Establishing Metrics for Kansei Responses: An Approach Using the Rasch Model

F R Camargo[1] and B Henson[2]

[1] University of Leeds, School of Mechanical Engineering, United Kingdom, mnfrc@leeds.ac.uk
[2] University of Leeds, School of Mechanical Engineering, United Kingdom, B.Henson@leeds.ac.uk

Abstract: Although qualitative comparisons are necessarily part of the process to elicit users' kansei responses to products, they are insufficient to provide a more fine-grained interpretation of users' interaction, which can require a measurement system. However, kansei variables cannot be measured directly. Data obtained from kansei responses need to be transformed by statistical methods and meet measurement assumptions. An approach to validate the quantitative structure of kansei scales is the application of Rasch measurement theory. The Rasch model, which is referred to as a family of probabilistic models, provides mechanisms to test the hypothesis that the observations meet the assumptions for establishing a quantitative structure. In this paper a number of procedures in Rasch modelling are outlined. Different examples from empirical applications using some techniques of kansei engineering show that the establishment of measures for comparisons between individuals and between stimulus objects is not a trivial matter.

Keywords: Kansei Engineering, Kansei Measurement, Rasch Model, Validation

1. INTRODUCTION

In his annual letter for the Gates Foundation on how important measurement is to improving the human condition (2013), Bill Gates used the words from Willian Rosen about steam power (Rosen, 2010). To establish steam power as a powerful invention, innovative ways to measure energy consumed by engines were necessary. Measurement, therefore, allowed inventors and engineers to compare incremental changes, which were expected to lead to improvements in engines. To Gates, there is an important lesson to be learnt from Rosen's observation: "*Without feedback from precise measurement, Mr. Rosen writes, invention is doomed to be rare and erratic* (Gates, 2013)." Gates' understanding on the measurement issue can clearly be identified in current research in a diversity of domains. Precise measurement has therefore extrapolated the domains of natural sciences, such as comparisons between students' performance in education, responses to treatments for depression in clinical assessments and consumers' satisfaction with regard to products, to mention but a few. Kansei engineering (KE) is not an exception. Applications abound about the notion that more or less of an element affects in different degrees the users' impression with regard to a kansei attribute of the product.

In the process, however, formal measurement cannot be established if too little information is available, usually at the beginning of the development of a new product. In that case, measurement using qualitative tools might take place. This includes pilot studies for identifying unwanted interactions between variables and the relevant construct and qualitative dimensions using statistical tools such as principal component analysis (Barnes and Lillford, 2009). Furthermore, measurement theory characterises that not every property of an observed event can be numerically represented. Under this perspective, kansei responses cannot be established as objective measures if they do not meet measurement assumptions. On the other hand, quantitative properties can be achieved if a relevant scale presents additivity, constant unit and invariant comparisons, which are subjected to theoretical foundation and empirical validation (Andrich, 1988). In other words, scales for kansei words are typically established by convention (e.g., -3 to +3, 0 to 7 or strongly disagree to strongly agree) while the quantitative characteristics are determined by the mathematical model used to their construction.

The purpose of measurement interpreted in this paper is to provide a consistent way to establish a shortened account of responses that people make to express their attitude and feelings with regard to a kansei attribute. The aim of the paper is to outline a novel approach to construct measurement scales for kansei responses underpinned by Rasch measurement theory (Camargo and Henson, 2011), which is commonly referred to as the Rasch model (RM), after the Danish mathematician Georg Rasch who developed it in the 1950s (Rasch, 1960, 1980). The RM is the denotation of a family of probabilistic models associated with latent trait theory in more recent psychometric approaches. Measurement is obtained from a combination between persons' responses and independent variables, called items. The RM's property of separability of the parameters for persons and items allows the design of a range of kansei words or statements that are used as a yardstick in a scale. Furthermore, the separability of parameters allows the comparison between any pair of individuals of a sample independently of the items and the comparison between any pair of items independently of the persons used to establish the scale.

In contrast with more typical data modelling that tries to fit a model to data, the RM's procedures test the data against measurement assumptions. Thus, if kansei data fit the model, it is possible to establish a scale with quantitative properties, referred to as a metric. However, in kansei practice it is very unlikely to find data that perfectly fit the RM. For this reason, most of the procedures in Rasch analysis test the data for anomalies, identifying sources of misfit and to what extent they corrupt measurement.

2. MEASUREMENT ASSUMPTIONS

The assumptions for making measurement possible go side by side with what one wants to make with the values obtained from measuring an object. There is frequently a formal difference between the kind of assignment of numbers originating from different procedures of measurement. For example, one could be interested in the classification of athletes at the finish line of a marathon. This would be possible taking note of the order of each athlete when crossing the line. The assumption here is that the first athlete crossing the finish line is the fastest. However, one could not be able to compare time performances between athletes without a calibrated chronometer. The assumption for this case would be the fastest athlete spends less time to cover a certain distance. Performance between athletes could therefore be compared by objective measures. In terms of kansei attributes, the measurement assumptions depend on the empirical, qualitative hypothesis that supports the association of the relevant kansei variables with physical properties of a product.

Measurement is fostered in this paper as a way of making meaningful inferences on kansei variables based on the numbers obtained from observed events. The main assumption is that the numbers represent a property of the relevant attribute. That is, a metric 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 (Krants et al., 1971; Luce and Tukey, 1964).

Another assumption is that a relevant kansei process can adequately be modeled as a systematic element, which is established by the mathematical model over operating conditions, and a random component, which is established by the measurement error associated with the estimation of parameters. This distinguishes the model from more typical approaches in the domain, which fail when conveying information about different source of errors (i.e., systematic and random), providing just one estimate of standard error for all respondents.

3. ESTIMATION OF PARAMETERS

In Rasch modelling the calibration of measures for items and persons is based on estimation of parameters (Wright and Master, 1982). 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 to calculate the standard error for each estimate through a second derivative of a likelihood function (Linacre, 1999).

Estimates of the person locations and item locations are preliminarily made according to the rating scale model (Andrich, 1978) or the partial credit model (Masters, 1982) 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. However, the iterative process is laborious, requiring a computer-intensive solution for practical purposes (Camargo and Henson, 2012a). The framework of an analysis can therefore have slight differences according to the software package used. Nevertheless, in general terms, after the estimates have been made, Rasch analysis is carried out to evaluate the extent to which the data fit the model.

4. ANOMALIES IN A DATA SET

Data from kansei responses to product features can present many symptoms of anomaly to fit the model. In KE it is very likely that persons do not give ratings 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 RM perspective, this is an indicative of a structure that is not quantitative. In this case, inferences made from the statistical results cannot be generalised beyond the sample studied and the scores cannot be considered as an element of a measurement structure.

However, there are different degrees of misfit in terms of a measurement structure which is originated from physical interactions 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. Furthermore, theoretical cut-offs are useful benchmarks although they ought not to be taken as a sole basis to make a decision. In general, such as in physics, misfit ought to be considered an anomaly in the data and substantively investigated.

5. EXAMPLES OF DISCREPANCIES TO THE MODEL BASED ON TWO EMPIRICAL STUDIES

5.1. Validity of the category scoring system

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. The transitions between categories can be interpreted as though there was an independent response for each of the thresholds. This allows the identification of potential problems with the empirical order of categories. If the response patterns are consistent, each response category has a point along the ability continuum where the most probable response is identified.

An example of inconsistent transitions between categories is the item "it is easy to know how much of the product is left in the packaging" taken from a study on containers of some products (Camargo and Henson, 2012b). In the study, 120 participants gave their ratings for five squeezable, everyday products. The scope of the study was to measure the relative importance of the packaging material for obtaining an intuitive impression of a moisturizer cream as a product feature when squeezing its container. Respondents, without seeing the products, used a self-report questionnaire that contained 16 statements based on the understanding of the product context to assess the required kansei attribute for each confectionery. The scale was designed as five-point Likert-style response options, i.e., strongly disagree, disagree, neutral, agree, and strongly agree with associated scores of 0, 1, 2, 3 and 4, respectively. Figure 1 shows that respondents did not use the category system for the item as it was designed by the analyst. The transitions do not discriminate between adjacent categories.

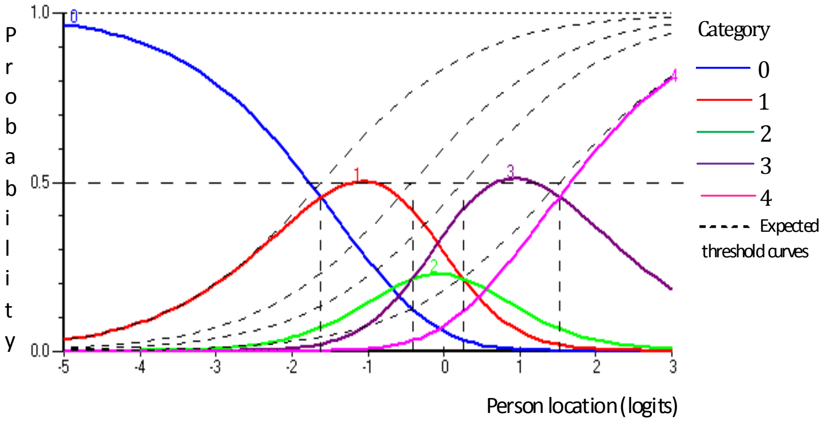


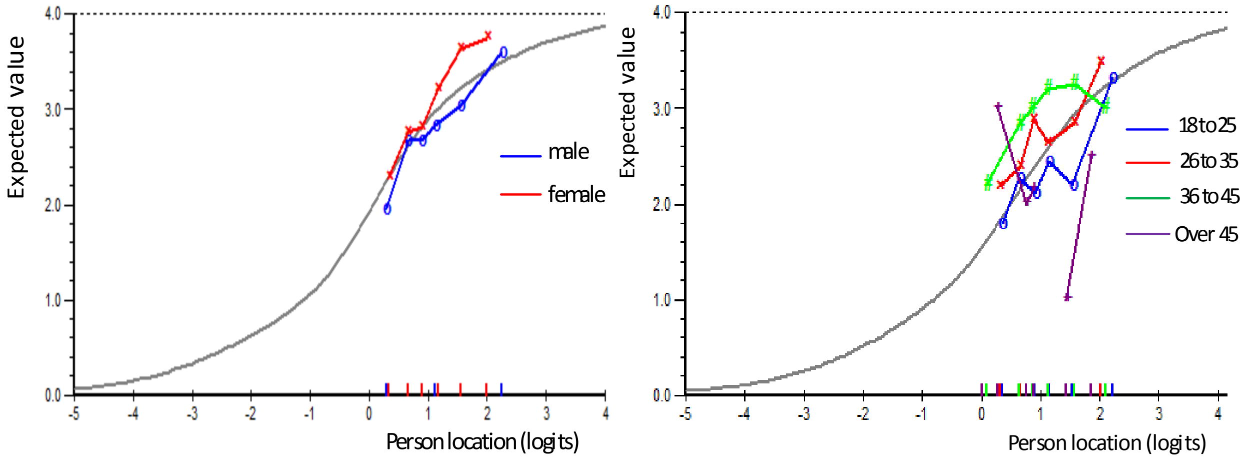
Figure 1: Disordered thresholds of an item from the study on packaging.

5.2. Differential item functioning

Another source of misfit in the data with regard to the model is denoted differential item functioning (DIF) or item bias (Broderson et al., 2007). DIF is presented when a group demonstrates consistently greater inclination to endorse an item than another group. This could be found in male and female groups, social classes, different professional groups, leisure activities groups, different age groups, cross-cultural investigations and cross-national studies.

DIF was, for example, identified in a study on the specialness of confectionery (Camargo and Henson, 2011). In the study, 306 participants gave their ratings for four pieces of wrapped confectionery. Respondents used a self-report questionnaire that contained 24 statements based on the understanding of the product context to assess the required latent attribute for each confectionery. DIF was detected through a two-way analysis of variance, indicating differences of responses between sexes as well as age groups.

Figure 2a shows that female respondents were more inclined to endorse the item “this chocolate is like a little present for me” than male respondents. Figure 2b indicates that the item “this chocolate would be nice during a break from housework” was more difficult to endorse by respondents in the age range from 18 to 25.



(a) (b)

Figure 2: DIF identified in the study on specialness of confectionery for (a) sex and (b) age groups.

5.3. Response dependence

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. Tests for local dependence identify anomalies that do not allow a scale to perform independently. 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 technique for assessing those violations is to identify patterns of high correlations among the standardised item residuals. The definition of high correlation can vary according to the context of use of the scale. One of the approaches is firstly to examine items with a coefficient of correlation higher than 0.30 between their standardised residuals (Tennant and Conaghan, 2007).

In the study on packaging mentioned in Section 5.1, for example, the preliminary analysis indicated dependence between the item “the product inside the container would spread easily” and the item “the product in this packaging is likely to flow easily” for one of the stimuli. Although the items were designed to measure two different impressions, the dependence suggests that they are redundant under the respondents’ point of view. Redundant items can mislead inferences or decisions made on account of means and standard deviations. Redundancy has also the effect of inflating reliability indices and item discrimination estimates (Marais and Andrich, 2008).

Another example from the study on packaging is the dependence indicated by a high correlation between the residuals of the item “the container feels only half filled when squeezing it” and the item “I might get a bit watery product in this container”. Apparently, those items are not redundant.

However, both of them presented misfit to the model. This could be a consequence of the absence of independent or relevant information in the context of the study.

5.4. Dimensionality

In Rasch modelling, trait dependence represents the violation of the assumption of unidimensionality in the measurement structure. The violation 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 KE studies, 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 for chocolate could, for example, contain items associated with peoples' affective responses to sharing among friends although the analyst is exclusively interested in the persons' feelings related to specialness.

Another point to be emphasised is that the concept of unidimensionality in the RM refers to independent variables that work well all together in a metric. It is noteworthy that in KE not only the variables ought to work well all together but also concurrently for every relevant stimulus object. An instance of violation of the assumption would be a pool of items that work well for two stimuli although in a different way for each stimulus. This situation can take place if the metric is calibrated individually. The reason is that differences in responses pattern can provoke different score systems across stimuli.

An example of violation of unidimensionality of this type is taken from the study on the specialness of confectionery (see Section 5.2). In that study a pool of 12 items was carefully calibrated for each confectionery separately. The system constituted a sound basis of measurement for each individual confectionery (Camargo and Henson, 2011). However, despite a common set of calibrated items, the stimuli cannot be directly compared through those scales. If the four 12-item scales had worked similarly, the locations of the items on the continuum for the different stimuli would be similar as well, taking into account their measurement errors. Figure 3 shows the results from t-tests between locations of items that belong to scales obtained from independent calibrations for two pieces of confectionery with higher endorsement for specialness. The high number of t-tests that fell out of 95% of the confidence interval indicates that the items drifted into different locations when participants considered the stimuli. That is, the statements in the scales had different interpretations according to the stimulus. One solution to overcome this problem is to use the stimulus objects as an independent parameter using the faceted RM.

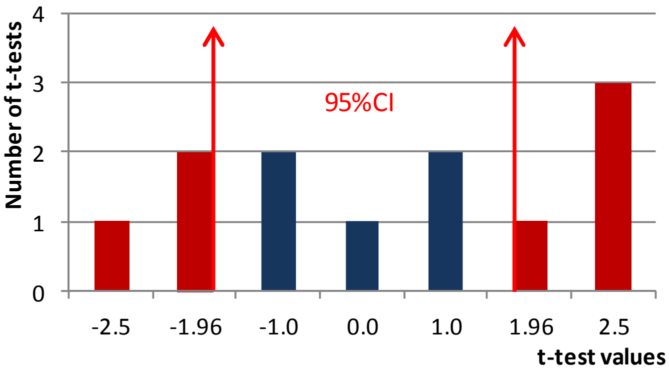


Figure 3: T-tests results from the comparison between two scales for the specialness of confectionery.

6. FUNCTIONING OF A RASCH-CALIBRATED METRIC FOR KANSEI RESPONSES

A derivation of the RM, namely the many-facet Rasch model (MFRM) developed by Linacre (1989), has been adapted by Camargo and Henson (2012b) for applications in KE. A facet is defined as a component or variable of the measurement condition that is assumed to affect the scores in a systematic fashion. There are at least three facets in the particular case of kansei metrics: persons, items and stimulus objects. The facets of the metric are represented by independent parameters sharing the same linear continuum. When such parameters are combined, it is possible to obtain the probability of a person's endorsement to any item for any stimulus.

A metric attains interpretation when the difference between persons as well as between items is established by the distance between different locations. Those differences are interpreted through the adapted MFRM, such that the probability of a response X is obtained by a function f of the locations (Camargo and Henson, 2012b) given by

$$\Pr(X) = f(\beta - \delta + \zeta) \tag{1}$$

where β represents the person parameter, δ is the item parameter and ζ represents the stimulus object. Thus,

1. The person parameter indicates the inclination of endorsement to any item and any stimulus object. That is, the more the readiness, the higher the probability of affirming an item.
2. The item parameter indicates the difficulty of endorsement. An easier item is endorsed by relatively more respondents than a more difficult item. That is, the easier the item, the more likely it will be affirmed.
3. The stimulus parameter indicates the kansei fulfilment. The more the attribute is fulfilled by the stimulus, the more likely it will be endorsed.

Figure 4 displays the generic representation of a metric for kansei responses using the RM. Column Facet 1 represents the persons' inclination of endorsement. Column Facet 2 indicates the difficulty of kansei words or kansei statements and Column 3 represents the fulfilment value or degree of the kansei attribute. It is noteworthy that all facets are on the same linear continuum.

The interpretation of probability of a response is established by the distance amongst a person, an item and a stimulus object. Zero distance amongst them indicates a probability of 50% of agreement. This is the case for Paul at location 0 with regard to Item 1 and Stimulus 4. Taking the same item and stimuli, Mary, David and Juan will have higher probability of endorsement, and Lucy a lower probability if they are compared with Paul.

One important characteristic of a Rasch-calibrated metric is the interpretation based on invariant comparisons. Note that the comparison between Mary and David or Mary and Lucy does not depend on any item in the metric and any stimulus object either. This allows the interpretation that David and Juan share the same degree of agreement with regard to the kansei attribute. Using a symmetric argument the comparison of difficulty between any pair of items and the fulfilment value of the kansei attribute between any pair of stimulus objects can also be made.

The unit in Rasch modeling is the log-odds unit, usually called logit. The logit is the distance on the continuum that indicates changes of the odds of observing the relevant event computed through Napierian logarithm (Linacre and Wright, 1989). The unit in logit denotes the same interval with regard to changes on the continuum.

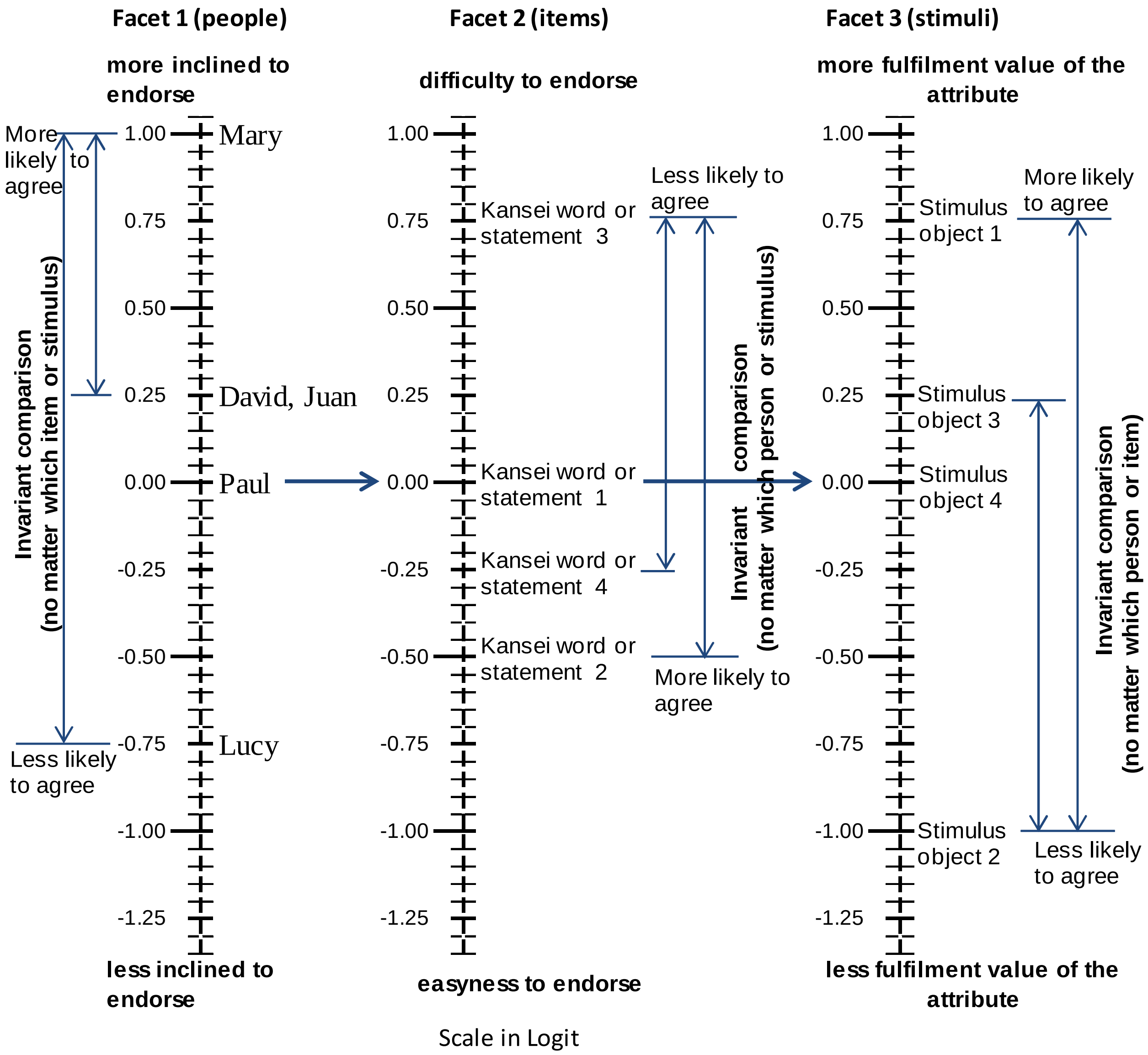


Figure 4: Generic representation of a metric for kansei responses using the faceted RM.

7. IMPLICATIONS

The feasibility of applying a measurement model in KE depends on the aims of an investigation. In the case of simple observations, the assumption could be whether the kansei attribute in a product was absent or present under certain conditions or whether it was manifested to some extent or to a greater extent. In this case, the investigatory observations are qualitative. As a result, the identification and classification interpreted from the data are nominal. On the other hand, if the assumption is that an observed event is better than another, then it is necessary to assign numerical values to those events. The numerical values, in that situation, follow an order for identifying which event will hold more or less of a kansei attribute. Furthermore, if algebraic operations on the numerical values are required, the measurement assumptions, which were briefly mentioned in Section 2, have to be met.

The RM provides theory and procedures to examine how well the kansei data fit together and cooperate to define the attribute being measured. The RM properties are important in KE studies when the consistency or stability of the scores are necessary for testing and comparing other similar objects using different sample of respondents. Those properties are also important to validate the inferences or interpretations that one makes from the test scores.

Given evidence from studies in different domains of knowledge using the RM, it is possible to envisage that a well-defined scale of measurement has potential applications in many KE settings. Defining subpopulations according to their differences is a far more advantageous approach to manage kansei attributes of a product than an entire population. Nevertheless, users' experience with regard to a kansei attribute is frequently idiosyncratic. Therefore, research taking into account different sub-groups of people, such as cross-cultural and cross-national studies, is strongly recommended for generalising the findings to more worldwide products.

One of the benefits from the application of the RM in KE is to avoid a number of pilot studies using individual variables in test groups. After calibration, smaller samples are necessary to validate results when comparing with typical approaches in the domain. Furthermore, calibrated metrics can measure oscillations in kansei responses to a product as a consequence of contextual conditions throughout the life-cycle of a product.

8. CONCLUSION

In this paper a number of the procedures in Rasch modelling to test a data set for a quantitative structure were outlined. Different examples from empirical applications using some techniques of kansei engineering show that the establishment of measures for comparisons between individuals and between stimulus objects is not trivial. A novel approach to measure kansei responses is the application of the RM. The purpose is to establish a consistent way to interpret and compare a shortened account of responses that people make to express their attitude with regard to a kansei attribute.

One outcome from the application of the RM in KE is the establishment of calibrated metrics. Those metrics have allowed the comparison between any two persons independently of the kansei variables used in a study and the comparison of the any pair of variables independently of the persons used to calibrate the scale. This property is relevant to validate the inferences or interpretations that one makes from test scores.

REFERENCES

Andrich, D. (1978). Application of a psychometric rating model to ordered categories which are scored with successive integers. Applied Psychological Measurement, 2(4), 581 – 594.

Andrich, D., (1988). Rasch models for measurement. Sage university papers series on quantitative applications in the social sciences, No. 68, London: Sage.

Barnes, C. and Lillford, S. (2009). Decision support for the design of affective products. Journal of Engineering Design, 20 (5), 477 – 492.

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), 309 – 324.

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), 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), 204 – 219.

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

Fisher, R. A. (1922). On mathematical foundations of theoretical statistics. Philosophical Transactions of the Royal Society of London (A) 222, 309 – 368.

Gates, B. (2013). Measuring progress. Retrieved November 15, 2013, from Bill and Melinda Gates foundation web site: http://annualletter.gatesfoundation.org

Krantz, D.H., Luce, R.D., Suppes, P. and Tversky, A., 1971. Foundations of measurement, Vol. 1. New York: Academic Press.

Linacre, J.M. (1989). Many-facet Rasch measurement. Chicago: MESA Press.

Linacre, J.M. (1999). Understanding Rasch measurement: estimation methods for Rasch measures. Journal of Outcome Measurement, 3(4), 382 – 485.

Linacre, J.M. and Wright, B.D. (1989). The "length" of a logit. Rasch Measurement Transactions, 3(2), 54 – 55.

Luce, R.D. and Tukey, J.W. (1964). Simultaneous conjoint measurement: a new type of fundamental measurement. Journal of Mathematical Psychology, 1(1), 1 – 27.

Marais, I. and Andrich, D. (2008). Effects of varying magnitude and patterns of response dependence in the unidimensional Rasch model. Journal of Applied Measurement, 9(2), 105 – 124.

Masters, G.N., 1982. A Rasch model for partial credit scoring. Psychometrika, 47 (2),149 – 174.

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.

Rosen, W. (2010). The most powerful Idea in the world: a story of steam, industry and invention. London: Random House.

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), 1358 – 1362.

Wilson, K., Lizzio, A. and Ramsden, P. (1997). The development, validation and application of course experience questionnaire. Studies in Higher Education, 22(1), 33 – 53.

Wright, B,D. and Masters, G.N. (1982). Rating Scale Analysis. Chicago: Mesa Press.