MEASURING AFFECTIVE RESPONSES TO CONFECTIONARIES USING PAIRED COMPARISONS

Farzilnizam AHMAD[a], Raymond HOLT[a] and Brian HENSON[a]

[a]*Institute Design, Robotic & Optimizations (IDRO), School of Mechanical Engineering University of Leeds, Leeds, LS2 9JT, United Kingdom*

ABSTRACT

The use of category scales on self-report questionnaires in Kansei Engineering can be subject to biases and errors. In this research, the use of Rasch analysis of paired comparisons of products to derive linear measurement of affective response is tested. Four pieces of confectioneries and twelve evaluative statements measuring the dimension of specialness that was validated in previous researches were used. A computer-based self-report system presented one hundred and fifty-seven participants with pictures of pairs of confectionery and the evaluative statements in all combinations, and the participants were asked to indicate which confectionery satisfied the statement best. The analysis demonstrates the viability of using Rasch analysis to obtain measures of affective response from paired comparisons, that people find it easier to make paired comparisons compared with evaluating products separately against Likert statements, but that in this case the fit of the data to the Rasch model is very poor.

Keywords: Paired comparison, Affective Engineering and Rasch Model

1. INTRODUCTION

In Kansei Engineering aproach Type III (Nagamachi, 1995), one attempts to measure people's affective responses to products so that relationships between the affective responses and the products' physical properties can be identified and used to improve design. In Nagamachi's approach, based on Osgood's semantic differential technique (Osgood, Suci and Tannenbaum 1957), people are asked to rate products against a number of adjective pairs on typically either five or seven point scales. A data reduction technique such as the Principal Components Analysis are used to establish a semantic space in which the physical properties of the products and the affective responses can be correlated.

Whilst intuitive for the research participant, the use of category scales can suffer from inaccuracies and biases. It is assumed that adjective pairs plotted in semantic space are equidistant from the neutral point (Heise 1969) and that common word-pairs are antonyms (Mordkoff 1963). Osgood (1971) demonstrated that the use of the semantic differential technique with the same stimulus that produced large numbers of failures to respond and decreased the speed of responses. Other sources of variance associated with reliability, factor scores and group means, can cause inaccuracies (Borsboom 2006). Category scales are usually treated as interval data, when they are at best ordinal (Wright and Linacre 1989, Stevens 1946).

There is some evidence from our own research that participants are often unable to clearly discriminate between the different categories of semantic differential scales during product evaluation (Camargo and Henson, 2011). In this research, a probabilistic model is applied, the Rasch model (Rasch, 1960, 1980), in order to measure people's affective responses to products. In this approach, rather than constructing a statistical model of the response data, it is determined whether the response data fits the Rasch model and, if it does, some measurement properties are demonstrated. The Rasch model, in the context of product evaluation, calculates the probability that someone will endorse a product as a mathematical function of the person's ability or willingness to endorse and the difficulty of endorsing the particular product. The result is a linear scale of how easy it is to endorse a product along the affective dimension the instrument is designed to measure. The Rasch model had been used widely in education and in medicine. Previously, various instances were observed in which the probability of participants endorsing each category on a response scale is not sequentially ordered, when it was expected to be (Camargo and Henson, 2011). Thus, while categorical response scales are intuitive, participants are not able to be easily discriminated by the categories of the response scale.

One reason why category responses might be troublesome could be because participants are asked to evaluate products separately, without reference to a benchmark product. One of the aims of this work is to establish whether participants might find it easier to evaluate products if the evaluations were made as paired comparisons. When making a paired comparison, the participant merely has to indicate which of the two products they endorse is more ready, rather than thinking about which category of response one would elicit separately. The challenge is then to derive a linear scale of affective response from such comparisons. There is a body of work from discrete choice theory which is based on making paired comparisons of products (Train, 2009). The aim in discrete choice theory, however, is to determine the relative importance of properties of the choices, which are often assumed to vary linearly, rather than to derive measurement. Thurstone's law of comparative judgement' (Thurstone, 1927) can be used to establish measurements from pairwise comparisons. It has been widely used in psychophysics to determine the relationship between perception and intensity of stimulus. The Bradley-Terry-Luce model (Bradley and Terry, 1952) is derived from Thurstone's law, but uses a slightly different statistical basis. In the context of education, it can be used to derive measurement from whether answers to questions are right or wrong. The Bradley-Terry-Luce model can be shown to be

equivalent to one of the forms of the Rasch model. However, the model is not directly applicable to evaluation of products because, whilst in the educational case there is a response associated with each question for each person, in the context of product evaluation there are questions and responses for each person for each product. In other words, the product forms an extra independent factor or facet for which the Bradley-Terry-Luce model and Thurstone's law cannot account. There are, however, forms of the Rasch model that might be able to account for the extra facet (Linacre, 1989).

The aims of the research are therefore to establish whether linear measurement of affective response can be derived from paired comparisons of products, to assess whether making paired comparisons are easier for participants, and to determine the quality of people's responses, which in this case is determined by responses that fits the Rasch model.

The approach taken is to use statements developed in previous research intended to measure the specialness of confectionary (Camargo and Henson, 2011). The previous research established that the statements can be used as a unidimensional instrument for measuring affective response. In the research reported here, the statements are used again to evaluate the same confectionary, but instead of rating each confectionary separately against Likert statements, the user is presented with all confectionary in all pair combinations, and the participant indicates which pair satisfies which evaluative statement. The responses were then analysed to determine their fit to the Rasch model. The analysis of the paired comparison data is compared with those of the original research which used Likert statements. It is concluded that the use of the Rasch Model analysis to derive a linear measurement of affective response from paired comparisons is viable, that people find it easy to make paired comparisons, but that in this case the fit of the data to the Rasch model is very poor.

2. METHOD

In previous research, 306 participants are selected to rate four pieces of confectioneries against twenty-four Likert statements on a five-point scale, related to the construct of specialness (Camargo and Henson, 2011). Four items of confectioneries that are readily available were used for these experiments. These confectioneries were Ferrero Rocher®, Lindor®, Caramel®, and Milky Way® from a Mars Celebrations® assortment (Figure 1). The confectionery was chosen because, it is viewed that they are likely to elicit different responses to statements pertaining to their specialness. The statements used in the experiment were determined through UK-based consumer research by a large confectionery company. The responses to the four pieces of confectioneries were analysed using the multi-faceted Rasch model to establish a unified scale, for which twelve statements fitted the model for all four pieces of confectionery (Table 1). The experiment established a linear scale for the measurement of the specialness of confectionery.

In the new research study, participants were asked to endorse the confectionery against the statements for specialness by paired comparisons. One hundred and fifty-seven participants,

(eighty-three males, seventy-four females) were recruited to take part in this study with an age range from 17 to 57 years old. Participants received £5 as compensation for taking part in the study. This study was conducted in the Affective Engineering Laboratory School of Mechanical Engineering, University of Leeds. Ethical approval for the study was obtained from the University of Leeds Research Ethics Committee (Reference MEEC 15-027).

The twelve statements were developed in the previous study were used in the new study. Data from participants' affective responses were collected using a bespoke, computer-based, self-report system (Figure 2). Some of the statements were modified slightly to better suit the method of paired comparisons. The system presented each participant with each pair of confectionery in all combinations, against each of the twelve statements concerning specialness, and the participant was asked to indicate which of each pair best satisfies the statement. The pair combinations, the order of statements and the order of each pair on the screen were randomized. Thus, for each participant, there were seventy-two statements in total. Participants were encouraged to look at and touch (but not eat) physical samples of the confectionery which were located next to computer terminal whilst filling in the online questionnaire (Figure 2).

Table 1 : Statements used in the experiments

Code	Statements
I0001	A box of these chocolates would be an appropriate 'thank you' gift.
I0002	A box of these chocolates would make a thoughtful gift.
I0003	This is premium chocolate
I0004	This chocolate does not need to shout about how good it is.
I0005	This chocolate would show that someone took the time to choose just the right chocolate for the occasion.
I0006	I would keep chocolates like this one for myself.
I0007	The chocolate in this wrapper is likely to exceed people's expectations.
I0008	This chocolate is like a little present for me.
I0009	With this chocolate, you feel like you are getting more than just chocolate.
I0010	This chocolate is stylish.
I0011	This chocolate would be nice at the end of a dinner party.
I0012	This chocolate would be good to enjoy with my loved-one on a quiet night

Fig. 1: Four items of confectionery

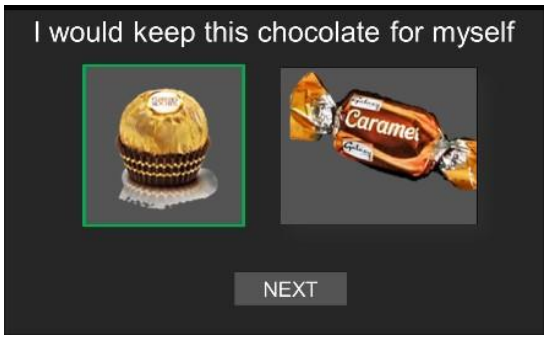

Fig. 2: Example of self-report interface

The data were analysed using the software, RUMM2030 (Andrich, Sheridan and Luo, 2012). The data from the paired comparison study were analysed to two ways. In the first way, each of the six pair combinations were treated as a separate facet, and the statements coupled with each pair were treated as the items (i.e. there were six levels in the facets and seventy-two items). The first approach does not yield linear measurement of locations of the confectionery on a scale of specialness, but was intended to reveal information about the difficulty of the task. In the second approach, the statements were combined with one of each pair to form the items, and the confectionery were separated out as a facet (i.e. there were forty-eight items and four levels in the facet). The challenge with the second approach is that the data structures required for the RUMM software needed each confectionery to be compared with itself against each statement. Initially in the analysis, this was coded as missing data, but the software was unable to process the data. The problem was solved by randomly coding each self-comparison with a one or a zero, after which the software was able to process the data. From a conceptual point of view, this is accurate because a random endorsement is equivalent to a fifty percent chance of endorsing the item, and this is equivalent to the confectionery being of equal difficulty of being endorsed when compared with itself. The second approach was intended to produce linear measurement of the affective response to the confectionery.

3. RESULTS

Figures 3 to 5 show the person-item distributions for the three analyses. There are two main regions in each of the graphs. The lower region shows the distribution of the difficulty of the items. The horizontal axis indicates increasing difficulty of endorsement in units of logit. A logit is an expression of the probability that a particular item will be endorsed. The items to the left of the graphs are easier to endorse than those on the right. The upper portion of the graphs shows the distribution of people's willingness to endorse the items. The horizontal axis indicates increasing willingness to endorse items in units of logit. Normally, one would attempt to develop items such that the spread of the difficulty of endorsement matches the distribution of the participants to endorse the items.

In the previous research, the data demonstrated a good fit to the Rasch model (Camargo and Henson, 2011). In the current research, in the analysis in which the comparison pairs were treated

as a separate facet, the data demonstrated an adequate fit to the model, but fit parameters would not be sufficient to be acceptable as a psychometric instrument. In the new analysis in which the individual pieces of confectionery were treated as a separate facet, the data demonstrated an extremely poor fit to the Rasch model.

The locations of the facets for the analysis that treated the six pair comparisons as a separate facet are shown in Table 2. The locations of the pieces of confectionery identified by the previous research and this new study are shown Table 3.

4. DISCUSSION

Figure 3 is the person-item distribution for the previous research in which participants were asked to rate the confectionary on five-point category scales against Likert statements (Camargo and Henson, 2011). The items in this case are a combination of each Likert statement with each product. It can be seen that overall, the participants found it easy to endorse the items, but that the targeting of the difficulty of the items to participants could be improved; there are many items that few people could endorse and many items that most people were able to endorse.

The person-item distribution for the analysis of the current data in which the six comparison pairs were treated as a separate facet (Figure 4) shows that the participants found it very easy to endorse the items and that the matching of the difficulty of the items with participants' willingness to endorse is quite poor. The facet locations appear to be ordered according to the ease with which the confectionery in each pair can be discriminated (Table 2). In other words, those confectionery pairs that are most similar and are difficult to discriminate have higher positive values, whereas those that are very different have more negative values. For example, the confectionery pair of Milky Way® and Ferrero Rocher®, has a large negative location relative to the other pairs, whereas Ferrero Rocher® and Lindor®, which are more similar in terms of specialness, has the most positive value of the pairs. It is speculated that it might be possible to interpret the location of the pairs as an indication of how easy it is for the participants to discriminate between those pairs, in which case the overall person-item distribution would be an indication of the ease of overall discrimination during paired comparisons. If this is the case, then the person-item distributions demonstrate that participants find it much easier to carry out paired comparisons than rating confectioneries individually against Likert statements.

The person-item distribution for the analysis of current data in which individual pieces of confectionery were treated as a separate facet (Figure 5) shows that approximately half of the items were very easy to endorse and the other half were very difficult to endorse. However, because of the way the data are coded for analysis, each comparison is represented by two data points, and consequently, each easy item has a mirror-image difficult item. In theory, therefore, the item distribution in Figure 5 should be symmetrical. The coding of the same-confectionery pairs by random ones and zeros might account for the small amount of asymmetry in the distribution. The very narrow spread of the persons' willingness to endorse shows that there was

not much variation in people's affective assessments of the confectionery. Together with the person-item distribution in Figure 4, it is interpreted as demonstrating that people find the paired comparison task very easy. The wide spread of the distribution of the items compared with the distribution of the participants indicate that these paired comparisons are perhaps too easy and that people were too consistent in making the comparisons.

The narrow spread of people's affective evaluations is compared with the wide spread of specialness of the stimuli results in a poor fit of the data to the Rasch model. Nevertheless, the locations of the confectionery on a scale of specialness derived from the paired comparisons (Table 3) is consistent with those derived from the experiment using the Likert scale. The confectioneries have identical ranks and similar relative locations within measurement error.

Measurement from paired comparisons suffer from some conceptual issues (Linacre, 1997; Wauthier and Jordan, 2013), and future work will have to address the issue of the rapid increase in the number of comparisons required when there is an increasing number of products.

Table 2: Locations of facet levels from analysis in which the confectionery pairs were treated as a separate facet.

Confectionery pair	Location (logit)	Standard error
Milky Way® - Caramel®	0.835	0.19
Milky Way® - Lindor®	-0.151	0.25
Milky Way® - Ferrero®	-0.961	0.38
Caramel® - Lindor®	0.346	0.22
Caramel® - Ferrero®	-0.959	0.39
Lindor® - Ferrero®	0.89	0.19

Table 3: Comparison of locations of confectionery between use of Likert scale (from (Camargo and Henson, 2011)) and paired comparisons.

Confectionery	Likert Scale		Paired comparisons	
	Location (logit)	Standard error	Location (logit)	Standard error
Ferrero Rocher®	1.080	0.10	1.407	0.260
Lindor®	0.780	0.10	0.552	0.190
Caramel®	-0.500	0.10	-0.720	0.230
Milky Way®	-1.370	0.10	-1.239	0.230

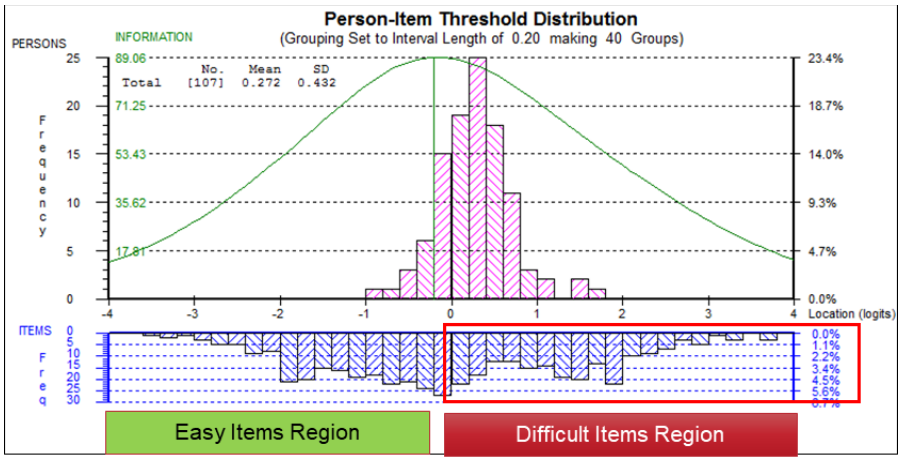

Fig. 3 Person-item distribution for data from experiment that used Likert scales (Camargo and Henson, 2011).

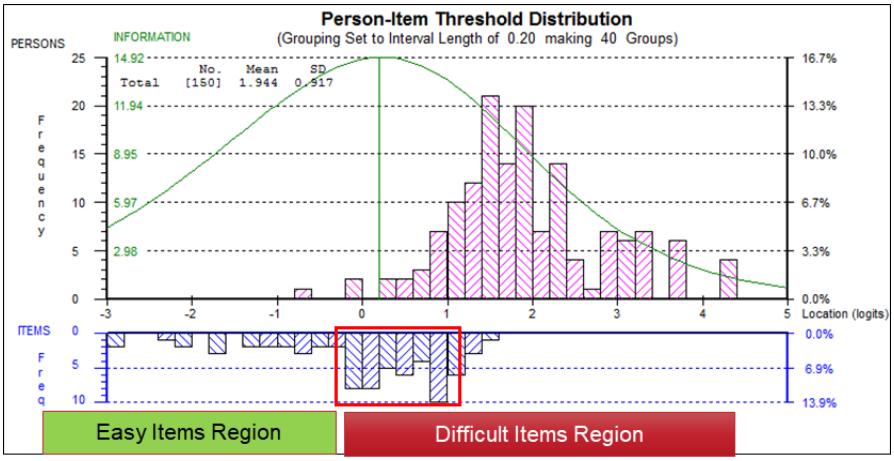

Fig. 4 Person-item distribution for analysis of current data in which confectionery pairs were treated as a separate facet.

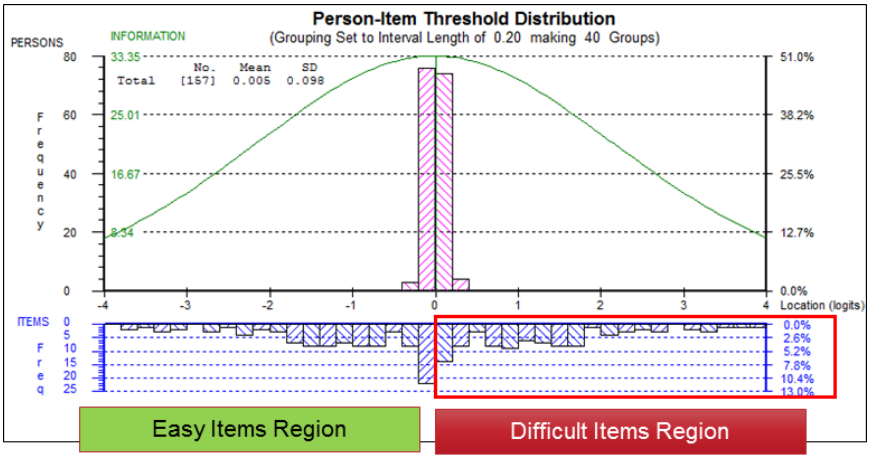

Fig. 5 Person-item distribution for analysis of current data in which individual pieces of confectionery were treated as a separate facet.

5. CONCLUSION

This work demonstrates the viability of using Rasch analysis to derive linear measurements of affective response from paired comparisons of products, although the challenges remain. If the products are too different and participants find it too easy to discriminate the products along the affective dimension of interest, then it is likely that the data will be a poor fit to the Rasch model. Participants in the research found it much easier to make paired comparisons than to evaluate the products separately against Likert statements. Although the resulting measures of specialness of the confectionery from paired comparisons were similar to those derived by the use of Likert statements within measurement error, the data were a poor fit to the Rasch model. Future research should apply the use of paired comparisons in a less-contrived context like the quality of materials for vehicle interiors.

ACKNOWLEDGMENTS

This work was partially funded by The Ministry of Rural and Regional Development of Malaysia (MARA).

REFERENCES

Andrich, D., Sheridan, B. E., & Luo, G. (2012). *RUMM2030: Rasch unidimensional models for measurement*. RUMM laboratory, Perth, Australia.

Borsboom, D., (2006). When does measurement invariance matter. *Medical Care, 44*(11), 176–181.

Bradley, R. A. & Terry, M. E. (1952). Rank analysis of incomplete block designs: The methods of paired comparisons. *Biometrika, 39*(3/4): p. 324-345.

Camargo, F. R., & Henson, B. (2011). Measuring affective responses for human-oriented product design using the Rasch model. *Journal of Design Research, 9*(4): 360-375.

Heise, D. R. (1969). Some methodological issues in semantic differential research. *Psychological Bulletin, 72*(6) ,406–422.

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

Linacre, J.M. (1997). Paired comparisons with standard Rasch software. *Rasch Measurement Transactions, 11*(3), 584-585.

Mordkoff, A.M. (1963). An empirical test of the functional antonymy of semantic differential scales. *Journal of Verbal Learning and Verbal Behavior, 2*(5 – 6), 504–508.

Nagamachi, M. (1995). Kansei engineering: a new ergonomic consumer-oriented technology for product development. *International Journal of Industrial Ergonomics, 15*, pp.3–11.

Osgood, C.E. (1971). Exploration in semantic space: A personal diary. *Journal of Social Issues, 27*, 5-63.

Osgood, C., Suci, G., & Tannenbaum, P. (1957). *The Measurement of Meaning*. University of Illinois Press, Urbana.

Rasch, G. (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.

Stevens, S.S. (1946). On the theory of scales of measurement. *Science, 103*, 677-680.

Thurstone, L.L. (1927). Psychophysics Analysis. *The American Journal of Psychology, 100*(3/4), 587-609.

Train, K. E. (2009). *Discrete choice methods with simulation. 2nd edition*. Cambridge University Press.

Wauthier, F. L., & Jordan, M. L. (2013). Efficient ranking from pairwise comparisons. *Proceedings of the 30th International Conference on Machine Learning (ICML)*, Atlanta, Georgia USA, June 2013.

Wright, B.D., & Linacre, J.M. (1989). Observations are always ordinal; measurements, however, must be interval. *Archives of Physical Medicine and Rehabilitation, 70*(12), 857–860.