Proceedings Paper:
Camargo, FR and Henson, B (2015) Conceptualising computerized adaptive testing for measurement of latent variables associated with physical objects. In: Journal of Physics: Conference Series. 2014 Joint IMEKO TC1-TC7-TC13 Symposium: Measurement Science Behind Safety and Security, 03-05 Sep 2014, Madeira, Portugal. Institute of Physics Publishing .

https://doi.org/10.1088/1742-6596/588/1/012012

Conceptualising computerized adaptive testing for measurement of latent variables associated with physical objects

F R Camargo[1] **and B Henson**[2]
[1,2]School of Mechanical Engineering, University of Leeds
Woodhouse Lane, Leeds, UK, LS2 9JT

[1]E-mail: mnfrc@leeds.ac.uk

Abstract. The notion of that more or less of a physical feature affects in different degrees the users' impression with regard to an underlying attribute of a product has frequently been applied in affective engineering. However, those attributes exist only as a premise that cannot directly be measured and, therefore, inferences based on their assessment are error-prone. To establish and improve measurement of latent attributes it is presented in this paper the concept of a stochastic framework using the Rasch model for a wide range of independent variables referred to as an item bank. Based on an item bank, computerized adaptive testing (CAT) can be developed. A CAT system can converge into a sequence of items bracketing to convey information at a user's particular endorsement level. It is through item banking and CAT that the financial benefits of using the Rasch model in affective engineering can be realised.

1. Introduction

The value of affective responses with regard to underlying attributes of products can be defined for practical purposes as the users' evaluation of physical characteristics that expresses some degree of positive or negative response with certain consistency. In affective engineering (AE) the information obtained from users' responses is converted by mathematical models into an improved design of products. However, measuring those impressions is not straightforward as it is typically to measure the properties of the physical elements that aggregate the product. One reason is that the underlying attribute solely exist as an element of a hypothesis or a premise, such as pleasantness or comfort, which are referred to as latent variables. Those attributes cannot directly be measured and inferences based on their assessment can be error-prone.

An approach in AE for eliciting affective responses is to establish a pool of variables, collect data from people's responses to those variables and apply statistical methods of analysis [1]. Typically, analysts have employed self-report questionnaires where persons give a rating of agreement against a set of words or statements, which will be termed ***items*** hereafter. Statistical treatments are given to the scores, resulting in a numerical representation of the observations subject to the analyst's interpretation. In many cases, the items emerge empirically from a multivariate analysis rather than being prescribed. This implies in difficulties to identify items that provide useful information in terms of measurement. Furthermore, scores from two different pools of items for measuring the same attribute cannot usually be compared directly. One of the reasons is that the treatments of scores can be sample dependent, limiting a quantitative interpretation [2].

One solution to overcome those difficulties is to construct through a stochastic framework a wide range of well designed items referred to as an item bank, which covers a variety of situations. Item

bank is, therefore, a repository of variables that can be used to measure a number of contextual applications. The approach follows theoretical propositions and successful applications in the fields of education and health sciences [3][4].

As a result of the item bank approach, the development of computerized adaptive testing (CAT) is possible. The concept of the CAT is concerned with establishing a sequence of items that seem most appropriate for a particular person. Although it is possible to draw upon the techniques for the development of CAT from educational testing and clinical assessment, applications in AE require particular characteristics in a CAT system such as the inclusion of stimulus objects.

The aim of this paper is to present a concept of a CAT system for the measurement of latent variables associated with physical objects. An anticipated benefit of this approach is that analysts can develop an ad hoc 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.

Item calibration is a set of procedures attached to the development of an item bank and CAT. That is, the procedures test whether independent items work well all together for measuring the latent attribute as a unidimensional structure. The concept of item bank and CAT proposed in this paper is underpinned by the Rasch measurement model (RM). The RM refers to a family of probabilistic models that provide mechanisms to test the hypothesis that the observations meet the necessary assumptions for a structure with quantitative properties. The Rasch model, named after the Danish mathematician Georg Rasch who developed it in the 1950s [5], holds the property of separation of the parameters for persons and items, which allows the design of a range of variables used as a yardstick in scales of measurement.

2. Rasch-calibrated metrics for underlying attributes of physical objects

Measurement is fostered in this paper as a way of making meaningful inferences on latent 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 being measured [6].

The RM's procedures test the observations against necessary measurement principles for quantifying the numerical validity of the data [7]. Such procedures are denoted calibration, a term coined by Wright and Panchapakesan [8], 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. The resultant probabilities allow the analysis of the expected values and what is actually observed because some results will certainly not follow the expected pattern. The relationship between person locations (β) on the continuum and the probability of a positive response is represented in the RM as item characteristic curve (Figure 1), where δ is the item parameter.

A number of RMs have been used for different applications. Camargo and Henson [9] have adapted for applications in AE the many-facet Rasch model (MFRM), a derivation of the RM developed by Linacre [10]. Thus, assuming that the data fit the model, it is possible to transform a categorical scale into an interval level with a constant unit of measurement.

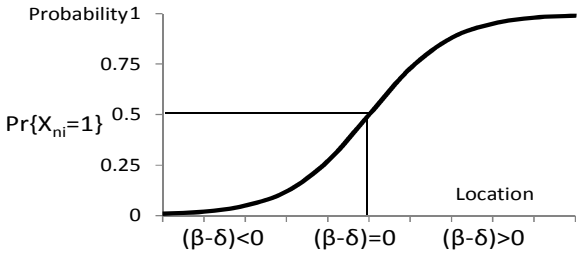

Figure 1– Probability of a positive response associated with persons' locations on the continuum.

3. How a Rasch-calibrated metric works in AE

A metric attains interpretation when the difference between persons as well as between items and between stimulus objects is established by the distance between different locations on the linear continuum [9]. A generic representation of a metric for latent variables associated with physical objects is shown in Figure 2.

Person, item and stimulus locations are based on estimation of parameters. Most of the estimation procedures are based on the method of maximum likelihood [11], calculating the standard error for each estimate through a second derivative of a likelihood function [12]. Parameters are preliminarily obtained and 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. Subsequently, Rasch analysis is carried out to evaluate the extent to which the data fit the model [7][13].

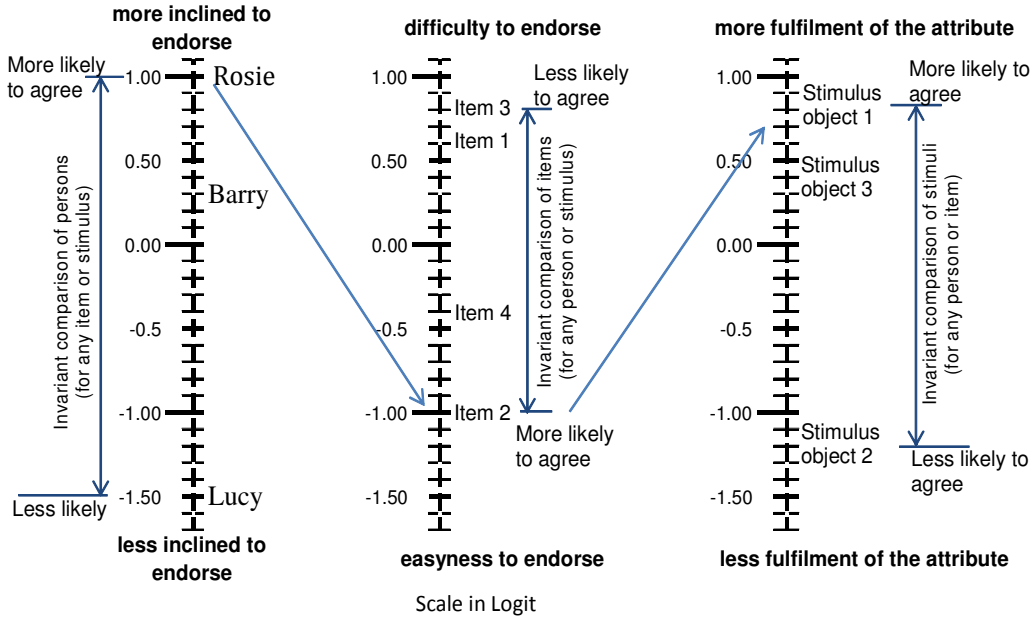

Figure 2 – Generic representation of a metric operation.

The interpretation of metric is based on the model's property of invariant comparisons. In Figure 2, for example, it is possible to interpret that Rosie has a higher probability of endorsement than Lucy independently on any item and stimulus object. Those differences are interpreted through the adapted MFRM, such that the probability of a response X is obtained by a function f of the parameters given by $\Pr(X) = f(\beta - \delta + \zeta)$, where β represents the person parameter, δ is the item parameter and ζ represents the stimulus object [9]. Using a symmetric argument, the comparison of difficulty of endorsement between any pair of items and the fulfilment of latent attribute between any pair of stimulus objects can also be made.

It is noteworthy that the unit in Rasch modelling is the log-odds unit or logit. The unit logit represents the distance on the continuum that indicates changes of the odds of observing the relevant event computed through natural logarithm [14]. The unit in logit denotes the same interval with regard to changes on the continuum.

4. A concept of computerized adaptive testing in AE

4.1. Construction of item banks for underlying attributes of products

The development of an item bank [15] entails a repository of variables represented by words or statements that hold different levels of readiness of endorsement when people are using a scale. Each

item is chosen as an independent element of a strand, which is defined to represent the relevant underlying attribute of a product. After calibration of the scale, responses for each item will indicate the degree of endorsement to a physical characteristic of the product at a specific point on the linear continuum.

The initial step toward an item bank is to obtain the strand of words or statements that can provide sufficient information in relation to a particular purpose. Those words or statements, referred to as items, are obtained from observations of the users' interaction with a product, online reviews, interviews, search in relevant literature and expert opinion [1]. Those targeted items will be established as a yardstick for measurement and therefore, they should carefully be designed through an essential specification within the context of the scale, ensuring consistency and replication validated by the RM.

4.2. Representation of the CAT system embodying physical characteristics

The calibrated items of an original structure can be used as the core of a preliminary item bank. The original, calibrated items determine anchoring points that yield the same predictions. Parameters may be anchored for either persons or items. Anchoring the person's parameter emphasises score distribution. In contrast, anchoring the item's parameter gives emphasis to the meaningfulness because items are more directly interpreted [16].

In the CAT system conceptualised in Figure 3, items are selected through a computer program 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 at the respondent's effective endorsement level. Consequently, each respondent responds to neither all items of the item bank nor all stimulus objects, but only a subset bracketing the threshold of endorsement.

The system can allow different stopping rules. In Figure 3 three rules have been used. The test can, for example, be ceased if a determined number of items are responded. A second rule can determine that a certain standardised measurement error be met. A third rule can establish a limited number of stimulus objects being assessed.

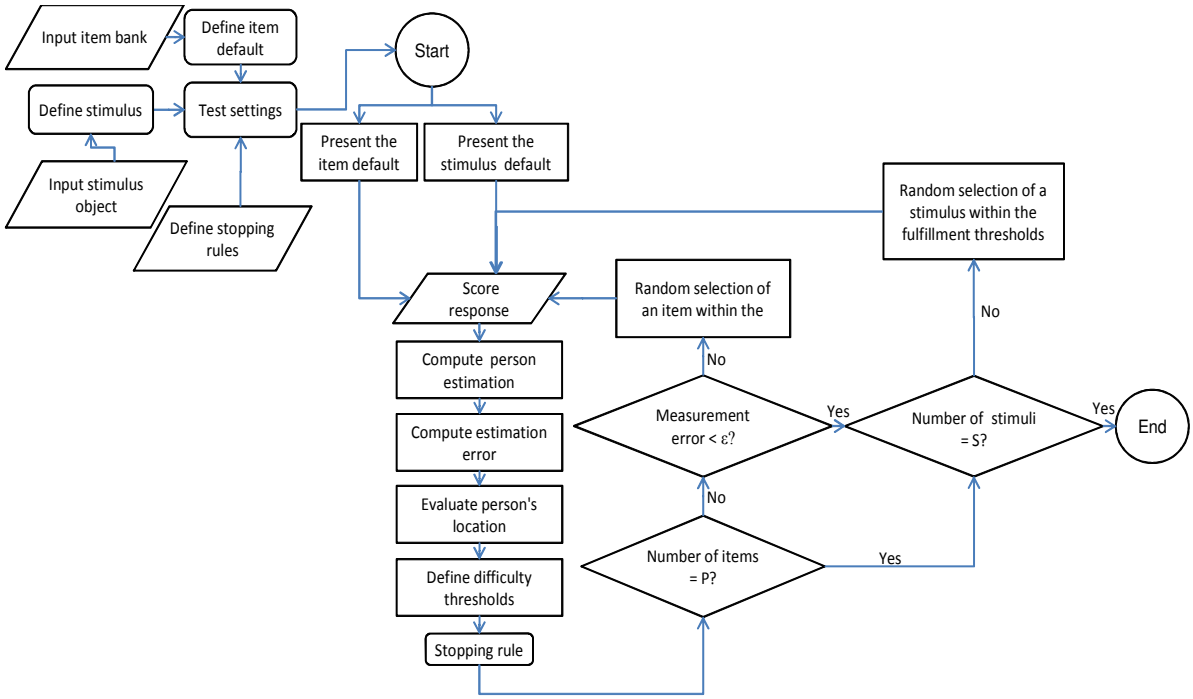

Figure 3 – Flowchart of a conceptual CAT system for latent variables associated with physical objects.

4.3. Operation and quality of measurement

Preliminarily, all items in an item bank are calibrated, i.e., they are tested for meeting the requirements of the RM. The locations of items on the linear continuum are anchored and the metric can then be used with a far smaller sample of individuals.

Figure 4 represents a generic example of the system operation based on the calibrated scale shown in Figure 2. The computer program presents the first item and the first stimulus by default. Items are subsequently selected according to two objectives. One objective is to minimise the test length. Another objective is to maximise the information function at all person locations. Information function is a technical term firstly defined by R.A. Fisher as the maximum precision which a person parameter can be estimated by the model. The precision is statistically obtained by the variability of the estimates around the value of the parameter [17], in which the maximum value at person level is found when $\beta - \delta = 0$ (see Figure 1).

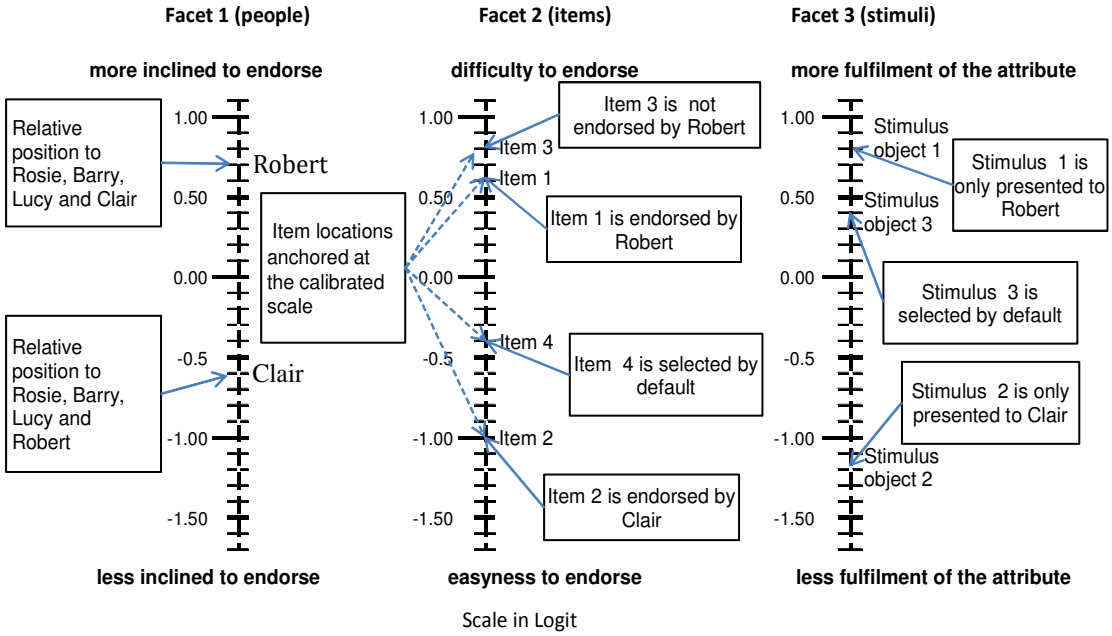

Figure 4 – Example of the CAT system operation in AE.

Empirical evidence to support the quality of measurement using the CAT system conceptualised in this paper can be given by tests of repeatability and reproducibility of measurement [18].The first replicates measurements on the same stimulus objects over a short period of time under the same conditions used to calibrate the scale. The latter involves setting conditions using the same or similar stimulus objects, different locations and operator to replicate the test, including a reasonably targeted sample of respondents with dissimilar characteristics.

5. Discussion and Implications

5.1. Potential difficulties

The concept of CAT presented in this paper for the measurement of latent variables associated with products has included in the system the assessment of physical features, distinguishing it from other domains. However, the system itself cannot ensure that the relevant measure of an underlying attribute of the product is accurate. Accuracy requires the comparison against external measurement systems or agreed standards, which hold robust evidence of validity. In the domain, such systems have been absent and precise definitions of affective attributes have hitherto been limited or inexistent. As a consequence, external systems of affective responses, to date, cannot be used as a criterion of acceptance entirely adequate to assess the quality of the measures.

Another potential difficulty of implementation of the CAT system in AE is to resolve outstanding issues with regard to dependence of words or statements when constructing an item bank. Response dependence can be found in redundant items and when 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.

5.2. Anticipated benefits

A practical value from a Rasch-calibrated metric is the possibility of construction of an item bank with elements that vary in a range of difficulty for different endorsement levels. The immediate benefit is that an analyst can tailor another structure without losing the properties of the core of the original calibrated structure to make comparisons between persons and between any pair of physical characteristics of a product. Another benefit is to make comparisons between two different studies using parallel scales.

Another benefit of item banking is to know the characteristics of the information for each item, allowing thus the specification of a subset of items with better discrimination. This feature supports the incorporation of further items to elicit affective user experiences as well as additional stimulus objects with different or improved physical characteristics.

It is through item banking and CAT that the financial benefits of using the RM in AE can be realised. The CAT system using Rasch-calibrated measurement structures is potentially useful to reduce cost in consumer research. Fewer items are necessary for estimation along with higher measurement precision. The system, after calibration, will allow the use of smaller samples and offer the advantages of convenience to respondents concerning flexible scheduling, improved security and data collection. Nevertheless, research on CAT in AE is still limited and, therefore, its application will require further investigation.

6. References

[1] Barnes C and Lillford S 2009 *J. of Eng. Design* **20** 477 – 492.
[2] Camargo F R and Henson B 2012 *Int. J. of Hum. Factors and Ergon*. **1** 204 – 219.
[3] Choppin B H 1978 *Item banking and the monitoring of achievement* (Slough: National Foundation for Educational Research).
[4] Eckes T 2011 *Psychol. Test and Assessm. Model*. **53** 414 – 439.
[5] Rasch G 1960, 1980 *Probabilistic models for some intelligence and attainment tests* (Copenhagen: Danish Institute for Educational Research), expanded edition (1980) (Chicago: The University of Chicago Press).
[6] Krantz D H, Luce R D, Suppes P and Tversky A 1971 *Foundations of measurement* **1** (New York: Academic Press).
[7] Andrich D 1988 *Rasch models for measurement* **68** (London: Sage Publications)
[8] Wright B and Panchapakesan N 1969 *Educ. and Psychol. Meas*. **29** 23 – 48.
[9] Camargo F R and Henson B 2013 *J. Phys.: Conf. Ser*. **4** 59.
[10] Linacre J M 1989 *Many-facet Rasch measurement* (Chicago: MESA Press)
[11] Fisher R A 1922 *Philos. T. R. Soc.* A **222** 309 – 368.
[12] Linacre J M 1999 *J. Outcome Meas.* **3** 382 – 485.
[13] Tennant A and Conaghan P G 2007 *Arthritis Rheum.* **57** 1358 – 62.
[14] Wright B D 1993 *Rasch Measurement Transactions* **7** 288.
[15] Hahn E A, Cella D, Bode R K, Gershon R and Lay J 2006 *Med. Care* **44** S189 – S197.
[16] Embretson S E and Reise S P 2000 *Item response theory for psychologists* (Mahwah: Lawrence Erlbaum).
[17] Timminga E and Adema J J 1995 *Rasch models: foundations, recent developments and applications* ed G H Fisher and I W Molenaar (New York: Spring-Verlag) 111 – 127
[18] VIM - International Vocabulary of Metrology 2012 *Basic and general concepts and associated terms*, 3rd ed (The Joint Committee for Guides in Metrology – JCGM)