res		Person-fit res		Chi-square	df	p	n	PSI	<95%CI
		Mean	σ	Mean	σ						
ST1	Preliminary	1.44	1.62	0.46	1.39	201.67	32	<0.05	66	0.62	>0.05
	Calibrated	0.35	0.51	-0.17	1.30	30.78	28	0.33	66	0.72	0.037
ST2	Preliminary	0.49	1.20	-0.01	1.35	68.88	32	<0.05	66	0.73	>0.05
	Calibrated	0.24	0.41	-0.10	1.05	20.32	28	0.85	60	0.69	0.046
ST3	Preliminary	0.25	1.26	-0.10	1.25	60.40	32	<0.05	66	0.81	>0.05
	Calibrated	0.14	0.68	-0.24	1.24	21.25	28	0.82	61	0.79	0.028
ST4	Preliminary	0.15	1.51	-0.18	1.44	91.58	32	<0.05	66	0.76	>0.05
	Calibrated	0.12	0.61	-0.28	1.35	29.53	28	0.39	63	0.75	0.035
ST5	Preliminary	1.33	1.47	0.46	1.28	152.97	32	<0.05	66	0.72	>0.05
	Calibrated	0.39	0.93	-0.16	1.40	41.7	28	0.05	54	0.67	0.027

Table 6.20 - The 14-item set for the co-calibrated scales

Code	Items
I1	The product in this container would give me a heavy, greasy film on my skin.
I2	The product in this container is likely to look and smell delightful.
I4	I feel the product in this container would hydrate my skin.
I5	The product in this packaging might be pricey.
I7	The container makes me feel like I would be buying a great product.
I8	The product inside the container would spread easily.
I11	The product inside this container could be sticky.
I12	The product in this packaging is likely to flow easily.
I14	It is quite hard to explain the product when touching its packaging.
I15	The product in this container could give me a refreshing sensation.
I16	The product in this packaging could be a bit boring.
I17	I could get just the right amount of the product when I squeeze its container.
I19	I feel this container as a skin care product.
I21	I could find no consistency in the product inside this container.

The fit of data to the model was examined from the set of the remaining items and the sample distributed into three groups of ability for each and every stimulus. Rasch analysis identified a non-significant item-trait interaction (Table 6.19, Column p). DIF was tested with Bonferroni adjustment for sex and age after re-scoring and removing items, indicating non-significance. Analysis of individual person-fit indicated that the pattern of responses from six participants for Stimulus 2 presented high residuals, five participants for Stimulus 3, two for Stimulus 4 and 12 participants for Stimulus 5 (Table 6.19, Column *n*).

The assumption of unidimensionality was met through a binomial test, which indicated that less than or equal to 5% of observations were expected to fall outside of the *t*-range of ±1.96 for the confidence interval for every stimulus (Table 6.19, Column <95%CI).

6.5.7.2 Establishing a metric with additional items using the faceted rasch approach

Using the faceted Rasch approach through the software package RUMM2030® the set of the co-calibrated items was equated without anchoring items on the continuum.

Individual item-fit residual indicated no absolute extreme values (i.e., ≥±2.50). Data from seven persons were removed from the analysis because they presented high value of fit-residuals (i.e., ≥±2.50). The scoring system was adjusted concerning disordered thresholds patterns. Thus, all items were collapsed to four categories either during the co-calibration of items or during the equating procedure.

Invariance across the measurement structure was indicated by a non-significant item-trait interaction with *p* =0.96. The PSI of 0.77 indicated that at least two groups of persons can be distinguished by the metric (Fischer, 1992). The threshold distribution pointed to a good spread of items thresholds on the continuum, indicating that the respondents are well targeted to the set of calibrated items (Figure 6.16). However, some thresholds are working at extreme locations.

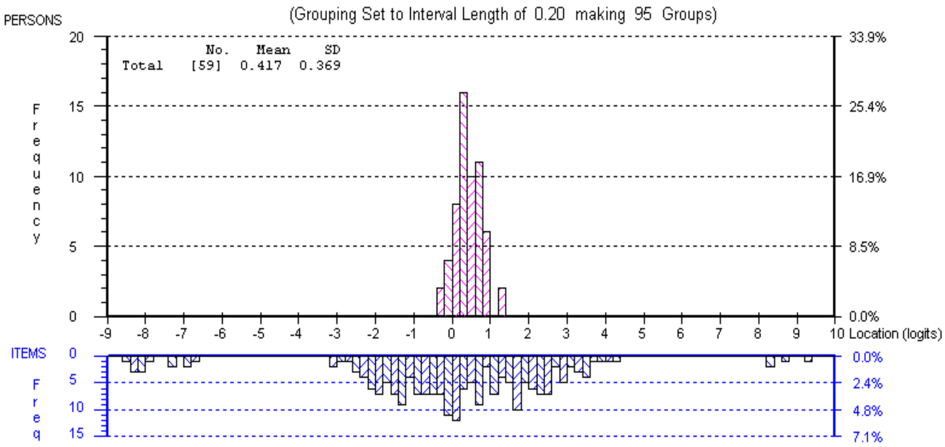


Figure 6.16 – Item-person threshold distribution.

The facets map (Figure 6.17) presented slight differences between locations of the same item obtained from the first sample after calibration of the 11-item set and from the second sample after calibration with the additional items. Nevertheless, items were located within a range of ±0.35 logit of difference comparing against the metric obtained from the first sample, as expected for the size of the sample (i.e., *SE*=±0.50 logit; Section 6.5.2.2). However, an exception arose for Item I5 that presented a location difference of 1.19 logits.

It is also noteworthy that the metric obtained from the second sample using the 14-item set indicated more pronounced difference between locations of Stimulus 3 and Stimulus 4. Furthermore, the metric points to a lower degree of endorsement to *perceptiveness* for the Stimulus 5 as comparing against the other stimulus containers.

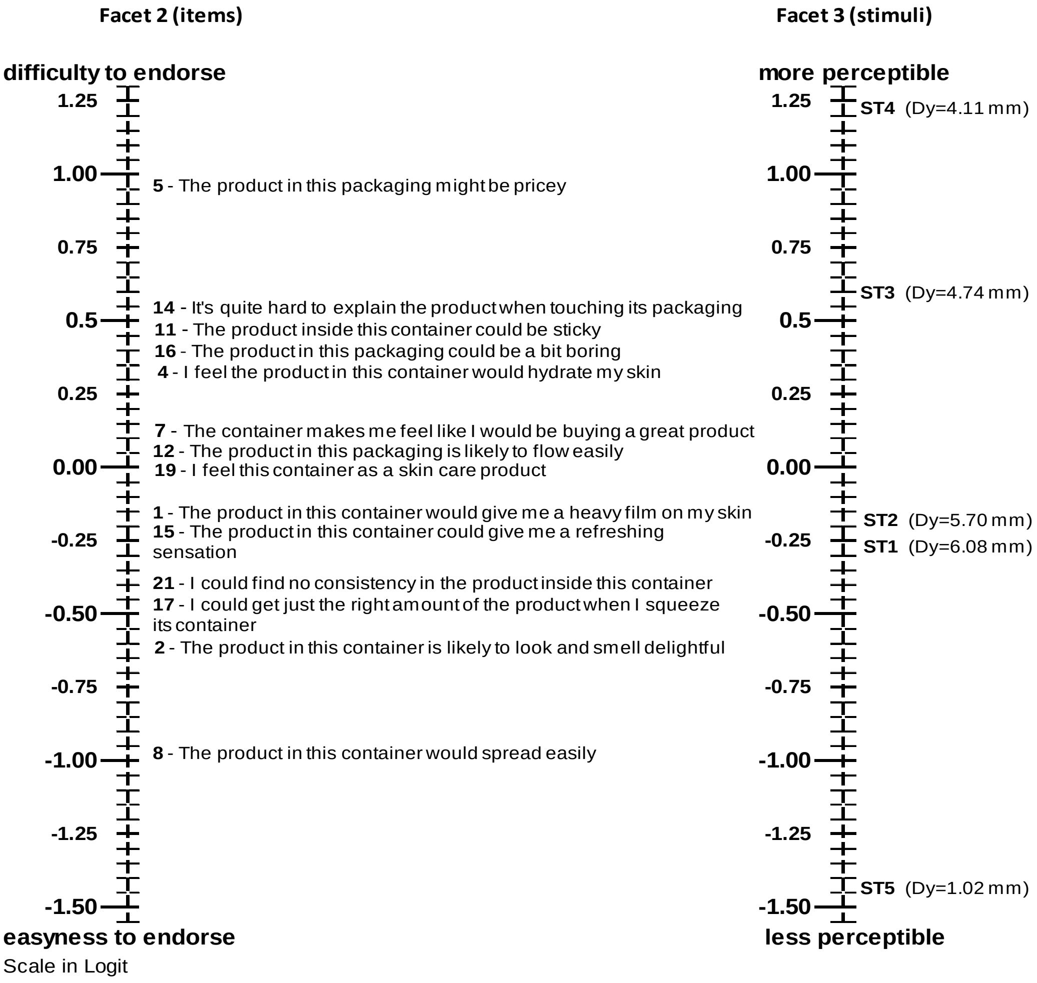


Figure 6.17 – Metric with a 14-item set for the affective attribute of the product containers.

6.6 REMARKS

6.6.1 Limitations of the Empirical Approach

During the calibration of the datum scale five items were removed from the preliminary pool. Analysis of misfit to the model suggested different reasons for discarding items. For example, Item I6 *the container feels only half filled when squeezing it* and Item I10 *it is easy to know how much is left in the packaging* could have been related by respondents to a different perception, rather than while squeezing product containers. Thus, those items could belong to another sub-scale of containers *perceptiveness*. However, this different sub-scale was not investigated in this study, taking the 11-item set as the basis for the *n*-calibrations.

Another limitation of the study was the sample size adopted to calibrate the scale. The sample size was based on items calibration stability related to its modelled *SE* and with moderately off-target observations. The scale's *SE* reflects the precision of the measurement. The precision of the calibrated items on unseen data when the scale is re-calibrated using all the available data will likely be higher because calibration normally improve the precision as more data becomes available.

It is also noteworthy that even though the individual scale for Stimulus 5 improved through the calibration, its PSI suggests a low statistical power to distinguish at least two groups of respondents. Thus, if the scale for Stimulus 5 was considered individually, a larger sample size could be required or additional items should be incorporated. However, the overall metric for all stimuli presented a good spread of thresholds on the continuum, indicating a reasonable sample-item targeting of the structure with the 11-item set.

6.6.2 Stability of the Measurement

The *t*-tests have not indicated a significant statistical difference between calibrations while cross-validating the scale although the ability locations did not achieve identical estimates. In Figure 6.14a, for example, score differences are identified at the higher and at the lower ability locations. Figure 6.14b, c and d indicate score differences at higher ability locations when comparing against the identity line. These differences could be associated with the measurement error in the compared calibrations. Another reason might be the methods for pooling items are not precise enough. Thus, although some items could lie on the continuum at different locations, they also could be excessively close, overlapping measurement errors and therefore, subject to non-significant statistical difference. Consequently, the locations of those items subjects to different samples could present slight fluctuations.

6.6.3 Concluding Comments

The results of the cross-validation in this study suggest the measurement structure of affective responses can be generalised, confirming the first hypothesis of the empirical study. The empirical study presented in this chapter has given evidence that the measurement of affective responses to design elements using the RM is independent of the sample distribution. That is, the measurement through calibrated structures will present empirical consistency within a similar context even when different groups of respondents are used.

In this chapter the second hypothesis of the empirical study has also been confirmed. When calibrated items are kept as a mainstay of the measurement structure, further items can be calibrated and accommodated into the structure to convey further information about the affective attribute.

The stability of the measurement structure obtained through calibration supports the establishment of the correlation between affective responses and sensory information. The location of stimuli on the continuum demonstrated that the container with lower compliance is located at the bottom of the scale in relation to the stimulus with higher compliance, indicating a lower degree of endorsement to *perceptiveness* for the first. On the other hand, the metric indicates that according to participants' perception there is the likelihood of the intermediate range of compliance being associated with a container of a moisturizer cream. Nevertheless, that range does not follow the order of the physical measurement for the stimulus compliances. In addition, the mean value of the force applied on the containers by participants in this study indicated a similar order of those for the affective responses (see Table 6.17). This is evidence that the participants' impression of a moisturizer cream when squeezing the containers is associated with the force applied on them even though the association is not linear with the physical property of compliance. However, the association between sensory information and affective responses could merely be made by overlapping metrics rather than through a direct correlation (see Chapter 7).

The invariant comparisons property of the RM has been reached within the frame of reference of the study. This statement is made possible given that items' difficulty are independent of the distribution of abilities in the relevant group of respondents and person ability estimates are independent of the set of items used for estimation. Because there is a stable relation between items after their calibration, the ratio of their relative endorsement ratings will remain statistically equivalent. Nevertheless, it is still unclear whether such stability will last when predicting the outcome regarding the pool of calibrated items based on the sensory variable for a different set of containers designed to obtain the same particular affective response.

CHAPTER 7

Modelling product features for affective responses using a Rasch-calibrated metric

In this chapter a correspondence for the metrics of physical properties, sensory information and a latent trait is established. The correspondence was achieved through an empirical approach that used the linear 14-item scale obtained from the preceding study to measure the relative differences of persons' impressions of a moisturizer cream when squeezing a collection of everyday product containers. The physical element compliance was established as a shared component between the metrics. The correspondence specified a set of container prototypes designed and manufactured to stimulate particular affective responses. A new Rasch-calibrated scale compared responses from a different group of persons to the prototypes. Although there were differences between the magnitudes of force applied on the existing products and on the prototypes, the results indicated that the prototypes fulfilled the affective attribute within the range of modelled compliances.

7.1 ESTABLISHING THE CORRESPONDENCE BETWEEN METRICS

Modelling the correspondence between a physical property of a product and a person's attitude associated with his or her interaction with the product intuitively leads one towards the idea of measurement. This requires that the comparisons between two elements in a measurement structure be independent of the instruments used for measuring them.

Throughout the thesis evidence has been given that the formal structure of the RM allows independence between person parameter, item parameter and stimulus parameter and therefore, their mathematical separation (Rasch, 1960, 1980; Linacre, 1989). In this research, the persons' parameter is associated with the persons' inclination to endorse the relevant affective attribute to a design element or a product feature. The item parameter refers to the adjectives or to the statements used as independent variables to quantify the affective attribute. The stimulus parameter is related to the different characteristics of objects presented to the respondents.

Accordingly, using an empirical approach, the previous study established a scale for measuring the relative differences of a collection of the containers associated with the users' affective responses. Those responses originated from the persons' intuitive impression of a moisturizer cream when squeezing a set of everyday product containers containing different characteristics of compliance. The study additionally collected sensory information through the force applied on the containers (see Chapter 6). Based on the fundamentals of Rasch measurement theory, it is possible to assume that the calibrated structure can linearly correlate affective responses with physical elements on a scale continuum.

7.2 HYPOTHESIS OF THE EMPIRICAL APPROACH

The linearity and stability of the unidimensional continuum achieved through the RM in the previous empirical study gave rise to the hypothesis that a new set of containers could be designed according to the compliances modelled on the basis of affective responses to the existing products (see Section 1.6.1).

7.3 AIMS

The empirical study aimed to test the hypothesis above through a correspondence between the metrics for affective responses and for sensory information. This can allow modelling the containers within a range of values of compliance for an impression of a moisturizer cream. Furthermore, the study aimed to investigate whether the calibrated metric could be the core of a different scale with measurement properties to compare a set of container prototypes designed to stimulate different degrees of affective impression.

7.4 METHOD

7.4.1 Scope and Metric for Modelling Compliances of the Containers

The scope of the experiment was to measure the relative importance of the packaging material to obtain an intuitive impression of a moisturizer cream as a product feature. The 14-item scale calibrated through the RM in the previous study was used to establish the metric for the current empirical approach (see Figure 6.17). In the preceding study the scale was initially developed on the basis of 11-calibrated items using five everyday products available in the market (see Section 6.5.2.1). Later, three more calibrated items were incorporated in the scale using the same stimuli although computing data from a different sample of persons (Table 7.1).

7.4.2 Force Applied on the Containers

The magnitudes of force applied by participants on the existing containers were obtained from the previous study. Participants from two different administrations of the preceding experiment were asked to squeeze the containers wearing tactile sensors on three fingers (see Section 6.5.8). The resulting forces when participants squeezed the containers are found in Table 6.17 based on values without taking into account the cases that exerted excessive influence on the trend of the data in the linear regression model (see Section 6.6.6).

Table 7.1 – Set of 14-calibrated items used in the study.

Code	Items
I1	The product in this container would give me a heavy, greasy film on my skin.
I2	The product in this container is likely to look and smell delightful.
I3	I feel the product in this container would hydrate my skin.
I4	The product in this packaging might be pricey.
I5	The container makes me feel like I would be buying a great product.
I6	The product inside the container would spread easily.
I7	The product inside this container could be sticky.
I8	The product in this packaging is likely to flow easily.
I9	It is quite hard to explain the product when touching its packaging.
I10	The product in this container could give me a refreshing sensation.
I11	The product in this packaging could be a bit boring.
I12	I could get just the right amount of the product when I squeeze its container.
I13	I feel this container as a skin care product.
I14	I could find no consistency in the product inside this container.

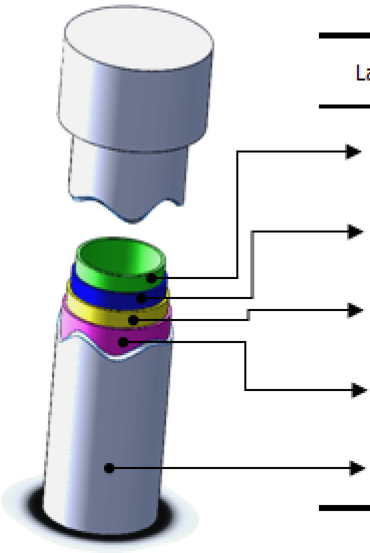
Notes: The item set was obtained from a previous calibration (see Chapter 6). Items were recoded for this study.

7.4.3 Manufacture of the Container Prototypes

Five prototypes were designed to test the hypothesis that a new set of containers would be consistent with the compliances modelled for affective responses. All containers were manufactured within the same dimensions, adopting a cylinder shape with the body diameter of 35mm and height of 160mm. A cap with diameter of 46mm and height of 52mm was used to seal the container. Every container received 139.5cm^3 of the same moisturizer, filling about 90% of the container's internal volume.

The surface roughness of all containers was designed to be similar. The containers' surface roughness was measured through a stylus surface profilometer RTH Form Talysurf 120L. The diamond stylus with radius 2.5μm of the Talysurf machine scanned an area of 5mm × 5mm on the surface and recorded the peaks at a resolution of 1024 data points per mm^2. These were then filtered by the acquisition software to remove any apparent form. Post-processing software was finally used to extract the values of the arithmetical mean of roughness R_a (μm).

Layers of different materials were used to establish the range of compliances for each container (Figure 7.1). An acrylic adhesive was applied on the surface of the laminated materials to adhere the layers. The measurement of the layers' thickness was obtained with a micrometer screw gauge (Mitutoyo 0.001mm).



Layer	Ra (µm)	Material	Type	Thickness (mm)
e	-	High density polyethylene	Laminated sheet	0.743
d	-	Low density polyethylene	Laminated sheet	0.406
c	-	Low density polyethylene	Laminated sheet	0.306
b	-	Aluminium	Laminated foil	0.035
a	1.17	Polypropylene	Laminated film	0.060

Figure 7.1 - Layers and materials used for the composition of the container prototypes.

For the physical measurement of the compliances of the container prototypes, a testing system was used that consisted of a force platform multi-component dynamometer (Kistler type 9256C2), an X–Z motion table (Series 1000 Cross Roller, Motion link), a steel ball of radius 10mm, a controller, and a computer. The containers were positioned between the steel ball and the force platform. The ball was pressed against the surface of each stimulus and the ball's displacement D_y with increasing load F_y was recorded. The measure of compliance was empirically taken to be the value of D_y (mm) when F_y was 3N (Chen et al. 2009; Shao et al., 2010).

The aim of manufacturing the containers was to determinate the correspondence between the metrics when modelling products for affective responses. Thus, the scope for designing the multi-layer containers did not take into account any barrier against oxygen ingress and aspects of sterilisation or aseptic filling. The layers were, therefore, exclusively used to address the property of the materials' compliances.

7.4.4 Data Collection for Affective Responses to the Container Prototypes

The five prototypes were presented to 67 respondents, 41.8% females and 58.2% males, 13.4% with age in the range of 18 to 25, 61.2% with age in the range of 26 to 35, and 25.4% over 35. Participants received £5.00 as a compensation for taking part in the study. This value has been a typical compensation in previous similar experiments. The sessions took place in the engineering systems laboratory at the fifth floor of the School of Mechanical Engineering, University of Leeds, Leeds, UK, in June and July 2012. Ethical approval for this empirical study was obtained from the University of Leeds Research Ethics Committee (ethics reference number MEEC 11-036).

Participants gave their ratings after squeezing the prototypes using computer-based self-report questionnaires against the statements obtained previously (see Section 6.5.3). A physical barrier was installed between the respondents and the stimulus containers preventing visual contact. Furthermore, participants did not make contact with the product inside the containers. The computerised system established the order in which participants were required to consider the containers using a counterbalanced design. The order of the statements on the questionnaires was automatically randomised by the system. Written information about the activity was provided in advance on the experiment's website. A verbatim protocol was used for giving instructions before the test.

7.4.5 Calibration of Scales Using Rasch Analysis

To measure affective responses to the container prototypes the items calibrated previously were also used as a core to establish a new scale. Rasch analysis was carried out with the MFRM (see Chapter 4) through the software package RUMM2030® (professional edition, 2012)[35].

The calibration of items as well as the cut-off points adopted in this empirical study followed the procedures in a similar fashion to those established for the study in Chapter 3 and Chapter 6 (see Section 2.4.7). These included a verification of the score system, tests of fit, analysis of the item-person interaction, analysis of the persons' response pattern and test for differential item functioning. The test for differential stimulus functioning (DSF) used the rationale proposed in Section 4.4.3. In addition, the data were tested for the assumptions of response independence and unidimensionality, both through the alternative technique for subtests proposed for MFRM (see Section 5.3.2).

7.5 RESULTS

7.5.1 Modelling the Correspondence between Metrics

The metrics for persons' affective responses to the containers and for persons' force applied on the containers were in this study designed to share the element compliance.

[35] RUMM2030© professional edition is an updated version of RUMM2030© licensed edition used in the previous empirical study.

Figure 7.2 is the graphical representation of both metrics. The correspondence between the stimulus locations and its compliance level are plotted on the upper graph. The horizontal dashed line indicates the lower bound of endorsement to the affective attribute.

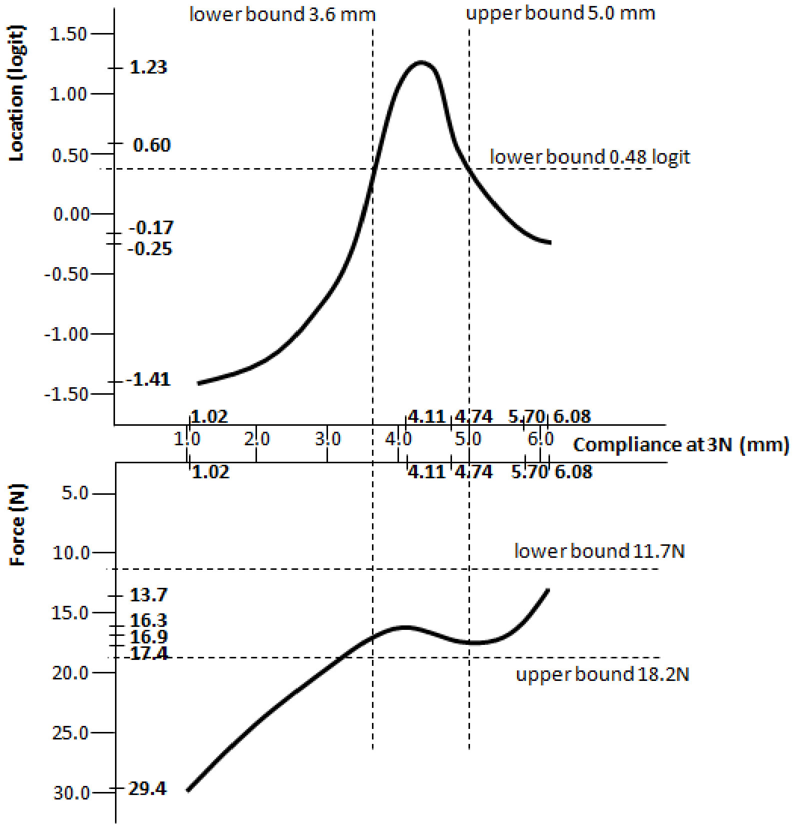


Figure 7.2 - Representation of the overlapping metrics. Lines were smoothened in the upper and lower graphs.

This boundary was taken as the lowest positive location (i.e., Stimulus 3) plus its measurement error in logits. The lower and upper bound indicated by vertical dashed lines were established by the intersections of the line representing the lower bound of endorsement with the curve estimated from the plots of affective responses. The lower graph represents the correspondence between the force applied on the containers and the containers' compliances. The range of force applied on the containers for obtaining particular affective responses was established by the modelled curve from the estimated plots based on Table 6.17. The horizontal dashed line indicates the lower and upper bound of applied force, taking into account the interval of compliances obtained from the previous upper graph. The boundaries were established by the persons' force in this interval and its standard error.

As a result of the compliances modelled for affective responses, represented in Figure 7.1 and 7.2, five container prototypes were designed to test the hypothesis that the new containers would capture the persons' impression of a moisturizer cream to some degree. Table 7.2 presents the general composition of layers that varied according to the modelled compliances (see Figure 7.1). Layer 'a' was used to present similar surface perception for every stimulus. The combination of different materials adhered together yielded different levels of compliance. Taking into account the characteristics of compliance and based on the modelled curves, different expected levels of endorsement were drawn to the affective attribute.

Table 7.2 - Layers composition of the container mock-ups.

Stimulus	Prototype 1	Prototype 2	Prototype 3	Prototype 4	Prototype 5
Code	ST1	ST2	ST3	ST4	ST5
Layers composition	a, b, c	a, b, d	a, b, c, c	a, b, c, e	a, b, d, e
Compliance D_y(mm)	4.17	4.12	3.81	2.24	1.14

7.5.2 Establishing the Metric for the Affective Responses to the Prototypes

7.5.2.1 Co-calibration of the set of items

The 14-item set from the previous study (see Section 6.6.7.1) was used as a reference to determine the calibration of the scale for the container prototypes. The likelihood-ratio test presented significance with p >0.05 indicating different intervals between categories and therefore, the partial credit was adopted (Masters, 1982) within the MFRM (Linacre, 1989).

Preliminary analysis identified significant item-trait interaction, evidencing some misfit to the model (Table 7.3). The chi-square probability <0.05 for Stimuli 1, 4 and 5 (Column p) pointed to lack of the invariance across the trait. Binomial tests indicated statistical significance of deviations from a theoretically expected proportion of observations. The expected proportion of paired t-tests that fell outside of the t-range of ±1.96 for the confidence interval was greater than 5% for all stimuli, indicating that the scales presented some degree of multidimensionality (Column <95%CI).

To co-calibrate the scales, the scores were recoded by applying reversed order for Items I1, I7, I9, I11 and I14. The analysis of the score system for the stimuli indicated an inconsistent response pattern for some items. Thus, after individual analysis (Table 7.4) those items were collapsed to four categories except Item I2 that kept five options of response. The person-item correlation analysis combined with the individual item-fit analysis evidenced no potential misfit to the model. Therefore, all items from the 14-item set remained after calibration.

Table 7.3 - Fit statistics from co-calibration using the 14-item set

Stimulus	Scale	Item-fit res		Person-fit res		Chi-square	df	p	n	PSI	<95%CI
		Mean	σ	Mean	σ						
ST1	Preliminary	0.38	0.88	-0.17	1.49	42.63	28	<0.05	67	0.78	>0.05
	Calibrated	0.25	0.84	-0.17	1.24	34.37	28	0.19	62	0.79	0.036
ST2	Preliminary	0.47	1.36	-0.19	1.52	40.30	28	0.06	67	0.85	>0.05
	Calibrated	0.32	1.15	-0.20	1.37	28.88	28	0.42	67	0.86	0.037
ST3	Preliminary	0.33	0.47	-0.22	1.42	25.99	28	0.57	67	0.82	>0.05
	Calibrated	0.24	0.55	-0.28	1.32	35.30	28	0.16	62	0.83	0.009
ST4	Preliminary	0.72	1.20	-0.12	1.51	42.81	28	<0.05	67	0.78	>0.05
	Calibrated	0.41	0.86	-0.18	1.40	39.89	28	0.07	63	0.76	0.036
ST5	Preliminary	0.73	1.68	-0.05	1.37	58.11	28	<0.05	67	0.74	>0.05
	Calibrated	0.29	0.73	-0.14	1.21	35.58	28	0.15	62	0.66	0.039

The sample was distributed into three groups of person's ability for each and every stimulus. The calibrated scales using Rasch analysis pointed to a non-significant item-trait interaction (Table 7.3, Column p). DIF was tested with Bonferroni adjustment for sex and age after re-scoring and removing items, indicating non-significance. Analysis of the individual person-fit indicated that the pattern of responses from five participants for Prototype 1 presented high residuals, five participants for Prototype 3, four for Prototype 4 and five participants for Prototype 5 (Table 7.3, Column n).

The model's assumption of unidimensionality was met through a binomial test for the calibrated scales, which indicated that less than or equal to 5% of observations were expected to fall outside of the t-range of ±1.96 in the confidence interval for every stimulus (Table 7.3, Column <95%CI).

Table 7.4 – Analysis individual of items and stimuli before collapsing to four categories.

Item	Reversed	Re-scored				
		St1	St2	St3	St4	St5
I1	Y	Y		Y		
I2						
I3						Y
I4			Y			
I5						
I6		Y	Y	Y	Y	Y
I7	Y	Y	Y	Y	Y	Y
I8		Y	Y	Y	Y	Y
I9	Y	Y	Y	Y	Y	Y
I10						Y
I11	Y					
I12			Y	Y	Y	Y
I13		Y	Y	Y	Y	Y
I14	Y	Y			Y	Y

7.5.3 Establishing a Metric Using the Faceted Rasch Approach

Data from affective responses to the prototypes were analysed through the faceted Rasch approach. Thus, the 14-item set was replicated for the five container prototypes yielding 70 items. The scoring system of every item was once again tested for disordered thresholds patterns. The whole sample of 67 participants was initially used in the MFRM; nevertheless, five persons were afterwards removed during analysis because they presented high fit-residuals. Individual item-fit residuals indicated no critical values (i.e., residual >±2.50). Invariance across the measurement structure was indicated by a non-significant item-trait interaction of p =0.51. The PSI of 0.90 indicated that the metric presented enough power for distinction between four groups of persons (Fischer, 1992).

The summary of the facet locations is presented in Table 7.5. An arbitrary zero was established as the default of the method applied in the analysis. The default for the origin constrained the stimuli facet[36] and the items facet at the centre of the logit scale. That is, both facets had measurement mean of zero. Also, the sum of the category coefficients was constrained to zero. Thus, solely the person facet floated on the continuum.

Table 7.5 – Fit statistics of Facet approach.

Stimulus	Location x(-1)	*SE*	Items	Mean location	*SE*	Chi - Square	Df	*p*	PSI
ST1	0.89	0.20	I14	-1.32	0.19	138.83	140	0.51	0.90
ST3	0.44	0.20	I1	-0.56	0.20				
ST2	0.23	0.20	I12	-0.28	0.20				
ST5	-0.60	0.18	I10	-0.15	0.20				
ST4	-0.96	0.20	I3	-0.15	0.18				
			I8	-0.09	0.20				
			I9	-0.09	0.21				
			I2	-0.07	0.18				
			I4	-0.03	0.19				
			I11	0.20	0.16				
			I13	0.31	0.20				
			I6	0.34	0.21				
			I5	0.86	0.18				
			I7	1.02	0.21				

Note: Locations and *SE*s in logits.

[36] The locations of stimuli obtained through RUMM2030® were multiplied by minus one. The reason was that the software package did not allow other configurations of the facets signal but that usually used for fairness of judgement. Thus, the stimulus locations are presented in a proper magnitude although they are placed at the reversed side on the continuum.

The metric is represented by the map of relative locations of all facets on the same logit scale (Figure 7.3). Person locations were plotted on the scale and represented in Column *Facet 1*. Participant locations that indicate more inclination to endorse the attribute of the container for giving an impression of a moisturizer are plotted on the top of the scale and those less inclined to endorse at the bottom.

The top of Column *Facet 2* of the map indicates items with more difficulty of endorsement. The Column *Facet 3* represents the location of stimuli on the continuum, indicating that the stimuli at the bottom of the scale are less likely to be endorsed as a container of a moisturizer, according to participants' impressions. It is noteworthy that the location of the stimuli does not follow the order of the physical measurement of compliance. Figure 7.4 indicates that the threshold distribution is widely spread, revealing that the respondents are well targeted to the set of calibrated items.

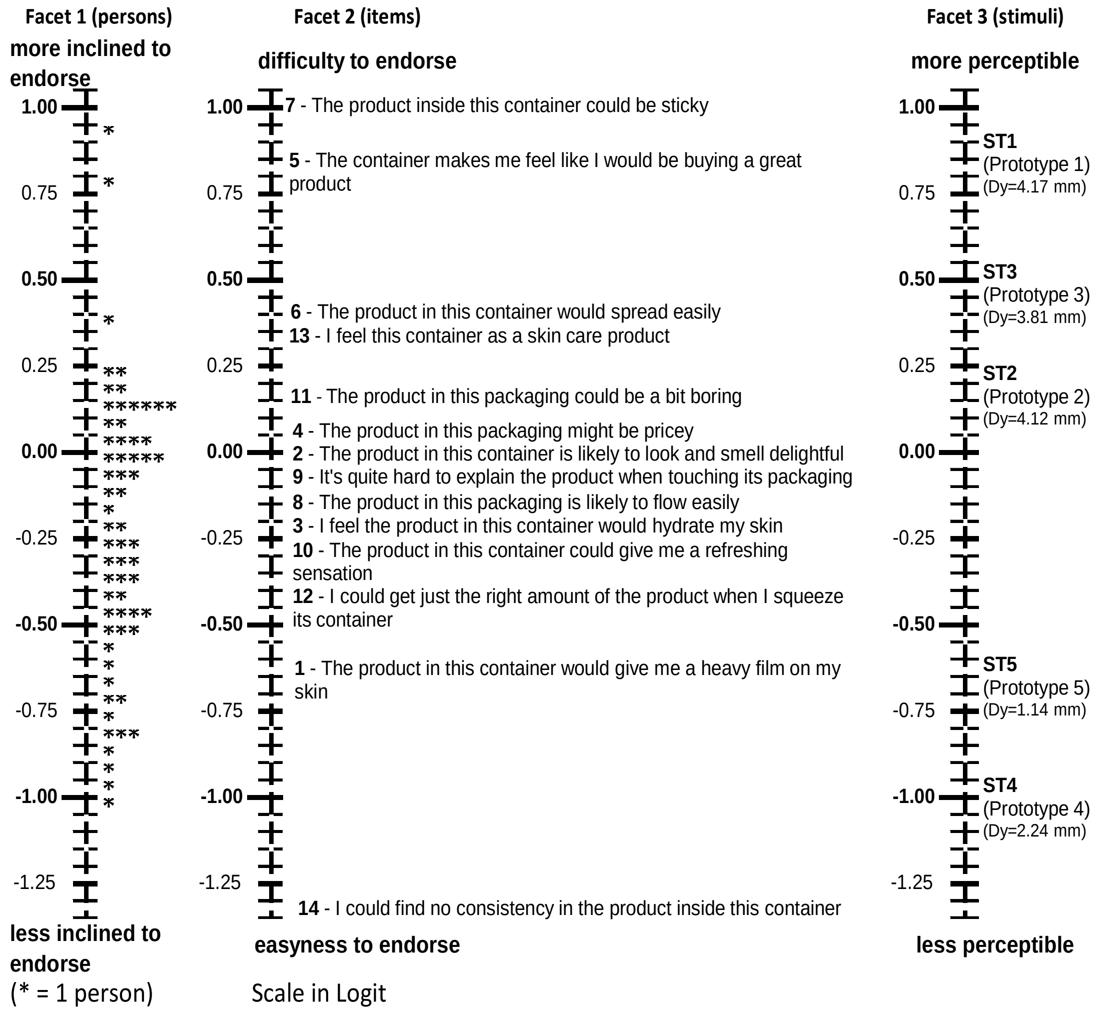


Figure 7.3 - Metric for the affective responses to the container prototypes.

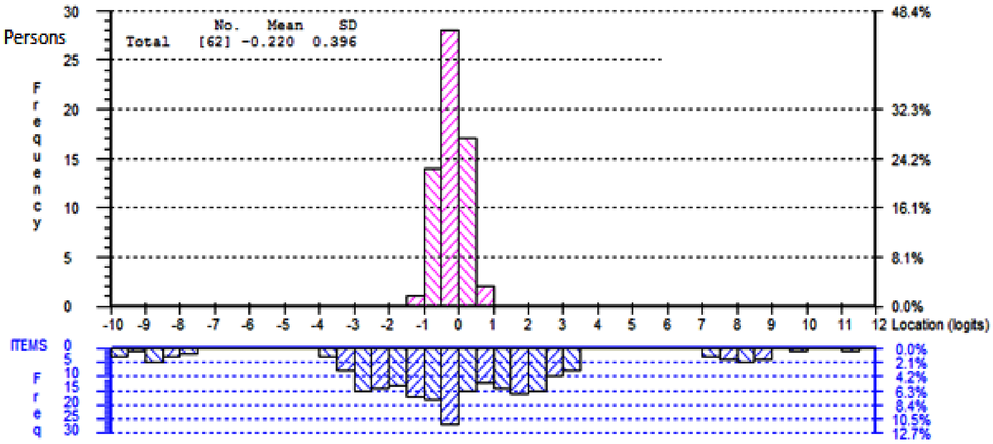


Figure 7.4 - Person-item threshold distribution using faceted Rasch model.

7.5.3.1 Differential stimuli functioning

A test for differential stimuli functioning (DSF) was carried out by comparing the enclosed area between SCCs and an imparity criterion (Figure 7.5). The areas were obtained by integrating the difference of the polynomials that define the SCCs (see Section 4.4.3). If two SCCs crossed over each other, then the reference point was the intersection of those two curves projected on the logit location axis (Table 7.6).

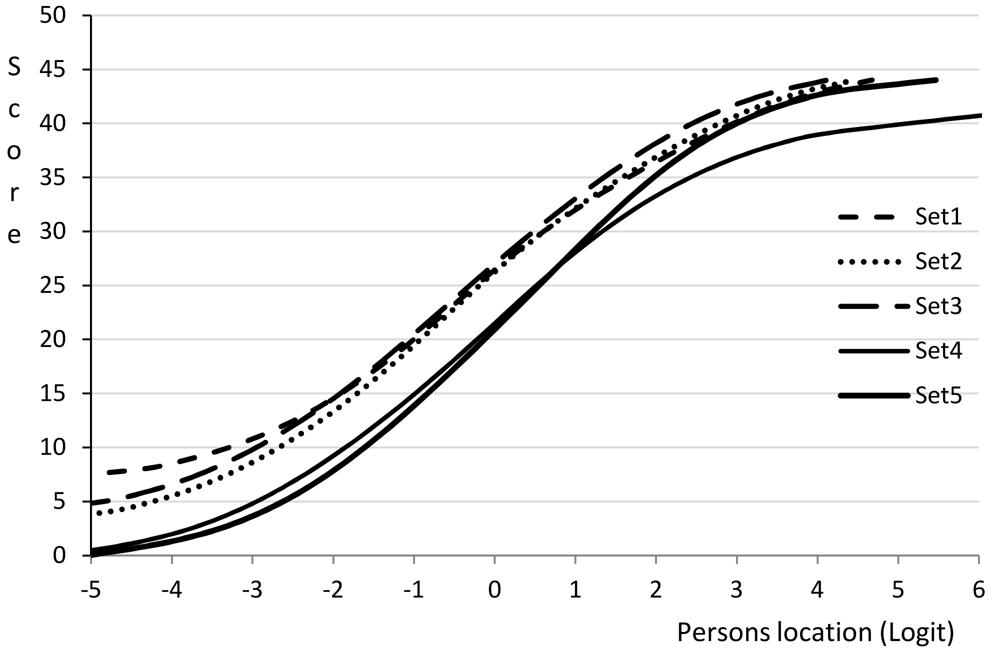


Figure 7.5 - Stimulus characteristics curves (SCCs).

If the difference was greater than zero, then the remaining area enclosed between SCCs was represented by the weak order $\succ$, such that upper SCC $\succ$ lower SCC. On the other hand, if the difference was less than zero then lower SCC $\succ$ upper SCC. Both of the cases indicated relative preference in a pairwise comparison.

Table 7.6 – Intersection points of the SCCs.

Stimulus	Prototype 1	Prototype 2	Prototype 3	Prototype 4	Prototype 5
ST1	x	x	x	x	x
ST2	-1.39	x	x	x	x
ST3	-2.62	-	x	x	x
ST4	-	-	-	x	x
ST5	2.65	-	-	0.66	x

The pairwise comparisons between stimuli through the enclosed area (A) contrasting with the enclosed area that included the measurement error (Ae) indicated that some of the stimuli did not present significant difference (Table 7.7). Looking at the facets map (Figure 7.3) and the comparisons of areas (Table 7.7) the affective attribute of the stimuli could be indicated as follows: $St1 \approx St3 \approx St2 \succ St5 \approx St4$.

Table 7.7 – Comparison of areas between SCCs and areas between SCCs along with errors.

	St1		St2		St3		St4		St5	
	A	Ae	A	Ae	A	Ae	A	Ae	A	Ae
St1	x	x	x	x	x	x	x	x	x	x
St2	**1.41**	**13.14**	x	x	x	x	x	x	x	x
St3	**8.57**	**13.04**	**9.99**	**13.88**	x	x	x	x	x	x
St4	45.43	11.10	44.02	11.94	54.01	11.84	x	x	x	x
St5	34.59	13.60	33.17	14.44	43.16	14.35	**10.85**	**12.41**	x	x

7.5.4 Sensory Responses when Squeezing the Product Containers

The values of force applied on the everyday product containers were taken from Table 6.17. The values considered were those without the influential cases flagged by the diagnostic mechanisms, which demonstrated a significant effect in the error variance.

The force applied by participants on each prototype is indicated in Table 7.8. The potential influential cases were identified by taking the values of D_i ≥0.06, CVR_i ≤0.91 and $DFFit_i$ ≥0.24 for n=67.

Table 7.8 - Mean and standard error of the force applied on the prototypes without influential cases.

Stimulus	Prototype 1	Prototype 2	Prototype 3	Prototype 4	Prototype 5
Mean force (N)	22.3	22.6	23.5	26.1	37.4
SE (N)	2.8	3.1	3.3	3.3	6.3

7.6 REMARKS

7.6.1 Comparisons between Areas and the Modelled Correspondence

Based on the comparisons between the areas established in Table 7.7, it is possible to draw the conclusion that there is no significant difference amongst Prototypes 1, 2 and 3 with regard to the affective attribute. There is also no significant difference between Prototypes 4 and 5. Nevertheless, according to the approach proposed in Section 4.4.3, there is difference between Group 1 that embraces Prototypes 1, 2 and 3, and Group 2 that contains Prototypes 4 and 5. This is not unexpected because in the modelled correspondence represented by Figure 7.2 the prototypes of Group 1 are located within the minimal interval of the lower and the upper bounds for the higher probability of endorsement of the affective attribute. Likewise, the prototypes belonging to Group 2 are out of the minimal interval, indicating lower probability of endorsement.

7.6.2 Concluding Comments

The affective responses were associated with the compliances of the existing products and of the container prototypes in a linear interval represented by the measurement scales. As a property of the RM the comparison between any two persons on the continuum is independent of the comparison between any pair of items and any pair of stimulus objects. This allows the assumption, for example, that the probability of a person at location 0.25 in Figure 7.4 to endorse Item 16 is higher than Item 7. Similarly, this person shows more readiness to endorse Prototype 1 because it presents more of the characteristic of a moisturizer cream than Prototype 4 does, based on a difference expressed in logits. The same is valid for a person at location -1.00 on the same continuum because the comparisons are invariant. On the other hand, the person at location 0.25 is more inclined to endorse the attribute than the person at location -1.00. Such individual comparisons in logits can be converted into probabilities through Equation 4.1. Furthermore, the locations in logits can be converted into scores and therefore, they can reliably be used in standard statistics for further analysis.

Evidence of the hypothesis that the compliances modelled for particular affective responses can specify features of new containers is given by the comparisons between the container prototypes. The scale represented in Figure 7.3 indicates that the Prototypes 1, 2 and 3, which are at the upper part of the scale, are in a range of displacement at a force of 3N between 3.81mm and 4.17mm. Observing the upper part of Figure 7.2 the range is within the lower and upper bounds of compliance for a favourable response to the affective attribute indicated by the locations in logit of existing products. The scale in Figure 7.3 also points to a range of displacement between 1.14mm and 2.24mm for Prototypes 4 and 5, which are at the lower part of the scale. Similar analysis in Figure 7.2 indicates Prototypes 4 and 5 are out of lower bound indicating less probability of a favourable response to the attribute.

Nevertheless, the range of force applied on the prototypes was higher than on the existing products. The lower part of Figure 7.2 indicates that the range of force applied on the existing products for a higher probability of a favourable affective response is between 11.7N and 18.2N. For the prototypes the force lies between 19.5N and 26.8N (i.e., force plus error), out of the range of the modelled compliances (Table 7.8). This might be a consequence of the influence of other factors than the containers' compliance when the persons processed the sensory information. The containers' characteristics of shape, for example, might have been combined with the characteristic of compliance for the respondents to form clear mental representations based on the tactile sensory information (d'Astous and Kamau 2010). Therefore, the participants might have had fewer hints about the impression of a moisturizer cream when they squeezed the prototypes than the existing products although this can vary from person to person. However, this assumption requires further investigation.

CHAPTER 8

Discussion

In the chapter, it is argued that although the comparison of relative precision amongst the different instruments produced in the previous chapters can be carried out, there has been no external reference value for comparing their accuracy as a consequence of the methodology used in affective engineering. Additionally, the problematic matter in the domain about the application of models for data reduction is demonstrated through the empirical data obtained during the research. Although the rationale developed throughout the research using the Rasch model can overcome most of the problems of measuring latent variables in the domain, there are limitations for predicting unknown observations. Nevertheless, important implications of the findings with regard to stability of the instruments and the incorporation of new variables to a calibrated core are explored in the chapter, such as item banks and computerized adaptive testing.

8.1 PRECISION AND ACCURACY OF THE MEASUREMENT SCALES

Quality of a measurement is highly associated with the precision and accuracy of the measurement instrument. The scale's precision indicates to what extent the measurement instrument agrees with itself and accuracy refers to the closeness of agreement of the results with an established value (ISO 5725, 1994; VIM, 2012).

The standard error (SE) of measurement has been referred to as the precision of an instrument for latent variables (Wright, 1995). The imprecision is associated with the misfit of data to the model[37]. Nevertheless, fit statistics in Rasch modelling has a stochastic component, i.e., the overall absence of errors leads to the deterministic form of the Guttman pattern (see Section 2.4.7.2), decreasing the instrument's precision. Thus, error is not unexpected in Rasch modelling. In other words, measurement error allows generalising the instrument in a determined context. Taking into account that after calibration (i.e., with data fitted the model) the measurement error is inherently a consequence of the modelled probability, the basis of comparison of instruments shall be made on the smaller standard error value.

[37] It is noteworthy that when using a very small set of items and small sample size, the model standard error can be far different of the actual standard error around the expected estimate measure.

To illustrate the discussion, Figure 8.1 shows the comparison of the 18 instruments obtained from the empirical studies in Chapter 6 and Chapter 7. Five instruments were obtained from co-calibrations of scales for the five existing product containers used as stimuli, all of them containing 11 items. The instrument with the same 11-item set calibrated through the multi-facet Rasch model was also compared. The comparison included five other instruments co-calibrated with 14 items for the five existing product containers and the scale calibrated through the faceted approach for the same 14-item set. Finally, the five instruments individually co-calibrated for the five prototypes using the 14-item set and the scale calibrated through the faceted approach were further included in the comparison. Comparison of the scales' precision was established by the root mean square error (*RMSE*) for each instrument, taking a lower index as relatively better (Linacre, 2005), such that

$$RMSE_{\hat{\delta}} = \sqrt{\frac{\sum_{i=1}^{I}\left(\delta_i - \hat{\delta}_i\right)^2}{I}} \tag{8.1}$$

where δ_i is the estimate for item i and $\hat{\delta}_i$ is the expected estimate for item *i*, with $i \in \{1,\ldots,I\}$. The *RMSE* of the person estimates was similarly computed, given that

$$RMSE_{\hat{\beta}} = \sqrt{\frac{\sum_{n=1}^{N}\left(\beta_n - \hat{\beta}_n\right)^2}{N}} \tag{8.2}$$

where β_n is the estimate for person n and $\hat{\beta}_n$ is the expected estimate for person *n*, with $n \in \{1,\ldots,N\}$.

Observing Figure 8.1 it is possible to identify that the scales varied in precision according to each stimulus. The individual calibrations using 11 items prevailed over the calibrations with 14 items with regard to the precision of person estimates. Furthermore, the calibration through the faceted approach resulting in an 11-item set pointed to a lower *RMSE* for the item estimates than that with 14-item set although it is unclear the improvement of precision for person estimates. These observations have suggested that the incorporation of items in a scale did not improve the precision of the scales necessarily even though the distinction of containers' characteristics was improved (see Section 6.6.7.2). One of the reasons is that there was a lower discrimination between items when using the 14-item set. That is, items clustered in the central location of the continuum and therefore, a better spread of items should be sought.

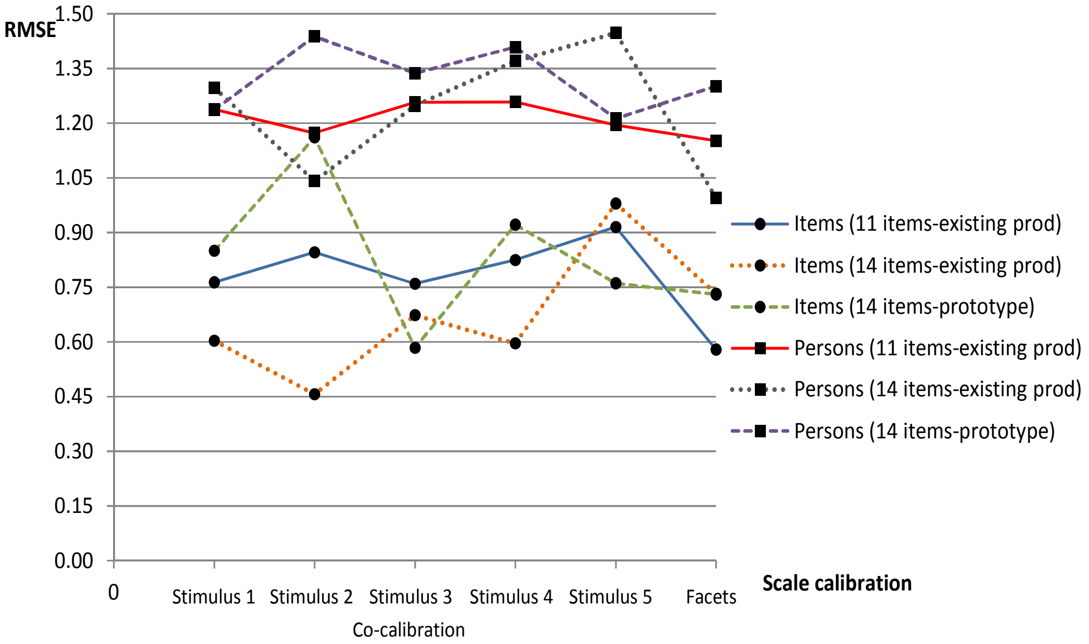


Figure 8.1 – Comparison amongst values of root mean square error (*RMSE*) of the co-calibrated scales and of the facet-calibrated scales using the 11-item set and the 14-item set for the existing product containers (see Chapter 6) and the 14-item set for the product container prototypes (see Chapter 7).

Precision of an instrument for measuring latent variables is not different from those to measure physical objects. The necessary precision shall be established by the context of measurement and by the expected outcomes. For this research, for example, the scales were precise enough to demonstrate the rationale for developing measurement instruments for latent variables in AE and for incorporating new variables in the instrument without loss of comparability. However, a necessary refinement for the scales should be carried out to improve precision if their application was demanded in another context.

Whereas precision is concerned with a relative value of comparison, accuracy, on the contrary, is not quantifiable (VIM, 2012). In education and social sciences accuracy has been associated to the goodness-of-fit of a model to the data or contrariwise in the case of the RM. This could be an approximation of the definition used in metrology where measurement accuracy is understood as a measurement that presents the smaller error. Accuracy is associated with the concepts of measurement precision and measurement trueness (ISO 5725, 1994). The latter represents the agreement between the average of an infinite number of replicate measured values and a reference value (VIM, 2012). However, in AE constructs emerge empirically from the multivariate analysis, rather than being prescribed. As a result, some variables that form the construct can mistakenly be exemplified by differences in degree. This prevents the establishment of values of reference and therefore, measurement trueness is impaired. This is not a consequence of the mathematical model used in an analysis; rather it is a flawed characteristic of the current methodologies in the domain to obtain a clear definition of the relevant latent variable.

8.2 REGRESSION MODELS AND THE RASCH MODEL

It is worth restating that the RM and multivariate approaches based on classical test theory belong to different paradigms and therefore, analysis of data requires a distinct interpretation. This research has not focused on comparing different statistical approaches using the same data set. The interpretation of data in this thesis is associated with measurement theory where a number of approaches have not met its principles (Wright, 1996; Salzberger, 2013). If a quantitative structure is established, statistical approaches could be employed in further analysis. Therefore, the RM does not disqualify statistical methods used in the domain, but it provides evidence of whether or not a data set holds a quantitative structure that allows mathematical operations. The most typical statistical approaches for data reduction in the domain are principal component analysis (PCA) and quantification theory type I and II (Hayashi, 1952) (see Table 2.1), which are methods used in KE similar to regression analysis and factor analysis (FA). Nevertheless, the drawbacks of those approaches for eliciting affective responses have been recognised since the middle of 1990s (see Section 2.2.7). Accordingly, a more advanced understanding on the complexity of affective data has recently addressed the treatment of uncertainty through applications of rough set theory.

8.2.1 Factor Analysis and Principal Component Analysis

The analysis of correlations between observed variables in typical applications of FA and PCA in affective engineering can lead to a misrepresentative description of the relationships amongst those variables. FA and PCA explain the relationship through analysis of correlation or covariance that can be modelled with a straight line similarly to multivariate linear regression models with observed, continuous test scores used as dependent variables and latent factors as independent variables (Spearman, 1904; Thurstone, 1947). However, such as in any regression model, the precision of results is conditioned to violations of statistical assumptions and anomalies in a data set. The linearity assumption of FA and PCA is, for example, necessarily violated when the common factor model is fitted to Pearson’s product-moment correlations amongst categorical, ordinal scaled items, including Likert-style scales (Flora et al., 2012). Furthermore, the usual procedure of rotating axes using varimax embodied into the principal components causes the distribution of data in the semantic space to force the factors to be orthogonal and the sum of squared slopes to be unity although these conditions might be untrue. Consequently, varimax or any other standard rotations incorporate unrealistic conditions in general and potentially suggest misleading conclusions (Swain, 1979).

As an illustration of this problem in AE, data reduction using PCA was carried out for responses from two different groups of persons (see Appendix C). The scores obtained from the affective responses to 24 items for the four confectionery stimuli previously employed in Rasch analysis (see Chapter 3) were also used in this analysis. The sample of 306 respondents was divided into two groups of 153 persons. After rotating axes using the varimax method, the loads for the first half and for the second half of the sample were compared. The results indicated that 14 out of 24 items analysed changed their loads for a different component when comparing the first half of the sample against the item loads of the second half of the sample (Table 8.1). This clearly suggests that the item loads for the components are sample-dependent. That is, if an affective engineering study is replicated and its data reduction method is based on PCA (or FA), the composition of the principal components will very likely differ to some important degree.

The differences between the RM and FA (or PCA) have largely been discussed in the literature of measurement of latent variables (Wright, 1996; Kyngdon, 2004). For example, the concept of measurement of FA (or PCA) is based on numbers assigned to persons' responses. In Rasch modelling measurement requires a process for establishing ratios,

Table 8.1 – Excerpt of the rotated component matrix for the first half and for the second half of the sample from the empirical study reported in Chapter 3 (only cases that changed component).

Variables	First half				Second half			
	Component				Component			
	1	2	3	4	1	2	3	4
Item 2	.099	***.755***	.150	-.097	***.741***	.052	.119	.106
Item 4	***.714***	.092	.181	-.146	.132	***.838***	.107	.089
Item 5	.251	.063	-.247	***.611***	.176	***.517***	.104	.392
Item 7	-.305	-.048	***.636***	-.074	.182	-.282	.152	***-.678***
Item 8	***.670***	.272	-.174	.158	.462	***.478***	-.013	.231
Item 9	***.681***	.426	.117	.020	.499	***.619***	-.006	.138
Item 10	***.793***	.123	-.091	.048	.278	***.712***	.129	-.087
Item 13	.397	***.645***	.144	.077	***.599***	.005	.226	.114
Item 14	.385	***.660***	.206	-.121	***.584***	.335	.156	-.087
Item 15	-.005	***.780***	-.113	.209	***.758***	.025	-.090	.017
Item 17	***.740***	.114	-.154	-.008	.433	***.574***	.279	-.101
Item 18	-.477	-.126	.087	***.593***	-.029	***-.570***	.103	.079
Item 19	.511	***.539***	.030	-.060	***.633***	.274	.046	-.300
Item 21	.326	***.627***	.103	.015	***.665***	.282	-.046	-.288

Notes: Loads after varimax rotation (component decision in bold). Full matrix is displayed in Appendix C.

rather than assigning numbers. Furthermore, FA (PCA) and the RM approaches are different with regard to the interpretation of person measures. The first makes reference to a sample mean with the weighted raw score considered to be a linear measure, which is directly related to the indicators. Differently, the RM refers to items that define the latent dimension and therefore, the raw score is not considered to be a linear measure. Thus, a transformation of raw scores into logits is necessary (Wright, 1996). Another difference between the approaches is that FA and PCA parameters are dependent on the sample. Therefore, representative samples of a population are essential. In the RM, item parameters are independent of the sample used although subject to model fit and sufficient targeting[38] (Ewing et al., 2005).

8.2.2 Rough Set Theory

In KE rough set theory has currently been considered as the most effective approach for the analysis of observations that can contain ambiguity, such as sensory and affective data (Nagamachi, 2008)(Table 2.1). According to Okamoto et al. (2007) rough sets allow the derivation of more specific decision rules than available tools of statistical regression analysis in KE. Rough set theory, pioneered by Pawlak (1991), has been developed as mathematical models to overcome uncertainty and inconsistent data. One of the main applications is the extraction of decision rules that can overcome the problem of sensory and affective data characterised by lack of linearity.

Rough set theory is related to the notion of imprecise concepts, called vagueness, such as the concept of beauty. This concept of vagueness is associated with the existence of objects characterised by the same information although they are indiscernible from the viewpoint of that available information. One of the benefits that has been claimed by the theory’s advocates is that analysis of data does not need any preliminary or additional information, such as probability distribution in statistics. This allows the model’s algorithms to find patterns in a data set to establish data reduction (Pawlak and Skowron, 2007). Rough sets are defined by approximations. The lower approximation of rough set comprises elements that belong to the available information. The upper approximation contains elements that have the likelihood of belonging to the set with regard to the available information (Pawlak, 1991).

[38] Shumacker and Linacre (1996), Ewing et al. (2009), Aryadoust (2009) and Wright (1991, 2000) have pointed to further differences.

Although rough set theory has been applied in different domains such as data mining, expert systems, pattern recognition and clinical diagnosis, it presents characteristics that could prevent reliable interpretations of data in AE. Hu et al. (2004) stated that rough set theory considers neither the statistical distribution nor the variance of the data when defining the lower approximation. Furthermore, if there are only few elements of lower approximation, the decision rules extracted from these few elements might be unreliable[39] (Nichino et al., 2005). Variance in a data set is one of the key concepts in measurement theory, as Allan (1987) stated "*since a measurement is no better than its uncertainty, specifying the uncertainty is a very important part of metrology*." As such, if the purpose of a study is to measure individual differences, then ignoring the effect of variance can yield flawed results for the uncertainties and draw mistaken interpretations. Differently, the RM deals with latent variables as potential objective measures. If such variables are indiscernible from the viewpoint of the available information, such as it is considered in rough set theory, then they are not elements of a measurement system.

8.3 PREDICTABILITY OF UNKOWN OBSERVATIONS

The main assumption in this thesis has been that if data fit the RM, ordinal raw scores are turned into objective linear measures. This assumption is, however, based on observed data. It is not clear in AE how the model would behave for unknown data from responses to idiosyncrasies of new sets of stimulus objects.

Comparisons between different calibrations demonstrated that the measures obtained from the scales are stable enough to denote them as instruments of measurement (see Chapter 6). Furthermore, evidence for the generalisation of the scale to a new set of similar stimulus objects has been given in Chapter 7. A straightforward interpretation of those results could lead one to claim anticipated benefits to any data set. However, such a generalisation should be drawn cautiously.

The Rasch-measures for the containers, based on their modelled compliance, allowed the establishment of a new scale to compare the container prototypes because they were designed and manufactured to fit within the measurement range obtained from the calibration of everyday products. The calibration was carried out in a controlled condition for detecting anomalies in data obtained in the empirical study. A different

[39] Nichino et al. (2005) have proposed a probabilistic approximation based on information gains of equivalent classes to overcome the problem. However, the theoretical foundation of his approach is not clear.

situation could require predicting ratings to magnitudes of characteristics that fall out of the modelled range. Another situation could require extending the scale based on responses from a small sample to a few stimulus objects to a larger number of objects. For both cases the results of the empirical studies in this thesis do not clarify whether or not ratings might precisely be predicted on the basis of the observed data.

However, the problem is not solved by other current analysis methods either. The condition established in current methods of AE is that a model ought to fit the available data where the smaller the divergence, the better the model. However, a problem arises when the model fit the data set very well (referred to as overfit). This means that the model has been conformed in the particularities of the sample. Therefore, the model would likely fit a different sample poorly. Overfit in Rasch analysis is consequence of redundancy, i.e., over-predictability (Walter, 2008). However, in Rasch analysis it is not unexpected that new estimates be more dispersed than the model predicts for them. The effects from anomalies are usually difficult to predict because of the singularities of each sample and stimulus object. Therefore, variance resulting from DIF and local dependence should carefully be investigated for the new data set.

More elaborate approaches to predicting ratings have been suggested by Linacre (2010). One of the approaches is the Boltzmann machines (Ackley et al., 1985). According to Linacre, the model can interact probabilistically with the unidimensional characteristic of the RM. Kastrin and Peterlin (2010) have proposed to use the RM to reduce dimensionality in micro array data in the domain of machining learning, where the number of variables is very large compared to the number of observations. Another possibility could be a data-mining technique associated with the RM. Typically the evaluation metric influences the feature selection algorithm, such as in rough set theory. Therefore, the lower approximation of rough sets (see Section 8.2.2) might be established by elements obtained from the logit Rasch-calibrated scale, solving the problem of statistical distribution and variance control.

8.4 ITEM BANKING AND COMPUTERIZED ADAPTIVE TESTING

One potential practical value from a frame of reference for measuring an underlying attribute of a product using Rasch theory in the domain of product design is the construction of an item bank. The empirical approach, reported in Section 6.6.7.2, inserted three items after a second calibration within a calibrated, 11-item scale to measure persons' impression of a moisturizer cream for a set of five everyday product containers. The outcome represented a better differentiation amongst the containers' characteristics of compliance. Furthermore, the major outcome from part of the empirical study is, perhaps, the potential of incorporating new variables in calibrated measurement structures that vary in a range of difficulty levels of endorsement for different persons' inclinations of endorsement. This supports the development of an item bank. The approach follows theoretical propositions and successful applications in the fields of education and health sciences (Chopin, 1978; Wright and Bell, 1984; Hahn et al, 2006; Eckes, 2011). An anticipated benefit of this approach is that analysts can develop a bespoken structure with additional items without losing the properties of the core of the original, off-the-shelf calibrated scale to make whatever general comparisons they require.

As a result of the item bank approach, the development of the computerized adaptive testing (CAT) is possible. It is through item banking and CAT that the financial benefits of using Rasch theory in AE can be realised. The concept of the CAT is concerned with establishing a sequence of items (i.e., adjectives, statements or questions) that seem most appropriate for a particular respondent. Items are selected through a computer such that if a respondent endorses an item, a slightly more challenging item for endorsement is automatically presented in the sequence, and contrariwise if the item is too difficult. This technique usually converges into a sequence of items bracketing and convey information on the respondent's effective endorsement level. Consequently, each respondent does not answer all statements in the item bank, only a subset bracketing the threshold of endorsement. The technique is well-developed in the field of education (Weiss, 1984) and applications in health sciences have frequently been published, such as in clinical diagnostics, where it has helped to assess patients through fewer items along with higher measurement precision (Elhan et al., 2008). In the domain of product design CAT using Rasch-calibrated measurement structures is potentially useful to reduce cost in consumer research because it allows using small samples and offers the advantages of convenience to respondents concerning flexible scheduling, improved security and data collection. Nevertheless, research on this topic in the domain of product design is still unknown and therefore, its application will require further investigation.

8.5 ALTERNATIVE STRATEGIES FOR CALIBRATION OF SCALES

The rationale to establish reliable measurement instruments developed throughout the thesis follows the principles of Rasch measurement theory. As seen in this research, Rasch theory provides mechanisms to test the construct to identify potential sources of anomalies in a data set used for calibration of instruments. Additional tests were proposed in Chapter

4 and Chapter 5 which although they can be applied in different domains, their purpose is to strengthen the measurement instruments of latent variables for applications in human-centred product design.

Nevertheless, there are different possible solutions to apply the RM in the domain. This is the case, for example, when different measurement constructs were found from the results of the experiment with the pieces of confectionery. One alternative solution could have been to establish equations for transformation of measures from one scale into other one and thus, to compare the chocolates. The solution would be similar to transform degrees Fahrenheit into degrees Celsius, as a metaphoric example. In the confectionery case, all the scales were measuring the same affective attribute specialness in unidimensional spaces. The difference amongst them is just their unit, which could be converted one into another. However, this might input more drawbacks than benefits when lecturing on the topic as a consequence of the unnecessarily complicated algebraic operations. It seems that the approach adopted in this thesis can offer a better understanding on the assumptions of RMT and its meticulous procedures to meet them.

It is noteworthy that what has been sought in the research is to contribute with a rationale to develop instruments for latent variables in the domain. Accordingly, there could be different strategies to calibrate the scales, maintaining the validity of the proposed rationale. These strategies could take place based on the Rasch tradition in a determined area of research. Furthermore, the software package used for analysis can induce certain strategies although their results should be similar. In most of the cases the software package FACETS® and alternatively the program ConQuest® have been established as preferences by Rasch-facets analysts. Nevertheless, the reason to foster the strategy reported throughout the thesis is the adoption of the computer program RUMM2030®. This Rasch-dedicated software package has shown a number of advantages, including its friendly user-interface and straightforward demonstration of results. However, the use the faceted structure in the program has been little explored by users as a consequence of its limitations. For this reason, the use of the program at the initial stages of the research drove a different solution for adopting the faceted approach. Thus, each stimulus was separately calibrated, followed by a co-calibration for adjusting a common pool of items. Finally, all the stimuli were re-calibrated in a faceted structure (see Chapter 4). On the other hand, the limitations at the beginning of the work turned into an insightful strategy which provided a window for understanding the behaviour of data and of the model. Nevertheless, the applications reported in the thesis have contributed to improve the facets feature in RUMM2030®. Modifications have recently been carried out by RUMM laboratory. This will allow a more straightforward strategy to calibrate instruments, renouncing the time-consuming, many-calibration procedure adopted by this research, upholding the proposed rationale, though.

8.6 ADDING VALUE TO THE AFFECTIVE ENGINEERING PROCESS

Eliciting information from affective responses to physical characteristics of products and transforming it into exploitable data can be an intricate process when formalising it in typical industrial product development processes. The interpretation of users' impressions and the establishment of requirements based on them throughout the process can differ as a consequence of different perspectives from a frequently variety of stakeholders. Thus, if the interest of an analyst lies on knowing individual levels of affective interaction independently of the scale used and the stakeholders who are assessing it, then a formal measurement of the latent variable using the Rasch model can be necessary.

One of the benefits for product developers from the applications of the RM in the AE process is to avoid a number of pilot studies using individual items in test groups, reducing costs of consumer research. In product development many trials are usually required to qualify the characteristics of a set of prototypes through groups of consumers. However, respondents can become disinterested, distracted and fatigued as a consequence of a high number of items and physical stimuli used in the trials, influencing the degree of uncertainty of the results. Furthermore, large samples (N>100) are necessary to obtain stable estimates in each trial. The application of the RM in this research suggests that a smaller sample could be used after the calibration of a measurement scale. Calibrated scales for objective measurement in AE allow the diagnosis of the best physical characteristic that fulfils the relevant affective attribute. The comparison between the costs attached to the calibration of a scale in AE and the current methods used for qualifying characteristics needs to be balanced against the cost of low measurement precision in each case of application.

Another characteristic associated with the design practice has been the improvement of existing features to add value to products (Cross, 2000). Values are associated with the attitude of users and the interpretation of them. Attitude can fluctuate according to social and cultural oscillations, technological advancements (see Section 1.1) and contextual conditions of each industrial application. The importance of context when experiencing materials has been related to persons' attitude that varies according to different stimulus products and stimulus sentences as well as environmental conditions (Chen et al., 2009). Such fluctuations, therefore, produce data from affective responses that could not be statistically stable for time enough and with different groups of persons to be utilisable for generalisation of the outcomes. Nevertheless, some works in AE seem to reside in a false sense of objective interpretation when using statistical reasoning, making hasty generalisations based on that interpretation (e.g., Hirata et al., 2004, Ogino, 2012). On the other hand, in many situations a latent regression model based on RMT will make more intelligible the associations between variables (Christensen, 2006) allowing afterwards the application of scores transformed by the RM into a diversity of statistical tools.

The RM adds mechanisms to the AE process that allow the validation of the structure of data from affective responses under certain empirical conditions. Such conditions can be seen through the works of Campbel (1920, 1928), Luce and Tukey (1964) and Suppes (2009) (see Sections 2.3 and 2.4.1), associated with the structure of numbers under corresponding algebraic operations. The most important perhaps, the RM can mathematically support the interpretation of the test scores and, as a consequence, to support the implications that the interpretation entails. On the contrary, if just the sum of scores was taken without any validation, the results would contain a high level of imprecision as an effect of measurement error and regression coefficients are likely to be attenuated. Thus, in affective engineering the RM can find the best number of independent variables in an instrument, maintaining the quality of measurement and avoiding problems with regard to the adverse effects of short instruments and small samples when using classical methods.

8.6.1 Implementation of the RM at Different Stages of the Design Process

Although affective engineering seems to be part of early stages of the design process, its incorporation into product development processes (e.g., preliminary and conceptual stages), has not formally been reported. Nevertheless, if affective engineering is taken into account in a project, opportunities for objective measurement of users' responses can be found in different tasks of formal or quasi-formal design processes (Table 8.2). Although Table 8.2 is not an exhaustive list of opportunities, it is possible to envisage many of the potential implementations of the RM in a design process.

Table 8.2 - Potential opportunities of implementation of the Rasch model in the product development process.

Opportunity during a design process	Implementation options
Project scope	Definition of a theory-driven coverage assessment for human-centred design.
Physical requirements	Measurement of latent responses to physical characteristics.
Performance requirements	Measurement and adjustment of the relationship between affective responses and performance.
Usability requirements	Measurement of latent variables in human-factors design.
Interface definition	Measurement and improvement of different interfaces and derivatives with regard to affective responses.
Labelling	Measurement of perception of meaning and comprehension of information.
Inter-changeability	Measurement of attitude toward the integration with other products or accessories of the same family.
Packaging	Measurement of attractiveness and relationship with the product.
Quality management	Measurement of perceived quality.
Reuse and refurbishment	Measurement of attitude toward reusable products.
Affective attenuation	Measurement of the level of affective attachment to a product or components of the product throughout its life-cycle.
Reference documentation	Traceability of calibration of scales for affective responses and affective requirements based on objective measures.
Decommissioning	Measurement of affective responses for premature decommissioning or disposal.
Installation and set-up	Measurement of the level of expectation with regard to temporary interruption of services, downtime or difficulty of set-up.

8.6.2 Cost-effectiveness Trade-offs

The approach using the RM and objective measurement itself in AE is novel. Thus, criteria for assessment of the value added to the AE process still require to be characterised. One of the criteria could be the trade-off between initial investment in the development of measurement scales for affective responses and how frequently they would be used throughout the life-cycle of a product. Other criteria to assess cost-effectiveness of the applications of objective measurement in affective engineering could be to what extent the novel approach using the RM would complement an industrial product development process and the time of carrying out the implementation of objective measurement.

The implementation of objective measurement in affective engineering can reduce costs of development by preventing rework or redesign of physical characteristics of products as a consequence of mismatch with what users' feel about them. However, the payoff of this effort might be one of difficult assessment. In private businesses, for example, quantitative data on how investment in the early stages of a product's development affects quality is in most cases labelled as *confidential* or *proprietary* information (Hooks and Farry, 2001). Another difficult assessment is the value added by a recording rationale, such as has been proposed in this thesis. Traceability of calibrations could, for example, reduce risks of users' acceptance of a derivative interface for a product and expose defective assumptions on the product features. Because the evolution of affective requirements can easily be followed, a derivative system can be compared against the original one by a same metric. Furthermore, communication between stakeholders when using affective engineering processes can be improved through the refinement of provided information by objective measurement using the RM.

Although cost-effectiveness still requires to be characterised when implementing the RM in design processes, relevant measurement will always lead to consistent requirements for the whole life cycle of a product. Objective measurement in AE can reduce costs that are originated from discrepancies between developers' concept of the attitude of consumers and the actual consumers' impressions. The implementation of the RM in AE will give finer levels of details with regard to human interactions with physical components. Unambiguous interpretation of such affective interactions is made possible only by objective measurement.

8.7 FURTHER STUDIES

A well-defined scale of measurement has potential applications in many design settings. Defining subpopulations according to their differences is a far more advantageous approach to manage affective attributes of a product than an entire population. Nevertheless, attitude to a product is frequently idiosyncratic (see Section 1.1). Therefore, persons' attitudes can vary in a period of time or can float according to the fluctuations of social characteristics. However, it is not trivial to understand whether the objectivity of the measurement structure is destabilised when the relative difference between item locations on the continuum varies over a period of time. The mathematical difference in estimates of item locations from a first empirical calibration and from a second calibration at a different point in time might, for example, be a consequence of a better familiarity with the product. This issue can be addressed through the comparison between two distinct measurements in certain period of time using different group of persons.

Another technical issue is associated with flexible content measurement structure, such as CAT, which might present some discrepancies for the aspects of individual differences. It is not clear, for example, whether a small set of items optimally selected to provide efficient information about a person's endorsement level represents all aspects of the attribute being measure as compared with the whole item bank. As a result, the characteristics of individual differences reflected from a subset of items can diverge to some extent according to the subset used. Furthermore, the context of a subset automatically generated in CAT may misrepresent the whole item bank. The subtest could, for example, be formed by items with a broader or a narrower context composition. This requires investigating whether inadequacies of model's fit in different calibrations may influence choices of subsets.

An important technical study is related to discrepancies associated with small sample size to develop assessment scales. The issue has generated a debate in the domain of health sciences because the US FDA has recommended preliminary small studies during the development of the patient-reported outcomes measures (PROMS). The reason is to minimize the risk that the instrument will not perform adequately in a new population and ascertain that the instrument holds measurement properties before establishing the content validity for labelling claims (FDA, 2009). PROMS instruments have been developed in a wide range of clinical pathways. However, when applying the RM to optimise clinical trials, the discussion lies on the calibration using small sample sizes that might lead to narrow the spread of items in the range of the continuum, preventing the generalisation to a broader population. However, this might be a consequence of the method to identify variables (or items) that ought to hold a scalar property, establishing a measurement structure. Further, the sample might not be targeted enough to represent differences in a relevant population. This technical issue of validation could be addressed in further investigation in the domain of AE although the findings could also be extrapolated to other domains of knowledge.

8.8 TRANSFERABLE SKILLS

Given the preceding evidence, it seems that there is a wide range of opportunities of investigation with regard to quantity values of latent variables using Rasch measurement theory in AE. The degree of knowledge on the subject depends on the purpose of study. For practical applications in students' projects or in more complex applications in industry settings and professional research, a variety of workshops has been growing every year, although focusing on different areas other than product design. The workshops are in general hands-on including a significant part on the theoretical background from basic statistics to probabilistic distribution and the principles that underpin Rasch measurement theory. Usually, the hands-on workshops vary according to the computer program used for analysis of data and are offered in different levels of information, in person or online, taking from a couple days to around a half year.

Following the statement by Andrich (2004) (see Section 1.4.2), the challenge in RMT is that the solution for measurement of latent variables should come from those who are involved in the relevant field of application. In product design this means aligning the physical elements with the user’s affective responses. Accordingly, application of the RM transfers the understanding of the statistical misfit and the validation of the measurement instruments from the analyst’s hands to the own designer or engineer. Perhaps, from this point of view, a complementary specialist workshop could be necessary in the domain.

CHAPTER 9

Conclusion

9.1 CURRENT PRACTICES OF QUANTITATIVE EVALUATION

Current practices in affective user experience have given evidence that there is an inconsistent application of assessment tools when eliciting users' responses to physical elements. One of the sources of inaccuracies is to assume that pairs of contrasting adjectives are true linguistic opposites. The precision of the method is associated with the degree to which this underlying assumption is fulfilled. However, this assumption has not been examined in the domain. A further problem with assessment tools is that scores are obtained by counting the ordinal position of the response possibilities in scales. This characteristic does not ensure by itself that data fulfil the assumption of interval property. Therefore, an analyst in AE ought to validate the assumption in every process of quantitative assessment.

Another issue raised in the research was that analyses in the domain have underestimated the influence of measurement errors. Differences between sex, age and cultural groups, misinterpretation, redundancy, ambiguity and other factors, for example, can be the source of systematic errors in a measurement instrument. In addition, to establish the relationship of variables expressed by qualitative dimensions, e.g., using PCA, one ought to take into account large and representative groups of persons because factors or dimensions are subject to the circumstances of sampling. Unrealistic conditions of linearity and rotation of axes can also suggest misleading interpretations. Thus, analysis of data should test the hypothesis that the empirical procedure is replicable at the same domain with rules of sampling not influenced by the dimensions previously found and that the same dimensional structure remains dominating.

These sources of inaccuracies when eliciting persons' affective interaction with design elements undermine the assumption that the relevant latent attribute of a product is quantitative. As a consequence, outcomes cannot be generalised.

9.2 APPLICATION OF RASCH MEASUREMENT THEORY IN THE DOMAIN

The main thesis pursued by the research is that Rasch measurement theory can overcome part of the inaccuracies in the domain. Rasch theory provided detailed procedures, called calibration, to identify anomalies in a data set which prevent the development of additive correspondences. Because data fitted the model, comparisons were made by the difference between the numbers associated with the persons' responses. As such, a particular difference is invariant, i.e., it has the same interpretation across a scale continuum. The RM's property of parameters separability allowed the development of scales that are independent of the sample of persons used to estimate item parameters and independent of the set of items used to obtain scale scores.

The RM has been vastly discussed and developed in education, health and social science. Nevertheless, the evolution of measurement of attitudes and the evolution of affective approaches to products have gone through parallel paths. Sophisticated solutions for the absence of linearity in affective and sensory information data have led to approaches that derive more specific decision rules, such as the rough set model, than statistical regression analysis frequently used in AE. However, such approaches have not held measurement properties. Other models, mainly belonging to IRT, present solutions that apparently offer better fit of the model although in detriment of measurement invariance. The stochastic framework of the RM is, therefore, the sole mathematical approach using the tools of standard statistics that fit data to the model, meeting the axioms of additive conjoint measurement to adjust non-physical objects to the concept of fundamental measurement.

However, because the purpose of the RM is to establish quantifiable relations of latent variables, the measures obtained through the model could not be necessarily a good description of the data. Part of the data obtained in the empirical approaches of the research fitted the model poorly. Consequently, a number of the preliminary items were removed from the analyses. Misfit was originated from the characteristic of lack of additivity in the data, rather than a problem associated with the model itself.

9.3 TEST OF THE RESEARCH HYPOTHESES

The measurement properties of the RM can be achieved if data fit the model. As such, the data obtained from affective responses to stimulus statements should be exemplified by differences in degree. However, in AE, the variables that represent the dispositions of mind such as feelings and preferences originate from multivariate analysis or empirically classified in clusters for affinity and meaning, rather than being prescribed. Therefore, the discrete observations formalised when individuals respond to determined stimuli could have no difference in degree but a qualitative characterisation. For this reason different hypotheses were tested during the development of the research as follows:

i. The primary research hypothesis was that the data obtained from persons' affective responses to design elements fit together and cooperate to define a quantitative structure established on the principles of RMT.

Results of the empirical approach indicated that the data fitted the model to some extent. Three hundred and six respondents gave their ratings on a five-point scale stimulated by 24 statements related to the attribute *specialness* of four pieces of wrapped confectionery. Twelve items from the preliminary pool presented some level of misfit and failed to meet the model's assumptions of response independence and unidimensionality. Sources of variance were associated with biased items, misinterpretation, ambiguity, redundancy and context. Those items were removed from the analysis. The remaining items constituted calibrated scales.

The empirical approach confirmed the hypothesis that affective data can establish a quantitative structure of measurement, producing invariant comparisons between any two persons independently of the statements within the scale and the comparison of any pair of statements independently of the persons. Nevertheless, respondents presented different response pattern for each confectionery, yielding different measurement structures. Therefore, the individual scales did not allow comparison across stimuli.

ii. Because of the difficulties found in the empirical study with confectioneries, a second hypothesis formulated that different stimuli could be measured on a sole continuum using a derivation of the RM.

To test the hypothesis the confectionery data set was used in the faceted Rasch model, which demonstrated to be an elegant theoretical solution.

The hypothesis was confirmed when the model parameterised the person estimates, the item estimates and the stimulus estimates independently. Nevertheless, new relationships between variables were observed during analysis. Within the individual frames of reference after calibration no indication of response dependence and trait dependence were observed; however, the enlarged frame of reference that was originated from the framework of the faceted approach embodying all stimuli on the same continuum produced some degree of response dependence. Although the analysis indicated anomalies, the problem could be a consequence of the faceted framework, rather than the data themselves.

This raised a further hypothesis that the stimuli would belong to the same frame of reference generated by the calibration of items using the multi-faceted Rasch model.

Evidence that the hypothesis is true was obtained through the investigation of the impact of the anomalies on the measure interpretation using an alternative technique developed during the research, which identified no high person-item residual correlations and indicated enough power to distinguish different groups of persons.

iii. The third research hypothesis was that the measurement of affective responses through calibrated structures using the RM does not vary within a same context even if different groups of persons are used.

To test the hypothesis an empirical approach was carried out with one hundred and ninety two persons. The study aimed to compare person, item and stimulus locations on the continuum using instruments to measure the persons' impression of a moisturizer cream when squeezing containers of everyday products. Firstly, six persons participated in a focus group in which semantic expressions were collected during tactile interactions with everyday product containers. Further sources for capturing semantics associated with touch followed the KE methodology. Sixteen statements related to a container of moisturizer cream were subsequently selected. A first sample with 120 participants rated their impression through a five-point scale. Eleven statements remained in the scale after calibration. The study was replicated through a second sample with 66 participants. Statements were not anchored on the continuum allowing re-calibrating the scale.

The replication of the study across two different samples confirmed the hypothesis. The results obtained from two-way ANOVA suggested that the difference between expected values of the two samples was not different from zero. Independent *t*-tests supported those results pointing to non-significant differences between score-person locations for the two samples, indicating empirical consistency. Furthermore, the results of a cross-validation through *t*-tests presented a non-significant statistical difference between separate calibrations using responses from different samples, endorsing the empirical consistency of the scales.

vi. The fourth research hypothesis formulated that if a calibrated set of items is kept as the core of the measurement structure, further items could be calibrated and accommodated into the structure.

The hypothesis was confirmed through an independent calibration of the 16-item set containing 11 items calibrated when testing the previous hypothesis and five additional items. Although the 11-item set was taken without anchoring, they remained fitted to the model after calibration, which resulted in a 14-item set. This demonstrated that the 11-item set was stable when incorporating more items into a new scale.

iv. The closing hypothesis formulated that if a stable measurement structure is obtained through calibration, then it is possible to model a design element of a product for particular affective responses.

An empirical approach tested the hypothesis through modelling the association of affective responses with the compliances of the existing products and with the compliance of a set of five container prototypes in a linear interval represented by the measurement scale. Using the 14-item set obtained from the test of the preceding hypothesis, 67 volunteers rated five non-functional prototypes with characteristics of compliance for stimulating determined affective responses for the impression of a moisturizer cream. The scale was once again calibrated through the faceted Rasch model. Furthermore, the force applied by participants was associated with the compliance of the existing containers. Affective responses and sensory information specified the design and the manufacture of the prototypes.

The hypothesis was confirmed by the comparison between the locations on the continuum for existing products and the locations on the continuum for the prototypes. The affective responses to the prototypes fell within the range of displacement at a force of 3N of the existing products, indicating that the latter can model new containers. Nevertheless, the range of force applied by the participants on the prototypes was higher than on the existing products. This might be a consequence of the influence of other factors, such as the containers' shape, when the persons processed the sensory information.

9.4 RATIONALE FOR THE DEVELOMENT OF MEASUREMENT INSTRUMENTS

Rasch measurement theory supported the rationale to the development of valid and reliable metrics for quantifying differences between individuals, between independent variables and between stimulus objects. The relationship amongst all those elements is established in terms of probability, embodying in this manner the inherent uncertainty when measuring latent variables.

The tests of the hypotheses formulated for determining the effectiveness of the RM confirmed that the rationale developed throughout the research can produce objective metrics in the AE. The rationale can be outlined as follows:

i. Context of measurement.

   The first objective in a study and perhaps, the most important one is to define what latent variable will be measured in the process. Usually physical systems are clearly defined to meet determined measurement conditions. However, it is not rare to find ill-formulated hypothesis of measurement. In most cases analysts want to establish just qualitative assessment although following quantitative methods. If measurement takes place, then the latent variable should carefully be defined in semantic terms. Furthermore, the measurement conditions should thoroughly be established in accordance with a well-acknowledged measurement theory.

ii. Words or statements that people will use to describe the product, product features or physical characteristics.

   The criteria for selection of words and statements should take into account the capability of the relevant latent variables to establish differences in degree rather than a qualitative classification. This could, however, not emerge from multivariate analysis, such as it is typically found in the domain. A more effective manner to select candidate variables would be to have rules and criteria to meet prescribed quantitative conditions.

iii. Physical stimuli and properties of the design element or product that will be investigated.

   An excessive number of stimulus objects can puzzle participants rather than provide more information. The decision on the number of stimuli should take into account whether respondents can clearly discern the different characteristics of the objects. Furthermore, if a large number of objects are presented to the respondents, disinterest and fatigue can take place and influence responses.

iv. Response options.

   The number of response options is conditioned on the clarity of the latent variable (i.e., based on the definition of the variable in terms of degrees of endorsement). Such as the number of physical stimuli, a high number of response options can obfuscate the actual person’s inclination in place of conveying more information.

v. Likelihood-ratio test.

   The likelihood-ratio test verifies which derivation of the RM should be applied if the scale offers more than two options of response. Rating scale or partial credit is then used within the multi-facet Rasch model.

vi. Score system.

Inconsistency associated with the respondents' use of the response categories shall be examined through the threshold patterns. If the response patterns are consistent, each response category has a point along the ability continuum where it is identified the most probable response. If the thresholds do not discriminate between adjacent categories, an analyst could consider combining their frequencies. However, this is not arbitrary. If data fit the model for some number of categories, then summing their frequencies can promote misfit.

vii. Differential item functioning.

A formal test for DIF should be carried out in every analysis. DIF is identified when a group demonstrates consistently greater inclination to endorse a statement than another group. A solution to deal with biased statements is to split the item in as many groups as the detected DIF. However, if too many items are split, the scale will be mis-characterised.

viii. Response dependence

Items that do not provide independent or relevant information can be the source of violations of the Rasch assumption of response independence. The test used in the research for detecting response dependence was the person-item residual correlation. The cut-off value for an indication of high correlation can vary according to the context of study.

This thesis has contributed to an alternative technique that discloses response dependence using a subtest protocol and an investigation of the item-person residual correlation when using the faceted framework.

ix. Test of fit.

Test of fit examines the degree to which the observed responses match in probabilistic terms with the Guttman pattern in a structure based on the expected values. The cut-off value of ±2.50 (i.e., representing 99% CI) was adopted in this thesis. However, data that do not fit the model should not be automatically rejected; rather they shall be investigated to identify their source of misfit and to what extent they corrupt measurement.

x. Unidimensionality.

After identifying and dealing with anomalies in the data set, there should be absence of any significant pattern in the residuals resultant of the relationship between items, excepting random associations. There are different methods to test the assumption.

This research adopted the method of comparison through *t*-tests of the two most divergent subsets of items identified by PCA of the residuals. The acceptable amount of deviating results was given by a binomial test.

It is noteworthy recalling that the assessment of dimensionality within the context of the RM has a different connotation than in the classical paradigm. The RM uses discrepancies between the observed responses and the expected responses by the model. That is, if data fit the model, items are said to work as a unidimensional measurement structure.

xi. Differential stimuli functioning

The taxonomy of the many-facet Rasch model was adapted to suit the characteristics of AE studies. Thus, the research has contributed to the development and incorporation of a test for identifying in-between stimuli heterogeneity. The test examines whether or not respondents are able of distinguishing different levels of the physical characteristic amongst all stimuli.

xii. Reliability

Two indices are typically used in Rasch analysis. The person separation index (PSI) from RMT and Cronbach's α from CTT. While reliability index represents essential information in the classical paradigm, in RMT it emphasises the precision of the individual estimates and therefore, the index is useful solely as an element of a comprehensive interpretation of a data set.

xiii. Calibrated metric for the relevant latent variable.

The outcome of the use of the rationale is a calibrated metric that associates affective responses with physical characteristics of products. A measure is obtained through the comparison of any two persons, pairs of items or any pair of stimuli, defined by the ratio of persons' endorsement level of the relevant affective attribute. One of the consequences of a Rasch-calibrated metric is the independence of the person, item and stimulus locations on the continuum, allowing invariant comparisons.

9.5 MAJOR IMPLICATIONS

9.5.1 Contributions to Knowledge

The original contribution of the research to knowledge is the demonstration that affective responses to design elements can be transformed into objective measures. The research pursued and has confirmed the hypothesis that observations from studies of affective engineering can meet measurement properties. Rasch measurement theory underpinned the research providing indirect tests for the numerical validity of the data in a quantitative structure. As a result, a theory-based rationale for measuring affective responses was developed throughout the research producing the following outcomes for the advancement of knowledge:

i. Transformation of affective responses into objective measures. As a consequence, permissible comparisons between results from different studies and the generalisation of research findings can be realised.

ii. Adaptation of the concepts, terms, definitions and equations typically found in academic publications and applications of the RM in other domains to the taxonomy used in AE. As a consequence, knowledge dissemination and the adoption of the model in the domain are facilitated.

iii. Development of an alternative technique to examine anomalies in data sets from affective responses using the faceted Rasch model. As a result, sources of local dependence in frames of reference that contain different stimulus objects as conditions for the affective responses can be identified.

iv. Demonstration that calibrated metrics are stable within a range of measurement error. The stability of a scale is shown to be a property of Rasch-based measurement structures.

v. Association of affective responses with sensory information through overlapping metrics. Independent parameterisations of persons, items and stimuli allow the correspondence of affective responses and sensory information using a shared physical element.

vi. Development of a rationale to establish a preliminary item bank. The incorporation of further variables (i.e., items and stimuli) is a consequence of the quantitative proprieties and independence of the estimation of parameters of a Rasch-calibrated metric.

9.5.2 Implications for Research and Practice

The research pursued the thesis that observations from affective user interaction with physical elements of products can objectively be evaluated. Rasch measurement theory has been the keystone to achieve the research aim. The theory has provided scientific procedures to test the hypothesis that the observations can be converted to objective measures of a latent variable.

Objective measures in AE will support analysts to understand the consumer's experience when interacting with products and eventually to their improvement. Applying the Rasch model allows the development of off-the-shelf scales of measurement because Rasch-calibrated metrics do not depend on sample distribution. If reasonably target samples of users are obtained and similar contexts are considered, then the scale can be replicated for measuring individual differences and drawn inferences embracing generalisations. Furthermore, different samples can reliably be compared, allowing an analyst to control the affective performance of a product during its life cycle.

The research demonstrated that the metrics calibrated through the RM allow the incorporation of further items to describe user experiences. This is evidence that it is possible to develop item banks in AE similarly to those in education and health science. Item bank will allow ad hoc applications in industry and research settings although maintaining a core metric for comparisons. An extrapolation of the benefit from item banks is the development of the computerized adaptive testing. The approach of item banks and CAT will reflect in lower costs and higher precision in consumer research.

Affective engineering is a multidisciplinary field of knowledge that has presented new challenges. This research has enlarged the role of an analyst to accommodate new skills. The expertise from the standpoint of the measurement paradigm should come from those who are challenged to transform individuals' latent expression to design elements into an improved product, understanding the statistical misfit and validating the metric for the relevant latent variable. Nevertheless, the rationale developed throughout the research does not supersede typical statistical approaches in AE. On the contrary, the rationale underpinned by Rasch measurement theory supports objective measures that in conjunction with statistical approaches can corroborate comparisons of results from different studies, strengthening the scientific investigation in the domain.

References

Ackley, D., Hinton, G. and Sejnowski, T., 1985. A learning algorithm for Boltzmann machines. *Cognitive Science*, 9 (1), pp.147 - 169.

Allan, D.W., 1987. Should the classical variance be used as a basic measure in standards metrology? *IEEE Transactions on instrumentation and measurement*, 36 (2), pp.646 – 654.

Alvarez, 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.

Alvarez, P. and Galera, C., 2001. Industrial marketing applications of quantum measurement techniques. *Industrial Marketing Management*, 30 (1), pp.13 – 22.

Anastasi, A., 1988. *Psychological testing*. New York, NY: MacMillan.

Andersen, E.B., 1973. A goodness of fit tests for the Rasch model. *Psychometrika*, 38 (1), pp.123 – 140.

Andersen, E.B., 1977. Sufficient statistics and latent trait models. *Psychometrika*, 42 (1), pp.69 – 81.

Andrich, D. and Hagquist, C., 2004. Detection of differential item functioning using analysis of variance. *2nd International conference on measurement in health, education, psychology and marketing: developments with the Rasch models*, 20 – 22 January, Perth, Australia.

Andrich, D. and Hagquist, C., 2012. Real and artificial differential item functioning. *Journal of Educational and Behavioral Statistics*, 37 (3), pp.387 – 416.

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., & Kreiner S., 2010. Quantifying response dependence between two dichotomous items using the Rasch model. *Applied Psychological Measurement*, 34 (3), pp.181 – 192.

Andrich, D., 1978. Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement*, 2(4), pp.581 – 594.

Andrich, D., 1985a. An elaboration of Guttman scaling with Rasch models for measurement. *In*: N. Brandon-Tuma, ed. *Sociological methodology*. San Francisco: Jossey-Bass, pp.33 – 80.

Andrich, D., 1985b. A latent trait model for items with response dependencies: Implications for test construction and analysis. *In*: S.E. Embretson, ed. *Test design,* New York: Academic Press, pp.245 – 275.

Andrich, D., 1988a. *Rasch models for measurement*. Sage university papers series on quantitative applications in the social sciences, No. 68, London: Sage.

Andrich, D., 1988b. A general form of Rasch's extended logistic model for partial credit scoring. *Applied Measurement in Education*, 1 (4), pp.363 – 378.

Andrich, D., 1989. Distinctions between assumptions and requirements in measurement in the social sciences. *In*: J.A. Keats, R. Taft, R.A. Heath and S. Lovibond, eds. *Mathematical and Theoretical Systems*, Amsterdam: Elsevier Science, pp.7–16.

Andrich, D., 2004. Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care*, 42 (1), pp.I-7 – I-16.

Andrich, D., 2006. Item discrimination and Rasch-Andrich thresholds revisited*. Rasch Measurement Transactions*, 20 (2), pp.1055 – 1057.

Andrich, D., 2010. Sufficiency and conditional estimation of person parameters in the polytomous Rasch model. *Psychometrika*, 75 (2), pp.292 – 308.

Andrich, D., Humphry, S.M. and Marais, I., 2012. Quantifying local, response dependence between two polytomous items using the Rasch model. *Applied Psychological Measurement*, 36 (4), pp.309 – 324.

Andrich, D., Sheridan, B. E., & Luo, G., 2012. RUMM2030: Rasch unidimensional models for measurement, Perth, Australia: RUMM Laboratory.

Aryadoust, S.V., 2009. The impact of Rasch item difficulty on confirmatory factor analysis. *Rasch Measurement Transactions*, 23 (2), p.1207.

Barnes, C. and Lillford, S., 2009. Decision support for the design of affective products. *Journal of Engineering Design*, 20 (5), pp.477 — 492.

Barnes, C., Childs, T., Henson, B. and Lillford, S., 2008. Kansei engineering toolkit for the packaging industry. *The TQM Journal*, 20 (4), pp.372 – 388.

Barnes, C., Childs, T.H.C., Henson, B. and Southee, C.H., 2004. Surface finish and touch: a case study in a new human factors tribology, *Wear*, 257 (7 – 8), pp.740–750.

Bechtel, G.G., 1985. Generalizing the Rasch model for consumer rating scales. *Marketing Science*, 4 (1), pp.62 – 73.

Belsey, D.A., Kuh, E. and Welsh, R., 1980. *Regression diagnostics: Identifying influential data and sources of collinearity*. New York: John Wiley & Sons.

Bernoulli, J. (1713). Ars conjectandi. Part 4. Basel. Excerpted in *Rasch Measurement Transactions*, 12 (1), p.625.

Birnbaum, A., 1968. Some latent trait models and their use in inferring an examinee's ability. *In*: Lord, F. M. and Novick, M. R., eds. *Statistical theories of mental test scores*, Reading: Addison-Wesley, pp.395 - 479.

Bland, J.M. and Altman, D.G., 1995. Multiple significance tests: the Bonferroni method. *British Medical Journal*, 310, p.170.

Bond, T.G. and Fox, C.M., 2007. *Applying the Rasch model: fundamental Measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum.

Borsboom, D., 2005. *Measuring the mind*. Cambridge: Cambridge University Press.

Borsboom, D., 2006. When does measurement invariance matter? *Medical Care*, 44 (Suppl), pp.176-181.

Briggs, D.C. and Wilson, M., 2003. An introduction to multidimensional measurement using Rasch models. *Journal of Applied Measurement*, 4 (1), pp.87 – 100.

Broderson, J., Meads, D., Kreiner, S., Thorsen, H., Doward, L. and McKenna, S., 2007. Methodological aspects of differential item functioning in the Rasch model. *Journal of Medical Economics*, 10 (3), pp.309-324.

Brodgen, H.E., 1977. The Rasch model, the law of comparative judgement and additive conjoint measurement. *Psychometrika*, 42 (4), pp.631 – 634.

Camargo, F.R. and Henson, B., 2010. Measuring the specialness of confectionery: a Rasch model approach in affective engineering. *In*: *International conference on probabilistic models for measurement in education, psychology, social science and health*, 13 – 16 June, Copenhagen.

Camargo, F.R. and Henson, B., 2011. Measuring affective responses for human-oriented product design using the Rasch model. *Journal of Design Research*, 9 (4), pp.360 – 375.

Camargo, F.R. and Henson, B., 2012a. The Rasch probabilistic model for measuring affective responses to product features. *International Journal of Human Factors and Ergonomics*, 1 (2), pp.204 – 219.

Camargo, F.R. and Henson, B., 2012b. Improving Kansei measurement using the Rasch model. *In*: *International Conference on Kansei Engineering and Emotion Research*, 22 – 25 May, Penghu, Taiwan.

Camargo, F.R. and Henson, B., 2012c. A rationale for comparing affective responses to stimulus objects using the faceted Rasch model. *In*: *International Conference on Probabilistic Models for Measurement in Education, Psychology, Social Science and Health*, 23 –25 January, Perth, Australia.

Camargo, F.R. and Henson, B., 2012d. Invariant comparisons in affective design. *In*: Ji, Y.G., ed. *Advances in affective and pleasurable design*, Boca Raton: CRC Press, pp.490 – 499.

Camargo, F.R., 2013. Response dependence in a multi-conditional frame of reference. *In: 2013 Rasch work group meeting*, 16 – 19 January, Cape Town.

Campbell, N. R., 1920. *Physics. The elements*. Cambridge: The University Press

Campbell, N. R., 1928. *An account of the principles of measurement and calculation*. London: Longmans, Green.

Chang, H. and Wu, S., 2010. Applying the Rasch measurement to explore elderly passengers' abilities and difficulties when using buses in Taipei. *Journal of Advanced Transportation*, 44 (3), pp.134 – 149.

Chang, H. and Yang, C., 2008. Explore airlines' brand niches through measuring passengers' repurchase motivation: an application of Rasch measurement. *Journal of Air Transport Management*, 14 (3), pp.105–112.

Chen, J., Wang, K., Liang, J., 2008. A Hybrid Kansei Design Expert System Using Artificial Intelligence. *In*: Ho. T., Zhou, Z., eds. *PRICAI 2008, LNAI 5351*, pp. 971-976.

Chen, W.H. and Thissen, D., 1997. Local dependence indices for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, 22, pp.265 – 289.

Chen, X., Shao, F., Barnes, C., Childs, T., & Henson, B., 2009. Exploring relationships between touch perception and surface physical properties. *International Journal of Design*, 3 (2), pp.67 – 77.

Cheng, Y., (2010). Exploring passenger anxiety associated with train travel. *Transportation*, 37 (6), pp.875 – 896.

Childs, T., 2010. *Kansei is a noun, affective is an adjective*. Affective engineering research group meeting, University of Leeds. Unpublished.

Choppin, B.H., 1968. An item bank using sample-free calibration. *Nature*, 219, pp.870 – 872.

Choppin, B.H., 1978. *Item banking and the monitoring of achievement*. Slough: National Foundation for Educational Research.

Christensen, K.B., 2006. From Rasch Scores to Regression. *Journal of Applied Measurement*, 7 (2): 184 – 191.

Christensen, K.B., Bjorner, J.B., Kreiner, S. and Petersen, J.H., 2002.Testing unidimensionality in polytomous Rasch models. *Psychometrika*, 67 (4), pp.563-574.

Cohen, J., 1988. *Statistical power analysis for the behavioral sciences*. 2nd ed., New Jersey: Lawrence Erlbaum.

Cohen, L., 1979. Approximate expressions for parameter estimates in the Rasch model. *British Journal of Mathematical and Statistical Psychology*, 32 (1), pp.113-120.

Cook, D. and Weisberg, S., 1982. *Residuals and influence in regression*. New York: Chapman & Hall.

Cook, D., 1979. Influential observations in linear regression. *Journal of the American Statistical Association,* 74 (36). pp.169 – 174.

Cortina, J. M., 2003. Apples and oranges (and pears, oh my!): the search for moderators in meta-analysis. *Organizational Research Methods*, 6 (4), pp.415 – 439.

Cortina, J.M., 1993. What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, 78 (1), pp.98 – 104.

Coxhead, P. and Bynner, J.M., 1981. Factor analysis of semantic differential data. *Quality and Quantity*, 15, pp.553 – 567.

Cronbach, L.J., 1951. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16 (3), pp.297 – 334.

Cross, N., 2000. *Engineering design methods: strategies for product design*. Chichester: John Wiley & Sons.

Cureton, E. E., 1965. Reliability and validity: basic assumptions and experimental design. *Educational and Psychological Measurement*, 25, pp.326 – 346.

d'Astous, A. and Kamau, E., 2010. Consumer product evaluation based on tactile sensory information. *Journal of Consumer Behaviour,* 9 (3), pp.206 – 213.

Davies, R.B. and Hutton, B., 1975. The effect of errors in the independent variables in linear regression. *Biometrika*, 62 (2), pp.383 – 391.

Deese, J., 1964. The associative structure of some common English adjectives. *Journal of Verbal Learning and Verbal Behavior*, 3, pp.347 – 357.

Desmet, P., 2004. Measuring emotion: development and application of an instrument to measure emotional responses to products. *In*: M.A. Blythe, A.F. Monk, K. Overbeeke and P.C. Wright, eds. *Funology: From Usability to Enjoyment*. Chapter 9. Dordrecht, The Netherlands: Kluwer.

DeVellis, R.F., 2006. Classical test theory. *Medical Care*, 44 (11), pp.S50 – S59.

Eckes, T., 2005. Examining rater effects in TestDaF writing and speaking performance assessments: a many-facet Rasch analysis. *Language Assessment Quarterly*, 2 (3), pp.197–221.

Eckes, T., 2009. Many-facet Rasch measurement. *In*: Takala, S., ed. *Reference supplement to the manual for relating language examinations to the common European framework of reference of languages: learning, teaching, assessment (Section H)*, Strasbourg: Council of Europe/Language Policy Division.

Eckes, T., 2011. Item banking for C-tests: A polytomous Rasch modeling approach, *Psychological Test and Assessment Modeling*, 53 (4), pp.414 – 439.

Elhan A.E., Öztuna, D., Kutlay, S., Küçükdeveci, A.A. and Tennant, A., 2008. An initial application of computerized adaptive testing (CAT) for measuring disability in patients with low back pain. *BMC Musculoskeletal Disorders*, 9 (166), pp.1-15.

Elhan A.H., Oztuna D., Kutlay S., Kucukdeveci, A.A. and Tennant, A., 2008. An initial application of computerized adaptive testing (CAT) for measuring disability in patients with low back pain. *BMC Musculoskeletal Disorders*, 9, p.166.

Elokla, N. and Hirai, Y., 2012. Developing new emotional evaluation methods for measuring users' subjective experiences in the virtual environments. *In*: Y.G. Ji, ed. *Advances in affective and pleasurable design,* Boca Raton: CRC Press, pp.425 – 435.

Embretson, S.E. and Reise, S.P., 2000. *Item response theory for psychologists*. Mahwah, New Jersey: Lawrence Erlbaum.

Engelhard, G., 1994. Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, 31 (2), pp.93 – 112.

Engelhard, G., 2011. Evaluating the bookmark judgments of standard-setting panelists. *Educational and Psychological Measurement*, 71 (6), pp.909 – 924.

Ewing, M.T., Salzberger, T. and Sinkovics, R.R., 2005. An alternate approach to assessing cross-cultural measurement equivalence in advertising research. *Journal of Advertising*, 34 (1), pp.17 – 36.

Ewing, M.T., Salzberger, T. and Sinkovics, R.R., 2009. Confirmatory factor analysis vs. Rasch approaches: Differences and measurement implications. *Rasch Measurement Transactions*, 23 (1), pp.1194 – 1195.

Fazio, R.H. and Olson, M.A., (2003). Attitudes: foundations, functions, and consequences. *In*: M.A. Hogg and J. Cooper, eds. *The Sage handbook of social psychology,* pp.131 – 160. London: Sage.

FDA - U.S. Department of Health and Human Services Food and Drug Administration, 2009. Guidance for Industry patient-reported outcome measures: use in medical product development to support labeling claims, p.21. http://www.fda.gov/downloads/Drugs/Guidances/UCM193282.pdf. Accessed on 12/05/2013.

Finkelstein, L. and Leaning. M.S., 1984. A review of the fundamental concepts of measurement. *Measurement*, 2 (1), pp.25 – 34.

Fischer, G.H. and Molenaar, I.W., 1995. *Rasch models: foundations, recent developments, and applications*. New York: Springer.

Fischer, G.H., 1995. Derivations of the Rasch model. & the derivation of polytomous Rasch models. *In*: G.H. Fischer and I.W. Molenaar, eds. *Rasch models: foundations, recent developments, and applications*. New York: Spring-Verlag.

Fisher, G.H., 1973. Linear logistic test model as an instrument in education research. *Acta Psychologica*, 37 (5), pp.359 – 374.

Fisher, R. A., 1922. On mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London* (A) 222, pp.309 – 368.

Fisher, W.P., 1992. Reliability statistics. *Rasch Measurement Transactions*, 6 (3), p.238.

Flora, D., LaBrishand, C. and Chalmers, P., 2012. Old and new ideas for data screening and assumption testing for exploratory and confirmatory factor analysis. *Frontiers in Quantitative Psychology and Measurement*, 3 (55), pp.1 – 21.

Ganglmair, A. and Lawson, R., 2003. Measuring affective responses to consumption using Rasch modelling. *Journal of Consumer Satisfaction, Dissatisfaction and Complaining Behavior*, 16, pp.198 – 210.

Garcia, 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.

Goldstein, H., 1979. Consequences of using the Rasch model for educational assessment. *British Education Research Journal*, 5 (2), pp.211 – 220.

Goldstein, H., 1980. Dimensionality, bias, independence and measurement scale problems in latent trait test score models. *British Journal of Mathematical and Statistical Psychology*, 33 (2), pp.234 – 246.

Goldstein, H., 2012. Francis Galton, measurement, psychometrics and social progress. *Assessment in Education: Principles, Policy & Practice*, 19 (2), pp.147 – 158.

Grayson, D., 2004. Some myths and legends in quantitative psychology. *Understanding Statistics*, 3 (1), pp.101 – 134.

Green, S.B., Lissitz, R.W. and Mulaik, S.A. 1977. Limitations of coefficient alpha as an index of test unidimensionality. *Educational and Psychological Measurement*, 37 (4), pp.827 – 838.

Gulliksen, H., (1950). *Theory of Mental Tests*. New York: Wiley

Guttman, L.A., 1950. The basis for scalogram analysis. *In*: S.A. Stouffer, L.A. Guttman, F.A. Suchman, P.F. Lazarsfeld, S.A. Star and J.A. Clausen, eds. *Studies in social psychology in world war II*, Vol. 4, *Measurement and Prediction*, Princeton: Princeton University Press, pp.60 – 90.

Hahn, E.A., Cella, D., Bode, R.K., Gershon, R. and Lay J., 2006. Item banks and their potential applications to health status assessment in diverse populations, *Medical Care*, 44 (11), pp.S189 – S197.

Hayashi, C., 1952. On the prediction of phenomena from qualitative data and the quantification of qualitative data from the mathematico-statistical point of view. *Annals of the Institute of Statistical Mathematics*, 3, pp.69 – 98.

Hays, R.D., Morales, L.S. and Reise, S.P., 2000. Item response theory and health outcomes measurement in the 21st century. *Medical Care*, 38 (9), pp.28 – 42.

Heise, D.R., 1969. Some methodological issues in semantic differential research. *Psychological Bulletin*, 72 (6), pp.406 – 422.

Helmholtz, H.V., 1887. *Epistemological Writings*. 1977 edition. R.S. Cohen and Y. Elkana, eds. Dordrecht: D. Reidel

Henson, B., 2009. The Rasch model as a measurement model in affective engineering. *In*: *MINET Conference: measurement, sensation and cognition*, 10-12 November, National Physical Laboratory, Teddington.

Henson, B., Barnes, C., Livesey, R., Childs, T. and Ewart K., 2006. Affective consumer requirements: a case study of moisturizer packaging. *Concurrent Engineering: Research and Applications*, 14 (3), pp.187 – 196.

Hirata, R., Nagamachi, M. and Ishihara, S., 2004. Satisfying emotional needs of the beer consumer through kansei engineering. *In*: *Proceedings of the 7th International conference on quality management and organizational development*, 4 – 6 August, Monterrey, México, pp.219–227.

Hooks, I. and Farry, K.A., 2001. *Customer-centered products: creating successful products through smart requirements management*. New York: AMACOM.

Horton, M. and Tennant, A., 2010 Assessing unidimensionality using Smith's (2002) approach in RUMM 2030. *In*: *International conference on probabilistic models for measurement in education, psychology, social science and health*, 13–16 June, Copenhagen.

Hotta, H. and Hagiwara, M., 2005. An automatic rule creating method for kansei data and its application to a font creating system. *In*: *2nd Conference on modelling decisions for artificial intelligence*, 25 – 27 July, Tsukuba, Japan.

Hu, X.C., Lin, T.Y. and Han, J., 2004. A new rough set model on database systems. *Fundamenta Informaticae*, 59 (2-3), pp.135 – 152.

Humphry, S.M. and Andrich, D., 2008. Understanding the unit in the Rasch model. *Journal of Applied Measurement*, 9 (3), pp.249 – 264.

Humphry, S.M., 2011. The role of the unit in physics and psychometrics. *Measurement: Interdisciplinary Research and Perspectives,* 9 (1), pp.1 – 24.

International Standards Office, 1994. *ISO 5725 – Part 1 – 6*: *accuracy (trueness and precision) of measurement methods and results*. Geneva: ISO.

Ironson, G.H. and Subkoviak, M.J., 1979. A comparison of several methods of assessing item bias. *Journal of Educational Measurement*, 16 (4), pp.209 – 225.

Ishihara, S., Ishihara, K., Nagamachi, M. and Matsubara, Y., 1995. An automatic builder for an engineering expert system using self-organizing neural networks, *International Journal of Industrial Ergonomics*, 15 (1), pp.13–24.

Ishihara, S., Ishihara, K., Nagamachi, M. and Matsubara, Y., 1997. An analysis of kansei structure on shoes using self-organizing neural networks. *International Journal of Industrial Ergonomic*, 19(2), pp.93 – 104.

Kaiser, H.F., 1970. A second generation little jiffy. *Psychometrika*, 35 (4), pp.401 – 415.

Karabatsos, G., 2001. The Rasch model, additive conjoint measurement, and new models of probabilistic measurement theory. *Journal of Applied Measurement*, 2 (4), pp.389 – 423.

Karwowski, W., 2005. Ergonomics and human factors: the paradigms for science, engineering, design, technology and management of human-compatible systems. *Ergonomics*, 48 (5), pp.436 – 463.

Kastrin, A. and Peterlin, B., 2010. Rasch-based high-dimensionality data reduction and class prediction with applications to microarray gene expression data. *Expert Systems Applications*, 37 (7), pp.5178 – 5185.

Kline, P., 1986. *A handbook of test construction: introduction to psychometric design*. London: Methuen.

Kline, P., 1998. *The new psychometrics: science, psychology and measurement.* London: Routledge.

Klöcker, A., Arnould, C., Penta, M. and Thonnard, J., 2012. Rasch-built measure of pleasant touch through active fingertip explorations. *Frontier in Neurorobotics*, 6 (5), pp.1 – 9.

Komazawa, T. and Hayashi, C., 1976. A statistical method for quantification of categorical data and its applications to medical science. *In*: F.T. Dombal and F. Gremy, eds. *Decision making and medical care*. Amsterdam: North-Holland.

Krantz, D.H., Luce, R.D., Suppes, P. and Tversky, A., 1971. *Foundations of measurement, Vol. 1*. New York: Academic Press.

Kreiner, S. and Christensen, K.B., 2004. Analysis of local dependence and multidimensionality in graphical loglinear Rasch models. *Communications in Statistics: Theory and Methods*, 33 (6), pp.1239 – 1276.

Kreiner, S. and Christensen, K.B., 2004. Graphical Rasch models. *In*: M. Mesbah, F.C. Cole and M.T. Lee, eds. *Statistical methods for quality of life studies*, Dordrecht: Kluwer, pp. 187 – 203.

Kreiner, S., 2007. Validity and objectivity: reflections on the role and nature of Rasch models. *Nordic Psychology*, 59 (3), pp.268 – 298.

Kyngdon, A., 2004. Comparing factor analysis and the Rasch model for ordered response categories: an investigation of the scale of gambling choices. *Journal of Applied Measurement*, 5 (4), pp.398 – 418.

Laurans, G., Desmet, P. and Hekkert, P., 2009. Assessing emotion in interaction: Some problems and a new approach. *In*: International conference on design pleasurable products and interfaces, 13 – 16 October, Compiegne, France.

Likert, R., 1932. A technique for the measurement of attitudes. *Archives of Psychology*, 140, pp.1-55.

Linacre J.M., 1994c. Constructing measurement with a many-facet Rasch model. *In*: M. Wilson, ed. *Objective measurement: theory in practice*. *Vol. II*. Newark, NJ: Ablex.

Linacre J.M., 2002c. Facets, factors, elements and levels. *Rasch Measurement Transactions*, 16 (2), p.880.

Linacre, J. M., 2012. *Facets computer program for many-facet Rasch measurement, version 3.70.0*. Beaverton, Oregon: Winsteps.com.

Linacre, J.M. and Fisher, P., 2012. Harvey Goldstein's objections to Rasch measurement: A response from Linacre and Fisher. *Rasch Measurement Transactions*, 26 (3), pp.1383 – 1389.

Linacre, J.M., 1989. *Many-facet Rasch measurement*. Chicago: MESA Press.

Linacre, J.M., 1994a. PROX with missing data, or known item or person measures. *Rasch Measurement Transactions*, 8 (3), p.378.

Linacre, J.M., 1994b. Sample size and item calibration stability. *Rasch Measurement Transactions*, 7 (4), p.328.

Linacre, J.M., 1995. PROX for polytomous data. *Rasch Measurement Transactions*, 8 (4), p.400.

Linacre, J.M., 1998. Structure in Rasch residuals: Why principal components analysis (PCA)? *Rasch Measurement Transactions*, 12 (2), p.636.

Linacre, J.M., 1999. Understanding Rasch measurement: estimation methods for Rasch measures. *Journal of Outcome Measurement*, 3 (4), pp.382 – 485.

Linacre, J.M., 2002a. Optimising rating scale category effectiveness. *Journal of Applied Measurement*, 3 (1), pp.85 – 106.

Linacre, J.M., 2002b. What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, 16 (2), p.878.

Linacre, J.M., 2005. Standard errors: means, measures, origins and anchor values. *Rasch Measurement Transactions*, 19 (3), p.1030.

Linacre, J.M., 2010. Predicting responses from Rasch measures. *Journal of Applied Measurement*, 11 (1), pp.1 – 10.

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L., 1981. Item bias in a test of reading comprehension. *Applied Psychological Measurement*, 5 (2), pp.159 – 173.

Liu, X., 2010. *Using and developing measurement instruments in science education: a Rasch modelling approach*. Charlotte, NC: IAP.

Lord, F.M. and Novick, M.R., 1968. *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Lord, F.M., 1952. *A theory of tests scores. Psychometric Monograph No. 7*. Iowa City: Psychometric Society.

Lord, F.M., 1980. *Applications of item response theory to practical testing problems*. New Jersey: Lawrence Erlbaum.

Lucadamo, A., 2010. Rasch analysis and multilevel models for the evaluation of the customer satisfaction. *Electronic Journal of Applied Statistical Analysis*, 3 (1), pp.44 –51.

Luce, R.D. and Tukey, J.W., 1964. Simultaneous conjoint measurement: a new type of fundamental measurement. *Journal of Mathematical Psychology*, 1 (1), pp.1 – 27.

Lunz, M.E., Wright, B.D. and Linacre, J.M., (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education*, 3 (4), pp.331 – 345.

Luo, G. and Andrich, D., 2005. Estimating parameters in the Rasch model in the presence of null categories. *Journal of Applied Measurement*, 6 (2), pp.128 – 146.

Mair, P. and Hatzinger, R., 2007. CML based estimation of extended Rasch analysis with the eRm package in R. *Psychology Science*, 49 (1), pp.26 – 43.

Marais, I. And Andrich, D., 2008a. Formalizing dimension and response violations of local independence in the unidimensional Rasch model. *Journal of Applied Measurement*, 9 (3), pp.200 – 215.

Marais, I. And Andrich, D., 2008b. Effects of varying magnitude and patterns of response dependence in the unidimensional Rasch model. *Journal of Applied Measurement*, 9 (2), pp.105 – 124.

Marais, I. And Andrich, D., 2011. Diagnosing a common rater halo effect using the polytomous Rasch model. *Journal of Applied Measurement*, 12 (3), pp.194 – 211.

Marais, I. and Andrich, D., 2012. *RUMMss – Rasch unidimensional measurement model simulation studies program*. The University of Western Australia, Perth.

Masters, G.N., 1982. A Rasch model for partial credit scoring. *Psychometrika*, 47 (2), pp.149 – 174.

Matsubara, Y. and Nagamachi, M., 1997. Hybrid kansei engineering system and design support. *International Journal of Industrial Ergonomics*, 19 (2), pp.81 – 92.

McKennel, A.C. and Bynner, J.M., 1969. Self-images and smoking behavior among school boys. *British Journal of Educational Psychology*, 39, pp.27-39.

Messick, S., 1989. Validity. *In*: R.L. Linn, ed. *Educational measurement*, 3rd ed., New York: Macmillan.

Michell, J., 1988. Some problems in testing the double cancellation condition in conjoint measurement. *Journal of Mathematical Psychology* 32 (4), pp.466 – 473.

Michell, J., 2009. The psychometricians' fallacy: too clever by half? *British Journal of Mathematical and Statistical Psychology*, 62 (1), pp.41 – 55.

Mindak, W.A., 1961. Fitting the semantic differential to the marketing problem. *Journal of Marketing*, 25 (4), pp.28-33.

Mordkoff, A.M., 1963. An empirical test of the functional antonymy of semantic differential scales. *Journal of Verbal Learning and Verbal Behavior*, 2, pp.504-508.

Moulton, M.H., 2003. *Rasch estimation demonstration spreadsheet*. www.rasch.org/moulton.htm. Accessed on 07/11/2012.

Myford, C.M., and Wolfe, E.W., 2003. Understanding Rasch measurement: detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, 4 (4), pp.386-422.

Myford, C.M., and Wolfe, E.W., 2004. Understanding Rasch measurement: detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, 5 (2), pp.189-227.

Nagamachi, M., 1974. A study of emotional technology. *Japanese Journal of Ergonomics*, 10 (2), pp.121 – 130.

Nagamachi, M., 1989. *Kansei Engineering*. Tokyo: Kaibundou.

Nagamachi, M., 1995. Kansei engineering: a new ergonomic consumer-oriented technology for product development*.* *International Journal of Industrial Ergonomics*, 15 (1), pp.3 – 11.

Nagamachi, M., 2008. Perspectives and the new trend of Kansei/affective engineering. *The TQM Journal*, 20 (4), pp.290 – 298.

Nagamachi, M., Okasaki, Y. and Ishikawa, M., 2006. Kansei engineering and application of the rough sets model. *Journal of Systems Control Engineering*, 220 (8), pp.763 – 768.

Nagamachi, M., Tachikawa, M., Imanishi, N., Ishizawa, T. and Yano, S., 2008. A successful statistical procedure on kansei engineering products. *In*: *11th Conference on quality management and organizational development*, 20–22 August, Helsingborg, Sweden.

Nishino, T., Nagamachi, M. and Ishihara, S., 2001. Rough set analysis on kansei evaluation of Color. *In*: Proceedings of the *International conference on affective human factors design*, 26 – 29 June, Singapore. London: Asean Academic Press, pp.109-115.

Nishino, T., Nagamachi, M. and Tanaka, H., 2005. Variable precision bayesian rough set model and its application to human evaluation data. *In*: Proceedings of the *10th International conference on rough sets, fuzzy sets, data mining, and granular computing*, 31 August – 3 September, Regina, Canada, pp.294 – 303.

Nunnally, J.O., 1978. *Psychometric theory*. New York: McGraw-Hill.

Ogino, A., 2012. A Model of User Preference for Personalization Service. *In: International Conference on Kansei Engineering and Emotion Research*, 22 – 25 May, Penghu, Taiwan, pp.163 – 168.

Okamoto, R.H., Nishino, T. and Nagamachi, M., 2007. Comparison between statistical and lower / upper approximations rough sets models for beer can design and prototype evaluation. *In*: *10th Quality management and organizational development conference*, 18-20 June, Helsingborg, Sweden.

Osgood, C., Suci, G. and Tannenbaum, P., 1957. *The measurement of meaning*. Urbana: University of Illinois Press.

Osgood, C.E. and Suci, G.J., 1955. Factor analysis of meaning. *Journal of Experimental Psychology*, 50, pp.325-338.

Osgood, C.E., 1952. The nature and measurement of meaning. *Psychology Bulletin*, 49, pp.192-237.

Osgood, C.E., 1964. Semantic differential technique in the comparative study of cultures. *American Anthropologist*, 66, pp.171-200.

Osgood, C.E., 1970. Speculation on the structure of interpersonal intentions. *Behavioral Science*, 15 (3), pp.237-254.

Oskamp, S., 1977. *Attitudes and opinions*. Englewood Cliffs, NJ: Prentice-Hall.

Osterlind, J.S. and Everson, H.T., 2009. *Differential item functioning*. Sage university papers series on quantitative applications in the social sciences, 2nd ed. London: Sage.

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.

Pawlak, Z. and Skowron, A., 2007. Rudiments of rough sets. *Information Series*, 177 (1), pp.3 – 27.

Pawlak, Z., 1991. *Rough sets: theoretical aspects of reasoning about data*. London: Kluwer.

Pedler, P., 1987. *Accounting for psychometric dependence with a class of latent trait models*. PhD thesis, Department of Education. The University of Western Australia.

Perline, R., Wright, B.D, Wainer, H., 1979. The Rasch model as additive conjoint measurement. *Applied Psychological Measurement,* 3 (2), pp.237 – 255.

Perneger, T.V. (1998). What's wrong with Bonferroni adjustments? *British Medical Journal*, 316: pp.1236 – 1238.

Picard, R.W., 1997. *Affective computing*. Cambridge: MIT Press

Raju, N.S., 1988. The area between two item characteristic curves. *Psychometrika*, 53 (4), pp.495 – 502.

Rasch, G., 1960, 1980. *Probabilistic models for some intelligence and attainment tests*, (Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by Wright, B.D. Chicago: The University of Chicago Press.

Rasch, G., 1961. On general laws and the meaning of measurement in psychology. *In*: Proceedings of the *4th Berkeley symposium on mathematical statistics and probability*. 20 June – 30 July 1960. Berkeley: University of California Press, pp.321 – 334

Rasch, G., 1968. *A mathematical theory of objectivity and its consequences for model contribution*. Amsterdam: European meeting on statistics, econometrics and management science.

Rasch, G., 1977. On specific objectivity: an attempt at formalizing the request for generality and validity of Scientific Statements. *In:* Proceedings of the *symposium in scientific objectivity* at "Rolighed", Vedbæk, 14 – 16 May 1976, Copenhagen: the Danish Yearbook of Philosophy, 14: pp.58-93.

Reese, T.W., 1943. The application of the theory of physical measurement to the measurement of psychological magnitudes, with three experimental examples. *Psychological Monographs*, 55 (3), pp.1 – 89.

Rizopoulos, D., 2006. ltm: an R package for latent variable modelling and item response theory analysis. *Journal of Statistical Software*, 17 (5), pp.1–25.

Rosenbaum, P. R., 1988. Item bundles. *Psychometrika*, 53 (3), pp.349-359.

Rossi, G.B., 2007. Measurability. *Measurement*, 40, pp.545-562.

Rudner, L.M., Getson, P.R. and Knight, D.L., 1980. Monte Carlo comparison of seven biased item detection techniques. *Journal of Education Measurement*, 17 (1), pp.1 – 10.

Salzberger, T., 2009. *Measurement in marketing research: an alternative framework*. Cheltenham, UK & Northampton, USA: Edward Elgar.

Salzberger, T., 2013. Attempting measurement of psychological attributes. *Frontiers in Quantitative Psychology and Measurement*, 4 (75), pp.1 – 4.

Scheiblechner, H., 1999. Additive conjoint isotonic probabilistic models (ADISOP). *Psychometrika*, 64 (3), pp.295–316.

Schifferstein, H.N.J. and Zwartkruis-Pelgrim, E.P.H., 2008. Consumer-product attachment: Measurement and design implications*. International Journal of Design*, 2 (3), pp.1 – 13.

Schütte, S. and Eklund, J., 2001. An approach to Kansei engineering: methods and a case study on design identity. *In*: Proceedings of the *International Conference on Affective Human Factors Design*, 26 – 29 June, Singapore. London: Asian Academic Press.

Schütte, S. and Eklund, J., 2010. Rating scales in kansei engineering: modifications for a European context. *In*: *International Conference on Kansei Engineering and Emotion Research*, 2 – 4 March, Paris.

Schütte, S., 2005. *Engineering emotional values in product design- Kansei Engineering in development*. PhD thesis. Institution of Technology. Linköping, University.

Schumacker, R.E. and Linacre, J.M., 1996. Factor analysis and Rasch analysis. *Rasch Measurement Transactions,* 9 (4), p.470.

Shao, F., Chen, X., Barnes, C.J. and Henson, B., 2010. A novel tactile sensation measurement system for qualifying touch perception. *Journal of Engineering in Medicine*, 224 (1), pp.97 – 105.

Shepard, L., Camilli, G. and Averill, M., 1981. Comparison of procedures for detecting test-item bias with both internal and external ability criteria. *Journal of Educational Statistics*, 6 (4), pp.317 – 375.

Shimizu, Y. and Jindo, T., 1995. A fuzzy logic analysis method for evaluating human sensitivities. *International Journal of Industrial Ergonomics*, 15 (1), pp.39 – 47.

Sijtsma, K., 2010. Psychological measurement between physics and statistics (Keynote). *In*: *4th International conference on probabilistic models for measurement in education, psychology, social science and health*, 13 –16 June, Copenhagen.

Smith R.M. and Miao, C.Y., 1994. Assessing unidimensionality for Rasch measurement. *In*: M. Wilson, ed. *Objective measurement: theory into practice, Vol. 2*. Norwood NJ: Ablex.

Smith, E.V., 2002. Understanding the Rasch model: detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement*, 3 (2), pp.205 – 231.

Smith, E.V., 2005. Effect of item redundancy on Rasch item and person estimates. *Journal of Applied Measurement*, 6(2), pp.147 – 163

Smith, R.M. and Plackner, C., 2009. The Family approach to assessing fit in Rasch measurement. *Journal of Applied Measurement*, 10 (4), pp.424 – 437.

Smith, R.M., 1996. A Comparison of methods for determining dimensionality in Rasch measurement. *Structural Equation Modeling*, 3 (1), pp.25 – 40.

Smith, R.M., 1996. Polytomous mean-square fit statistics. *Rasch Measurement Transactions*, 10 (3), pp.516 – 517.

Smith, R.M., Linacre, J.M. and Smith, Jr., E.V., 2003. Guidelines for manuscripts. *Journal of Applied Measurement*, 4 (Editorial), pp.198-204.

Smith, R.M., Schumacker, R.E. and Bush, M.J., 1998. Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement*, 2 (1), pp.66 – 78.

Soutar, G.N., Bell, R.C. and Wallis, Y.M., 1990. Consumer acquisition patterns for durable goods: a Rasch analysis. *European Journal of Marketing*, 24 (8), pp.31 – 39.

Spearman, C., 1904. The proof and measurement of association between two things. *The American Journal of Psychology*, 15 (1), pp.72 – 101.

Stevens, S.S., 1946. On the theory of scales of measurement. *Science*, 103, pp.677-680.

Su, J., Jiang, Y. and Wanga, P., 2008. Research on product image styling design method based on neural network and genetic algorithm. *In*: *4th International conference on natural computation*, 25 – 27 August, Jinan, China.

Swain, C.D., Bryndza, H.E. and Swain, M.S., 1979. Hazards in Factor Analysis. *Journal of Chemical Information and Computer Sciences*, 19 (1), pp.19 – 23.

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 & Rheumatism*, 57 (8), pp.1358 – 1362.

Tennant, A. and Pallant, J.F., 2006. Unidimensionality Matters! (A Tale of Two Smiths?). *Rasch Measurement Transactions*, 20 (1), pp.1048 – 1051.

Tennant, A., McKenna, S.P. and Hagell, P., 2004. Application of Rasch analysis in the development and application of quality of life instruments, *Value in Health*, 7 (1), pp.S22–S26.

Thurstone, L.L., 1927. The law of comparative judgment. *Psychological Review*, 34 (4), pp.278 – 286.

Thurstone, L.L., 1928. Attitudes can be measured. *American Journal of Sociology*, 33, pp.529 - 554.

Thurstone, L.L., 1947. *Multiple-factor analysis*. Chicago: University of Chicago Press.

Traub, R.E. and Rowley, G.L., 1991. NCME instructional module: understanding reliability. *Educational Measurement: Issues and Practice*, 10 (1), pp.37 – 45.

Traub, R.E., 1994. *Reliability for the social sciences: theory and applications*. Sage university papers series on measurement methods for the social sciences, Vol. 3. Thousand Oaks: Sage.

Van der Linden, W., 1994. Review of Michell-1990. *Psychometrika*, 59 (1), pp.139 –142.

VIM - International vocabulary of metrology, 2012. *JCGM 200:2012* - Basic and general concepts and associated terms, 3rd ed. www.bipm.org/utils/common/documents/jcgm/JCGM_200_2012.pdf. Accessed on 22/01/2013.

Walter, S., 2008. Conceptualizing overfit or over-parameterization. *Rasch Measurement Transactions*, 22 (2), pp.1165.

Wang, X., Bradlow, E. T., and Wainer, H., 2002. A general Bayesian model for testlets: Theory and applications. *Applied Psychological Measurement*, 26 (1), pp.109 – 128.

Warm, T.A., 1989. Weighted likelihood estimation of ability in item response theory. *Psychometrika*, 54 (3), pp.427 – 450.

Waugh, R.F. and Chapman, E.S., 2005. An analysis of dimensionality using factor analysis (true-score theory) and Rasch measurement: what is the difference? Which method is better? *Journal of Applied Measurement*, 6 (1), pp.80 – 99.

Weiss, D.J. and Kingsbury, G.G., 1984. Application of computarized adaptive testing to educational problems. *Journal of Educational Measurement*, 21 (4), pp.361 – 375.

Wilson M., and Adams, R. J., 1995. Rasch models for item bundles. *Psychometrika*, 60 (2), pp.181-198.

Wilson, K., Lizzio, A. and Ramsden, P., 1997. The development, validation and application of course experience questionnaire. *Studies in Higher Education*, 22 (1), pp.33 – 53.

Wilson, M. and Adams, R., 1995. Rasch models for item bundles. *Psychometrika*, 60 (2), pp.181 – 198.

Wilson, M., 1988. Detecting and interpreting local item dependence using a family of Rasch models. *Applied Psychological Measurement*, 12(4), pp.353 – 364.

Wolf, E.W., 2009. Item and rater analysis of constructed response items via the multi-faceted Rasch model. *In*: Smith, E.V. and Stone, G.E., eds. *Criterion referenced testing: practice analysis to score reporting using Rasch measurement models*. Maple Grove, MN: Jam Press, pp.71 – 88.

Wolfe F., (2006) Multiple Significance Tests. *Rasch Measurement Transactions*, 19 (3), p.1044.

Wright, B,D. and Masters, G.N., 1982. Rating Scale Analysis. Chicago: Mesa Press.

Wright, B. and Panchapakesan, N., 1969. A procedure for sample-free item analysis. *Educational and Psychological Measurement*, 29 (1), pp.23 – 48.

Wright, B.D. and Bell, S.R., 1984. Item banks: what, why, how. *Journal of Educational Measurement*, 21 (4), pp.331 – 345.

Wright, B.D. and Douglas, G.A., 1977. Best procedures for sample-free item analysis. *Applied Psychological Measurement*, 1 (2), pp.281 – 295.

Wright, B.D. and Stone, M.H., 1979. Best test design. Chicago: Mesa Press.

Wright, B.D., 1988. The efficacy of unconditional maximum likelihood bias correction: comment on Jansen, Van den Wollenberg, and Wierda. *Applied Psychological Measurement*, 12 (3), pp.315 – 318.

Wright, B.D., 1991. Factor analysis versus Rasch analysis of items. *Rasch Measurement Transactions*, 5 (1), pp.134 – 135.

Wright, B.D., 1995. 3PL or Rasch? *Rasch Measurement Transactions*, 9 (1), p.408.

Wright, B.D., 1995. Which standard error? Item-specific or general? Ideal or real? *Rasch Measurement Transactions*, 9 (2), p.436.

Wright, B.D., 1996. Local dependency, correlations and principal components. *Rasch Measurement Transactions*, 10 (3), pp.509 – 511.

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*, 65–104. Mahwah, NJ: Lawrence Erlbaum.

Wright, B.D., 2000. Conventional factor analysis vs. Rasch residual factor analysis. *Rasch Measurement Transactions*, 14 (2): p.753.

Wu, M.L., Adams, R.J., Wilson, M.R. and Haldane, S., 2007. ACER ConQuest Version 2.0: General item response modelling software: ACER Press.

Zhou, F., Wu, D., Yang, X. and Jiao, J., 2008. Ordinal logistic regression for affective product design. *In*: *International conference on industrial engineering and engineering management,* 8 – 11 December, Singapore.

Zwinderman, A. H., 1995. Pairwise parameter estimation in Rasch models. *Applied Psychological Measurement*, 19 (4), pp.369 – 375.

APPENDIX A

Subtest is the term used when two or more items are grouped producing a single polytomous item. Following the partial credit model (Masters, 1982) (see Section 2.4.5.3)

$$\Pr\{X_{nij} = x_{nij} | \beta, \delta\} = \frac{\exp \sum_{j=0}^{x} (\beta_n - \delta_{ij})}{\sum_{k=0}^{m_i} \exp \sum_{j=0}^{k} (\beta_n - \delta_{ij})} \tag{A.1}$$

where $\Pr\{X_{nij} = x_{nij} | \beta, \delta\}$ is the probability of a person n on location β_n responds in category x to item i. For parameterisation it is established that

$$\delta_{ij} = \delta_i + \tau_{ij} \tag{A.2}$$

where the average location of item i characterises a central single feature of the item, such that $\sum_{k=0}^{m} \tau_{ij} = -\tau_{i1} - \tau_{i2} - ... - \tau_{im} = 0$ where $\tau_{i0} = 0$ (Andrich, 1988b).

Given that the maximum score of a subtest is the sum of the maximum scores of the individual items involved such that $s_i = \sum_j x_{ij}$, Equation A.2 is then re-parameterised. Using the estimation procedure in RUMM2030 (see Section 2.4.6) and representing the spread of the scores by the parameter μ, which indicates the average half-distance between thresholds, the model takes the form

$$\Pr\{X_{nij} = x_{nij} | \beta, \delta\} = \frac{1}{\phi} \exp[x(m-x)\mu_i + x(\beta_n - \delta_i)] \tag{A.3}$$

where m is number of thresholds for $(m+1)$ categories and $\phi = \sum_{k=0}^{m_i} \exp \sum_{j=0}^{k} (\beta_n - \delta_{ij})$ is a normalising factor. The quadratic coefficient $x(m-x)$ of μ_i indicates the discrimination of the subtest. Additionally, the threshold locations can be recovered by the skewness. The model then is expressed by

$$\Pr\{X_{nij} = x_{nij} | \beta, \delta\} = \frac{1}{\phi} \exp[x(m-x)\mu_i + x(m-x)(2x-m)\eta_i + x(\beta_n - \delta_i)] \tag{A.4}$$

where η_i represents the average asymmetry from the average distance between thresholds (Andrich, 1985a).

The resulting location obtained from the re-parameterisation for the subtests can be represented by sum of the uncentralised thresholds derived by adding the location estimate to each centralised threshold. The mean of the set of uncentralised thresholds for a subtest is therefore the location estimate for a subtest (RUMM2030, 2012), such that

$$\delta_i^* = \frac{1}{(KM)} \sum_{k=1}^{KM} \tau_k \qquad \text{(A.5)}$$

where k represents a condition within a frame of reference, such that $k = 1,2...,K$, m represents the thresholds of a sub-item, such that $m = 1,2...,M$, and i represents the items in their subtest form, such that $i = 1,2,...,I$.

APPENDIX B

The novel approach developed in Chapter 5, Section 3.2 does not modify the original structure of the data. The alternative framework of subtests, which is formed by the original items instead of being structured by the conditions, does not conceal anomalies if they do exist in the data.

This can be demonstrated using the study of the confectionery stimuli based on an enlarged frame of reference (see Section 4.4.4). The preliminary non-calibrated 24-item set was replicated across the four wrapped confectioneries used as stimuli making 96 items. After estimating the stimulus locations, the sub-items were re-grouped in subtests formed by the original item. Table B.1 shows subtests that presented misfit in a preliminary analysis (Column *p*, with *p* <0.05, in bold). Misfit was also indicated by the subtests with residual ≥±2.50 (99% CI) (see Column *fit residual,* in bold). Analysis of response dependence, indicating a source of misfit, was carried out through the matrix of correlation for person-item residual. Table B.2 indicates correlations with *r* ≥ ±0.30 (in bold).

Table B.1 - Subtest fit statistics for the preliminary 24-item set from the previous confectionery study under the alternative framework.

Subtest	Location	*SE*	Fit residual	χ^2	Degree of freedom (*df*)	*p*
ST01	0.10	0.03	-1.26	12.73	4	**0.01**
ST02	-0.05	0.03	-0.82	7.55	4	0.11
ST03	0.31	0.03	**2.73**	16.50	4	**0.00**
ST04	-0.09	0.03	-0.67	11.37	4	**0.02**
ST05	0.33	0.03	1.40	6.05	4	0.20
ST06	-0.13	0.03	-2.28	29.56	4	**0.00**
ST07	-0.18	0.02	**6.09**	129.61	4	**0.00**
ST08	0.27	0.03	-0.84	10.08	4	**0.04**
ST09	0.10	0.03	**-2.55**	39.59	4	**0.00**
ST10	0.25	0.03	-0.79	15.87	4	**0.00**
ST11	0.13	0.03	-1.00	10.24	4	**0.04**
ST12	-0.30	0.03	-0.91	5.98	4	0.20
ST13	-0.24	0.03	-1.06	11.84	4	**0.02**
ST14	-0.36	0.03	-2.12	19.93	4	**0.00**
ST15	-0.19	0.03	-0.28	1.54	4	0.82
ST16	0.35	0.02	**4.21**	47.60	4	**0.00**
ST17	-0.27	0.04	-1.51	18.16	4	**0.00**
ST18	0.17	0.03	**5.78**	204.02	4	**0.00**
ST19	-0.32	0.03	-1.70	24.33	4	**0.00**
ST20	0.12	0.03	-1.39	24.19	4	**0.00**
ST21	-0.00	0.03	-1.07	15.71	4	**0.00**
ST22	-0.10	0.03	0.50	0.68	4	0.95
ST23	-0.09	0.03	-0.28	8.98	4	0.06
ST24	0.18	0.03	**3.66**	77.70	4	**0.00**
Overall fit				749.79	96	0.00

Table B.2 – Person-item residual correlation matrix for the confectionery subtests.

Subtest	ST1	ST2	ST3	**ST4**	ST5	**ST6**	**ST7**	**ST8**	**ST9**	**ST10**	ST11	ST12	ST13	**ST14**	ST15	ST16	**ST17**	**ST18**	**ST19**	ST20	ST21	ST22	ST23	ST24
ST1	1.00																							
ST2	-0.05	1.00																						
ST3	-0.14	-0.12	1.00																					
ST4	0.12	-0.18	-0.11	1.00																				
ST5	-0.03	-0.15	-0.05	0.10	1.00																			
ST6	0.23	-0.07	-0.16	0.24	-0.10	1.00																		
ST7	-0.23	-0.05	0.26	-0.23	-0.09	**-0.37**	1.00																	
ST8	0.06	-0.11	-0.22	0.05	0.10	0.09	**-0.35**	1.00																
ST9	0.02	-0.05	-0.21	0.23	-0.05	0.21	**-0.45**	**0.31**	1.00															
ST10	0.13	-0.19	-0.18	**0.37**	0.06	**0.30**	-0.29	0.17	0.25	1.00														
ST11	0.13	-0.13	-0.09	0.16	0.00	0.11	-0.21	0.02	0.06	0.09	1.00													
ST12	0.04	-0.05	-0.10	-0.05	-0.13	0.10	-0.16	-0.02	-0.01	0.06	-0.01	1.00												
ST13	0.01	0.02	-0.13	-0.21	-0.21	0.08	-0.16	-0.05	-0.03	-0.12	-0.11	0.15	1.00											
ST14	-0.05	0.29	-0.09	0.06	-0.10	0.04	-0.11	-0.10	0.25	-0.01	-0.06	0.07	0.05	1.00										
ST15	-0.11	0.19	-0.13	-0.26	-0.04	-0.15	-0.08	-0.10	-0.20	-0.21	-0.10	0.07	0.12	0.00	1.00	1.00								
ST16	-0.15	-0.07	0.20	-0.10	-0.08	-0.18	0.23	-0.13	-0.20	-0.22	-0.22	-0.24	0.03	-0.18	-0.18	-0.20	1.00							
ST17	0.16	-0.13	-0.16	0.29	0.05	0.27	-0.26	0.12	0.13	**0.31**	0.19	0.04	-0.07	0.04	-0.12	0.23	**-0.34**							
ST18	-0.29	-0.09	0.18	**-0.33**	0.11	**-0.42**	**0.49**	-0.21	**-0.39**	**-0.38**	-0.15	-0.15	-0.10	-0.29	-0.02	-0.27	0.10	1.00						
ST19	0.05	-0.03	-0.08	-0.03	-0.21	0.06	-0.16	0.06	0.11	0.04	-0.02	0.01	0.00	0.07	0.04	-0.18	0.05	-0.27	1.00					
ST20	0.07	-0.03	-0.24	0.13	-0.11	0.14	**-0.34**	0.26	**0.37**	0.17	-0.01	0.00	-0.03	-0.02	-0.05	-0.24	-0.07	**-0.39**	0.28	1.00				
ST21	-0.10	0.06	-0.10	-0.03	-0.10	-0.07	-0.05	-0.01	0.12	-0.05	-0.06	-0.09	0.00	0.05	0.12	-0.05	-0.05	-0.17	**0.31**	0.19	1.00			
ST22	0.11	-0.02	-0.05	0.01	-0.16	0.04	-0.04	-0.11	-0.09	0.05	0.11	0.03	-0.05	-0.07	0.07	-0.05	-0.20	-0.06	-0.08	-0.09	-0.11	1.00		
ST23	-0.07	-0.03	-0.21	-0.05	0.09	0.01	-0.22	0.03	0.06	0.01	0.03	0.02	0.08	0.02	0.11	-0.20	0.02	-0.05	0.05	-0.01	-0.07	-0.16	1.00	
ST24	-0.21	-0.02	0.16	-0.28	0.07	-0.29	0.26	-0.15	**-0.31**	-0.29	-0.14	-0.15	-0.12	**-0.30**	0.03	0.22	**-0.36**	**0.46**	-0.27	-0.21	-0.19	-0.06	-0.09	1.00

APPENDIX C

Comparison between data reduction using two samples was carried out through principal component analysis (PCA) using the software package SPSS version 20.0. The 24-item set of the confectionery experiment (see Chapter 3) was used for establishing the comparison. The original sample of 306 participants was split in two halves forming a group of respondents with identity from 1 to 153 and another group from 154 to 306. Person's scores for all stimuli were averaged for every statement. The matrix **Y** was constituted of *n x m* elements, where *n* is a person identity and *m* is an item. The correlation matrix $\mathbf{Y}^\mathbf{T}\mathbf{Y}$ with *m x m* elements allowed the extraction of the principal components. Varimax rotation method with Kaiser normalisation was used to obtain the item loadings.

Kaiser-Meyer-Olkin test pointed to the sampling adequacy for PCA presenting KMO >0.80 (Kaiser, 1970) (Table C.1). Bartlett's test of sphericity was significant, indicating that the correlations between items were sufficiently large for PCA (Table C.1). The extraction resulted in four components with eigenvalues greater than one. After rotation the variance explained by the extraction taking into account the four components was nearly 55% for the first half and nearly 52% for the second half (Table C.2).

Table C.1 – KMO test and Bartlett's test for the first and second half of the sample.

Sample		First half	Second half
Kaiser-Meyer-Olkin measure of sampling adequacy		0.90	0.89
Bartlett's test of sphericity	Approximate χ^2	1658.05	1402.66
	df	276	276
	Sig.	<0.001	<0.001

Table C.2 – Combined variance explained for the first half and the second half of the sample.

Component	First half						Second half					
	Initial eigenvalues			Rotation sums of squared loadings			Initial eigenvalues			Rotation sums of squared loadings		
	Total	% of variance	Cumulative %	Total	% of variance	Cumulative %	Total	% of variance	Cumulative %	Total	% of variance	Cumulative %
1	8.55	35.64	35.64	6.36	26.52	26.52	7.84	32.66	32.66	5.44	22.65	22.65
2	1.97	8.19	43.83	3.86	16.08	42.60	1.91	7.97	40.63	4.04	16.82	39.47
3	1.57	6.53	50.36	1.64	6.82	49.42	1.39	5.80	46.43	1.55	6.46	45.92
4	1.25	5.19	55.55	1.47	6.13	55.55	1.29	5.37	51.80	1.41	5.88	51.80

Table C.3 shows the rotate component matrix for the first half and for the second half of the sample. The item loads chosen for each component are displayed in bold. Loads displayed in bold and italic show items that changed the correspondence with a component when analysing the two halves of the sample. Figure C.1 and Figure C.2 display the scatter plot matrix for the first half and for the second half of the sample respectively. For both cases the bivariate relationship between components resembles normality, indicated by the curves in the diagonal of the matrix, although they are not related linearly.

Table C.3 – Rotated component matrix for the first half and for the second half of the sample.

Variables	First half				Second half			
	Component				Component			
	1	2	3	4	1	2	3	4
Item 1	**.660**	.309	-.022	-.125	**.456**	.393	.291	.090
Item 2	.099	***.755***	.150	-.097	***.741***	.052	.119	.106
Item 3	.214	.188	**.661**	.048	-.003	-.026	**.708**	-.212
Item 4	***.714***	.092	.181	-.146	.132	***.838***	.107	.089
Item 5	.251	.063	-.247	***.611***	.176	***.517***	.104	.392
Item 6	**.841**	.241	.032	-.064	**.531**	.374	.308	.147
Item 7	-.305	-.048	***.636***	-.074	.182	-.282	.152	***-.678***
Item 8	***.670***	.272	-.174	.158	.462	***.478***	-.013	.231
Item 9	***.681***	.426	.117	.020	.499	***.619***	-.006	.138
Item 10	***.793***	.123	-.091	.048	.278	***.712***	.129	-.087
Item 11	**.692**	.163	.138	.138	**.400**	.387	.072	.069
Item 12	**.470**	.460	-.009	-.016	**.623**	.180	.093	.006
Item 13	.397	***.645***	.144	.077	***.599***	.005	.226	.114
Item 14	.385	***.660***	.206	-.121	***.584***	.335	.156	-.087
Item 15	-.005	***.780***	-.113	.209	***.758***	.025	-.090	.017
Item 16	.096	.080	**.582**	.076	.120	.116	**.711**	.221
Item 17	***.740***	.114	-.154	-.008	.433	***.574***	.279	-.101
Item 18	-.477	-.126	.087	***.593***	-.029	***-.570***	.103	.079
Item 19	.511	***.539***	.030	-.060	***.633***	.274	.046	-.300
Item 20	**.647**	.466	.016	-.132	**.522**	.489	-.116	.147
Item 21	.326	***.627***	.103	.015	***.665***	.282	-.046	-.288
Item 22	**.418**	.140	.171	-.004	**.457**	.197	.161	-.153
Item 23	**.450**	.402	-.237	.296	**.606**	.218	-.090	.164
Item 24	-.080	.029	.249	**.680**	.241	-.359	.220	**.536**

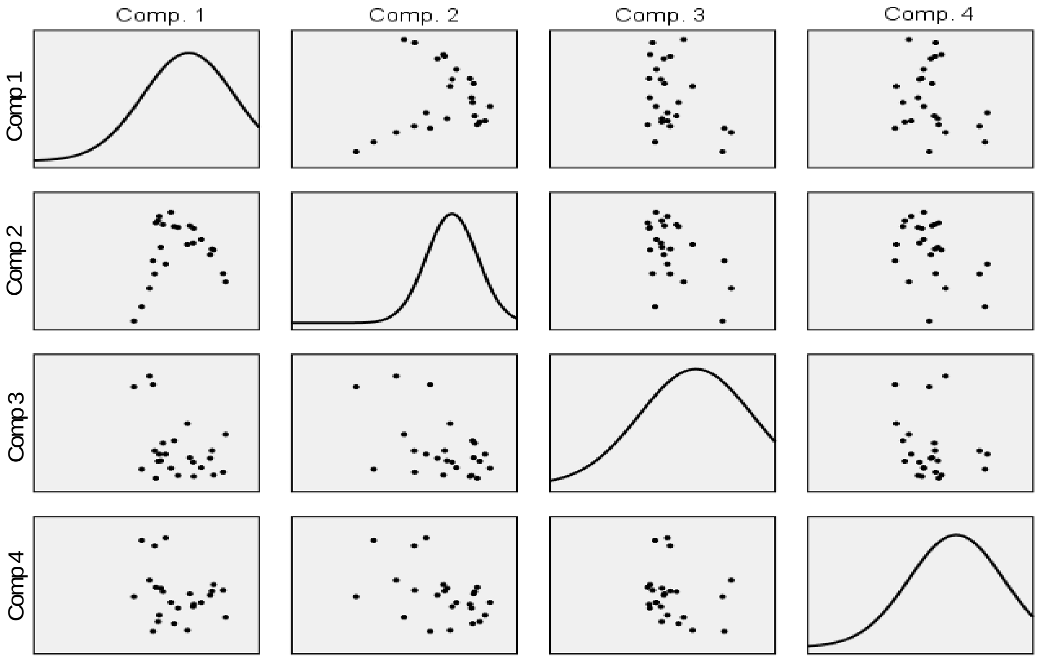


Figure C.1 – Component scatter plot matrix in rotate space for the first half of the sample.

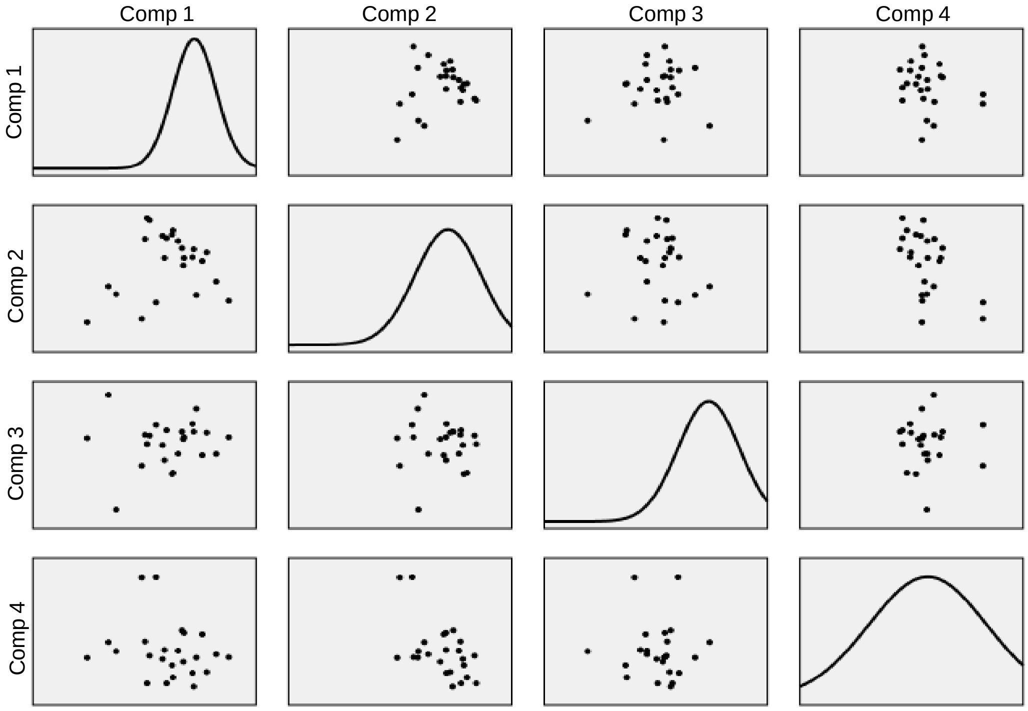


Figure C.2 - Component scatter plot matrix in rotate space for the second half of the sample.