African Journal of Research in Mathematics, Science and Technology Education, 2013
Vol. 17, Nos. 1–2, 126–138, http://dx.doi.org/10.1080/10288457.2013.828407

The Usefulness of the Rasch Model for the Refinement of Likert Scale Questionnaires

Liezel Retief[1*], Marietjie Potgieter[2] and Marietjie Lutz[1]

[1]*Department of Chemistry and Polymer Science, Stellenbosch University, South Africa*
[2]*Department of Chemistry, University of Pretoria, South Africa*
**Corresponding author, email: liz@sun.ac.za*

In this paper the use of the Rasch model is explored as a transparent, systematic and theoretically underpinned response to quality issues that are widely recognised as problematic in the refinement of Likert scale questionnaires. Key issues are the choice of length of scale, the pursuit of a favourable estimate of Cronbach's alpha at the possible expense of construct validity, and the fact that total raw scores arise from ordinal data but are used and interpreted as if measurement had occurred. We use a questionnaire under development for the measurement of perceptions of first-year chemistry students on demonstrator effectiveness to illustrate the process of Rasch analysis and instrument refinement. This process involves investigation of fit of the data to the model, possible violations of the assumption of local independence, and several aspects of item functioning. We identified disordered response categories as the probable reason for misfit in this data set and propose strategies for modification of items so that they can be retained rather than rejected.

Keywords: Rasch model; questionnaire refinement; Likert scale; construct validity; internal consistency

Introduction

Questionnaires are the most widely used type of instrument for collecting data from respondents due to the fact that they are relatively economical, present the same questions to all the respondents and can ensure anonymity (McMillan & Schumacher, 2010). Likert scales were developed by the psychologist Rensis Likert (Likert, 1932) and are commonly used in questionnaire surveys to collect ordinal data on strengths of attitudes, attributes, perceptions and beliefs. The design of an instrument whether for an attitude survey or for an ability test is never an easy task. Not only does it require extensive contextual work in developing the items that will measure the desired construct, but ethical considerations demand that the instrument should also show both good internal consistency and validity, thereby ensuring that inferences can be made that are valid and fit for purpose.

The concepts of validity and reliability are often presented as separate constructs in various text-books and articles. We believe, however, that these concepts do not stand alone but are integrated, especially during the stages of instrument design and refinement. Traditionally, validity issues are addressed during instrument design whereas reliability issues are dealt with once the instrument has been piloted. Draft instruments are often subjected to a panel of experts for judgment on different aspects of validity, for example construct validity (content representativeness), face validity (clarity of language and presentation), and suitability for the purpose intended. However, all of these judgments are inevitably subjective by their very nature and are seldom tested in an objective manner.

There are a number of issues associated with the design, refinement and scoring of Likert scale questionnaires that should be carefully considered because they can potentially influence the test instrument's validity and reliability. The first issue deals with the construction of the item scale: How many response categories should be included, and should the scale include a neutral option such as a midpoint on the scale? McKelvie (1978) reported from a review study that in respect of reliability, item scales with less than five categories have lower reliability coefficients than item scales with five or more categories, while item scales with more than 11 categories do not show an improvement. He concluded that there is no single optimal number of categories, and that the length of the item scale is influenced by the trait being measured, the respondents' familiarity with the trait and their educational and socioeconomic background. Busch (1993) notes that a scale with an odd number of categories usually includes a neutral option in order to provide a balanced choice of options, whereas an item scale with an even number of categories requires the respondent to make a decision in a specific direction. Inclusion of a middle option termed 'Neutral', 'Not Sure' or 'Undecided' has been shown to be problematic (Andrich, De Jong, & Sheridan, 1997). They recommend that pilot testing of questionnaires should be done to determine the optimal scale length of individual items and item reliability. Standard errors for estimates and standard deviations of item scores should also always be checked during the pilot phase (Busch, 1993).

The second issue deals with the accepted practice of rejecting items with poor item characteristics during instrument refinement. Survey and test construction usually involve the crafting of a large collection of items of varying intensities or difficulties, which are then piloted. Only those items with good characteristics are retained, i.e. items that show good discrimination between respondents, correlate strongly with the total score and contribute positively to internal consistency as reflected by the Cronbach's alpha statistic (Cronbach, 1946). It is not uncommon in high stakes large-scale testing that as few as 10% of the original collection of test items are retained in the final instrument. Two complications may arise from this practice of item selection. Highly discriminating items may provide an unfair advantage to students of greater ability (Masters, 1988) and the quest for high instrument reliability may sacrifice construct validity because items addressing essential construct dimensions are rejected without being replaced (Tennant, McKenna, & Hagell, 2004).

The third issue deals with the scoring and processing of data. For example, data collected with a questionnaire comprising five item response categories such as 'Never', 'Rarely', 'Sometimes', 'Often' and 'Always' are traditionally scored on a rating scale with arbitrary number sequences such as 1 to 5. This number labelling creates the illusion that distances between the categories are equal and also an exact known number. The concern here is that any such distance is a psychological distance (Busch, 1993) and upon careful consideration one would not expect the respondent to necessarily experience these distances to be the same. It is quite possible, for instance, that the distance between 'Rarely' and 'Sometimes' is smaller than the distance between 'Never' and 'Rarely'. This incongruity would happen when respondents are very reluctant to endorse 'Never', but much more willing to endorse either 'Rarely' or 'Sometimes'. After assigning a number to each response to reflect the ordinal position of the option selected, the raw scores are summed to generate a single total score for each respondent. The person-total is used to rank respondents in terms of the aggregated strength of their endorsement of the statements comprising the questionnaire. In traditional test theory (TTT) the respondents are ordered in terms of the strength of the variable being studied based on this total raw score. This strategy does not constitute real measurement, because the total raw scores alone cannot generally yield meaningful information about the distance between respondents on the scale that TTT uses. Total raw scores merely represent ordinal data which can only be legitimately used for the ranking of persons (Busch, 1993). The use of the total score also does not give any information about the relative strength of the statements, also referred to as the 'endorsability' of items, because it does not indicate which items are easier to support, confirm or endorse than others. The numbers used to score Likert scale item responses are usually processed in calculations such as addition and division and the results of such operations are interpreted as if they have an inherent mathematical meaning, which is clearly unjustified without a coherent and robust set of corroborating arguments.

These concerns about common practice in TTT can be addressed by using the Rasch model for instrument refinement and data analysis. In the Rasch model, raw scores for respondents and for items are transformed to log-linear interval measures which describe the differences between persons in measurement terms. Thus ordinal data are elevated to interval measures provided that the data fit the Rasch model (Boone & Rogan, 2005). An equivalent reliability index to that of Cronbach's alpha is calculated in the Rasch model (Andrich, 1982), which is called the Person Separation Index (PSI), so-named to indicate that it is a property not of the test but of the persons responding to the test instrument. The PSI can be estimated even in the presence of missing data (Boone & Rogan, 2005), not as a compromise, but as an inherent property of the Rasch model. It is calculated in an analogous way to Cronbach's alpha and has a theoretical maximum value of 1.0 indicating perfect internal consistency. During instrument refinement every attempt is made to modify and improve items rather than to reject them, based on the assumption that sound theoretical or practical reasons were the original rationale for their inclusion. In addition, the model generates empirical evidence for the functioning of response categories to inform instrument refinement, as will be demonstrated in this paper.

It is not our aim to contribute to the already existing vast methodological knowledge of the Rasch model nor is it to advocate the Rasch model as a novel concept for the development and refinement of test instruments. Our aim is rather to illustrate the use of the Rasch model as a transparent, systematic and theoretically underpinned response to the above mentioned quality issues, that are widely recognised as problematic, in the refinement of Likert scale questionnaires. This will be specifically useful for the researcher who is dealing with statistical analysis of questionnaire data and is dissatisfied with the limitations of constructing summary scales from raw score data.

We pose the following research questions and set out to answer them by means of a Likert scale questionnaire that we have developed:

(1) Is the reliability index a trustworthy statistic for internal consistency or is it artificially inflated?
(2) Are all response categories likely to draw responses from respondents, i.e. are they all useful?
(3) Do the psychological distances between response categories differ for distinct items?

In the rest of this paper the Rasch model will firstly be introduced, whereafter a specific example of data analysis from our work will be presented to demonstrate how empirical evidence is used to confirm that the fundamental assumptions of the Rasch model are met and to answer the above research questions.

The Rasch model

The Rasch model is a powerful tool for the analysis and refinement of survey and test instruments especially with regards to increasing reliability and validity (Boone & Rogan, 2005).The concept of Rasch analysis is not new to this journal. Two articles pertaining to Rasch analysis have already been published. The first was by Boone and Rogan (2005) who gave an outline and explanation of the Rasch model itself and motivated its use to achieve greater rigour in quantitative analysis. The second was by Potgieter, Davidowitz, and Venter (2008) who designed a performance instrument as a diagnostic and placement tool of first-year chemistry students and used the Rasch model to demonstrate its quality in terms of alignment and internal consistency. A number of mathematics education researchers have applied the Rasch model in their work in South African studies, e.g. Huntley, Engelbrecht, and Harding (2009), who chose the Rasch model for statistical analysis because it does not depend on the assumption of a normal distribution of scores, Julie, Holtman, and Mbekwa (2011), who used the Rasch model to verify the viability of a newly developed instrument for measuring teacher preferences, and Long (2009), who used the Rasch model to characterise learners individually in terms of their mastery of the multiplicative conceptual field. However, examples of the application of the Rasch model in science and technology education research in South Africa are very limited.

Numerous international studies have been reported demonstrating the application of the Rasch model in science and mathematics education research. For example, as editors of a special journal issue on the topic, Callingha and Bond (2006) have convincingly argued for the merit of the Rasch model in mathematics education research, Boone and Scantlebury (2006) have shown its application

at both the individual and systemic levels in achievement testing in science and Neumann, Neumann, and Nehm (2011) did a Rasch analysis to evaluate and improve a Likert scale instrument on nature of science. The Rasch model has also been implemented to evaluate the robustness of instruments in a wide variety of testing applications including TIMSS (Trends in International Mathematics and Science Study), PIRLS (Progress in International Reading Literacy Study) and PISA (Program for International Student Assessment).

The Rasch model, also referred to as the Rasch Latent Trait Theory, is part of modern test theory developed by George Rasch (Rasch, 1960). The term latent refers to an underlying or unobservable or hidden trait that is to be measured, such as the spelling ability of a respondent on a spelling test. The Rasch model measures the strength of the latent skill by harnessing the power of the measured items, to form a graduated scale with known intervals, as stated by Van de Grift and Van der Wal (2010).

The Rasch model is built on a number of basic assumptions such as local independence, unidimensionality, sufficiency of raw scores, invariance and explicit and implicit use of parallel item characteristic curves. Boone and Rogan (2005) discussed some of the basic assumptions and hence this paper will extend the discussion by exploring the assumption of local independence in more depth because of its relevance to this study. The Rasch model is built on the hypothesis that the items on a survey or test instrument must indirectly measure one and the same unidimensional variable or latent trait throughout the whole range of the instrument. The assumption of local independence implies that every item is expected to contribute unique information regarding the latent trait which is not captured by any other item. The information that is obtained from each item is relevant to the common trait and therefore related to, but statistically independent of the other information. As a result, in those circumstances, the responses to a number of different items can be summed to give a more valid and reliable measurement of the strength of the latent trait than the response on only one item. The assumption of local independence can be violated when one or more items are included in the instrument that measure variation on another latent variable in addition to the one that is being measured. This relationship is called *trait dependence*, more commonly known as multidimensionality. A second type of violation of local independence occurs when the response on one item depends on the response on another. In such a case a questionnaire respondent who endorses the first item is more likely to also endorse the second one than would be the case when the two items were completely independent. This violation is called *response dependence*. When violations of local independence occur, parameter estimation is affected in Rasch analysis and the apparent reliability of the test instrument can be artificially inflated (Tennant & Conaghan, 2007). If these assumptions are violated, the data will not fit the model and reliable inferences cannot be made. Measurement results may even be completely misleading. Basic assumptions such as unidimensionality and response independence are not assumed to be met, but can and must be confirmed by statistical analysis.

Methods

We have recently developed a questionnaire to probe the perceptions and attitudes of first-year Chemistry students regarding Chemistry practical sessions. This study provides a vehicle to illustrate the Rasch analysis process for instrument validation and refinement and to address the research questions presented above.

Laboratory practicals play an important role in the teaching and learning of Chemistry. Literature suggests that the students' perceptions of the usefulness and applicability of a subject, method or technique impacts their success rate in that subject (Henderleiter & Pringle, 1999; Johnstone, 2000). Well-designed laboratory work has been shown to have an impact on enhancing student attitudes towards science; stimulating interest, enjoyment and motivation of science learning (Hofstein & Lunetta, 2004). It is therefore of interest to us to investigate what the first-year students' perceptions and attitudes are towards our current practical training, what they perceive the intended outcomes of Chemistry practicals should be and whether any of these perceptions change during the year. We also want to

determine how their perceptions and expectations of the outcomes of Chemistry practicals compare with those of the demonstrators and lecturers.

Although literature includes several attitude studies using questionnaires as test instruments, Blalock et al. (2008) have reported that the majority of science attitude test instruments are plagued by deficiencies such as an absence of psychometric evidence, limited reliability and validity information and a disregard for missing data. Therefore we identified the necessity to develop our own questionnaire instrument and analyse and refine it by using a suitable Rasch model in order to address these concerns.

Instrument design

The items and response options in the questionnaire were informed by informal small group and individual interviews with students, written student feedback and email questions sent to lecturers as well as information obtained from literature. Using the students' and lecturers' own wording was intended to ensure both face and construct validity. The instrument consisted of three subscales to probe student perceptions and attitudes regarding (i) the laboratory experience, (ii) the demonstrators, and (iii) the outcomes expected from the laboratory training. In this paper only one of these subscales will be presented, i.e. the subset of questions probing student experience and perceptions of demonstrator effectiveness. In this subset we probed the students' perceptions and experience of the demonstrator system rather than of individual demonstrators. Although students were assigned a specific demonstrator, the students were free to interact with other demonstrators, technical assistants and lecturers. Polytomous data were collected by means of 8 items each with 5 response categories (strongly disagree, disagree slightly, neutral, agree slightly, strongly agree). A common item stem was used: 'What is your experience with regards to the demonstrator that helped you during Chemistry practicals?', with the following statement options: 'Helpful' (Item 17), 'Gave clear instructions' (Item 18), 'Encouraged students to ask questions' (Item 19), 'Made mistakes often' (Item 20), 'Understood the practical work and theory' (Item 21), 'Made practicals enjoyable' (Item 22), 'Available when needed' (Item 23), and 'Confused students regularly' (Item 24). Items 20 and 24 had negative statements and were reverse scored.

Sample and data collection

The questionnaire was distributed to 842 first-year Chemistry students at a prominent South African university at the start (Questionnaire 1) and the end (Questionnaire 2) of the first semester of 2012. The first questionnaire was administered at the end of the first practical session in order to give students an initial glimpse of what Chemistry practicals entail and to probe their expectations of Chemistry practicals. The second questionnaire was administered at the end of the last practical session in order to capture their actual experience. Participation was voluntary. Response rates were 80% for the first questionnaire (675) and 57% for the second questionnaire (481). The data from both questionnaires 1 and 2 were combined for the purpose of instrument refinement, which generated a data set consisting of 1,143 data records. Of those 1,143 data records 16 had no responses in the subscale we are analysing in this paper and were therefore removed. The prevalence of missing data in the entire data set was found to be only 2%.

Data analysis

Statistical analysis of the data was carried out using the Rasch model. In the next section the following procedures and steps will be discussed: the choice of a suitable Rasch model for the analysis, analysis of item and person fit statistics, local independence, differential item functioning and operation of response categories.

Choice of Rasch model

The questionnaire uses a Likert scale for responses, but the wording of the options suggest uneven distribution of intervals between response categories (strongly agree—slightly agree—neutral—

slightly disagree—disagree strongly). Such a Likert scale generates polytomous data which can be analysed using either the Masters Partial Credit Model (PCM) (Masters, 1982) or the Andrich Rating Scale Model (RSM) (Andrich, 1978). Both of these models belong to the family of Rasch models but they differ in terms of underlying assumptions. In the RSM it is assumed that the distances between successive response categories within items are unequal, but all items share the same unequal distribution of distances between response categories. In the PCM the distances between successive categories within items are not equal and the distances between response categories are unique for each item. The Partial Credit Model was selected because it was less restrictive than the Rating Scale Model and was expected to allow better fit of the data to the model, an assumption which was empirically confirmed. The PCM allows the distances between response categories to emerge from the data rather than being imposed on the data, as would be the case in TTT and to a lesser extent in the RSM as well. There are various software packages available operationalising the Rasch model including WINSTEPS, RUMM2020 and ConQuest (Tennant & Conaghan, 2007). In this study the RUMM2030, the latest version, was used.

Frequency distribution of responses

Respondents generally gave a very positive judgment of demonstrator effectiveness as can be seen in the category response frequency distributions reported in Table 1. All categories elicited responses, but only 7% of the responses populated the categories ‘Strongly disagree’ and ‘Disagree slightly’.

Analysis of fit statistics

Rasch analysis is a multistep process aimed at confirming the basic assumptions of fit of data to the preferred model. A cluster of fit statistics is examined for the instrument as a whole and for persons and items individually (Smith & Plackner, 2009). Misfits of either item or person responses are viewed as anomalies that warrant further investigation. Since the focus of this study is instrument development and refinement it will be possible to remove data records of persons whose responses are misfitting beyond an acceptable margin of deviation. However, misfitting items which elicit too many unexpected responses must be investigated both quantitatively and qualitatively for possible item improvement. All items are assumed to be designed for a unique purpose, i.e. to generate data essential for measurements and should therefore be retained if at all possible, but rendered more conformable with measurement properties.

The first analysis of all data indicated that the PSI is good (PSI = 0.75). The closer the reliability index value is to 1, the better the internal consistency of the test instrument. A reliability index should be at least 0.85 if the data will be used to make decisions about individuals, however if the data will be used to draw conclusions on a group of students (as is the case in our research) then the reliability index need only be greater than 0.65 (Frisbie, 1988). Despite the suggested good PSI, the data does not fit the model well (total item chi-square = 206.41, *df* = 72, probability = 0.0000). This means that the data fits the model is rejected. The origin of misfit will be further explored in the sections that follow.

Item fit

Rasch analysis involves a series of comparisons between *expected* and *observed* responses of individuals and groups to each of the items. The *expected* responses are obtained from mathematical equations for the calculation of the probability of observing a specific response based on three estimated variables; item difficulty, person ability and the category threshold parameter in the case of polytomous items. Should an item elicit responses that deviate consistently from what is predicted by the Rasch model then an item is flagged as misfitting. There are many reasons why items misfit, including data entry errors, instrument administration errors, ambiguous item phrasings, and the possibility that an item taps into dimensions other than only the underlying construct being measured.

The fit statistics for the items included in the subscale on demonstrator effectiveness, Items 17 to 24, are reported in Table 2. In this table the location refers to the level of endorsability of the item. During Rasch analysis, the mean of item locations is arbitrarily set at 0.00 logits, with items that are more

Table 1: Category response frequency distributions for Items 17 to 24

	Statement option	Category 1: Strongly disagree	Category 2: Disagree slightly	Category 3: Neutral	Category 4: Agree slightly	Category 5: Strongly agree
Item 17	Helpful	13	27	126	361	581
Item 18	Gave clear instructions	17	44	155	427	462
Item 19	Encouraged students to ask questions	38	76	328	346	312
Item 20*	Made mistakes often	24	75	208	307	491
Item 21	Understood the practical work and theory	14	31	139	355	566
Item 22	Made practicals enjoyable	21	58	385	363	279
Item 23	Available when needed	17	58	171	380	480
Item 24*	Confused students regularly	8	65	202	289	452

*Reverse scored—note that the category labels are reversed for these items.

easily endorsable located at negative values and those more difficult to endorse placed at increasingly positive locations. The item locations reported in Table 2 imply that Item 17 was most readily endorsed ('Helpful'; δ = -0.393) and Item 19 was the most difficult to endorse ('Encouraged students to ask questions', δ = 0.533). The fit residuals indicate how consistently the *observed* item responses approached *expected* values predicted by the model and as a rule of thumb these standardised residual values should fall within the range of –2.5 and +2.5. The chi-square values reflect the overall difference between the observed and expected values for responses to a specific item. A high chi-square value is an indication that the responses are not consistent with the model, especially when the probability is less than 0.05 as a first approximate cut-off value criterion. The probability value indicates the likeliness that such a large a chi-square value would be obtained merely due to chance. The results in Table 2 indicate that items 17, 18 and 20 show misfit as evidenced by their large fit residuals ($> \left| 2.5 \right|$), large chi-square values and the fact that these values are unlikely to have been derived by chance (prob < 0.05). These three items should be investigated qualitatively in order to determine the potential causes of the observed misfit to guide instrument refinement.

Person fit

In an approach analogous to investigation of item misfit, person data are analysed to identify data records with erratic patterns within the person responses that may complicate further analysis. Standardised fit residuals are calculated for the difference between *expected* and *observed* responses of a specific person to each of the items in the test instrument. In this case the fit residuals of 65

Table 2: Fit statistics for individual items

	Location (logits)	Fit Residual	Chi-Square	Probability*
Item 17	–0.393	–3.633	38.705	0.0000
Item 18	–0.083	–3.439	46.835	0.0000
Item 19	0.533	0.258	8.210	0.5132
Item 20	0.177	6.253	54.789	0.0000
Item 21	–0.218	–0.115	22.327	0.0079
Item 22	0.352	0.506	10.939	0.2799
Item 23	–0.059	–0.073	10.019	0.3490
Item 24	–0.309	1.468	14.591	0.1028

*Degrees of freedom for this analysis = 9.

persons (6% of 1,143 respondents) exceeded the desired boundaries of –2.5 to + 2.5 and their data records were therefore removed from the data. This elimination strategy produced a clearer picture of the operation of the instrument in order to guide its refinement.

Analysis of local independence

Two types of violations of local independence were tested for, i.e. *response dependence* and *trait dependence*, or multidimensionality, as it is more commonly known. Both types of violations will manifest themselves in correlations between the fit residuals of a subset of items thereby indicating that some of the items have something more in common than all of the items have in common with one another.

When the fundamental assumption of local independence is violated by response dependence, the data will no longer fit the Rasch model and instrument reliability. Then PSI would suggest a better internal consistency for the instrument than is justified. Response dependence is observed when a student's response on one item is affected by or dependent on a response given to another question and thus there is a non-zero correlation between the two items. Table 3 indicates that this data set shows no evidence of residual correlations above 0.4, the cut-off value of our preference.

The second possible violation is trait dependence or multidimensionality. The Rumm2030 software used checks for multidimensionality by means of principal component analysis of item fit residuals which is analogous to factor analysis in TTT. After extracting the latent trait and the item residuals associated with this dimension there should be no further pattern of correlations between the item residuals. However, if a principal component analysis indicates a meaningful pattern of correlations between fit residuals then multidimensionality is suspected. In our case no further multidimensionality within the subscale of demonstrator-effectiveness was detected.

Empirical evidence for the absence of violations of local independence indicates that this instrument is characterised by good internal validity—only one dimension is measured—as well as good reliability, i.e. that the estimated person separation index (PSI = 0.75) gives a fair reflection of internal consistency. The numerical value for PSI is interpreted similarly to the Cronbach's alpha: a value of 0.75 for an instrument with only eight items indicates good internal consistency. This finding answers research question 1, i.e. that the reliability index is a trustworthy statistic for internal consistency; it has not been artificially inflated. This favourable finding should not be taken for granted, but must be verified in every analysis.

Item functioning and operation of response categories

Further steps must be taken to determine what the origin of the misfit is between the data and the model as uncovered in the first Rasch analysis of the data in both the summary fit statistics and the individual item fit statistics (Table 2). There are two possible sources of item misfit which will be discussed next, i.e. differential item functioning and the operation of response categories.

Table 3: Residual correlations

	Item 17	Item 18	Item 19	Item 20	Item 21	Item 22	Item 23	Item 24
Item 17	1							
Item 18	0.229	1						
Item 19	–0.009	–0.011	1					
Item 20	–0.280	–0.252	–0.315	1				
Item 21	–0.136	–0.138	–0.238	–0.037	1			
Item 22	–0.092	–0.137	0.012	–0.391	–0.051	1		
Item 23	–0.115	–0.195	–0.154	–0.217	–0.129	–0.024	1	
Item 24	–0.198	–0.172	–0.309	0.134	–0.173	–0.288	–0.109	1

Differential item functioning (DIF) refers to the observation that subgroups of respondents within a sample may respond in a significantly different manner to an individual item despite equal levels of underlying characteristic being measured. DIF is an unwanted anomaly since it will indicate that one group of respondents is unfairly advantaged over another group with regards to a specific item.

The data were divided into subsets for three different analyses of DIF to check whether there are any significant differences between the response patterns of the subgroups.

The three data subsets are:

i) Questionnaire data collected at the start of the first semester (questionnaire 1) and questionnaire data collected at the end of the first semester (questionnaire 2);
ii) Respondents per laboratory session: Monday, Wednesday, or Friday;
iii) First-time students and repeaters.

DIF analysis involves a process whereby observed responses from each data subset are divided into 10 categories of strength of endorsement or ability, and the corresponding subset mean values are plotted on the item characteristic curve (ICC) for each test item. The ICC is a plot of the probability of a positive response at each location of person proficiency, with the person location on the x-axis and the expected value as a probability of endorsement on the y-axis. The person location is an indication of the tendency of the respondent to endorse statements projecting a positive perception of demonstrator effectiveness; the more positive the person location, the more enthusiastic the endorsement. If DIF is present for any test item then the observed means calculated for response categories will deviate from the ICC.

As an illustration, in Figure 1 the observed means are plotted for the data obtained for Item 21 from the first questionnaire (Quest1) and from the second questionnaire (Quest2). The deviation of the two curves from the ICC is not statistically significant as indicated by an analysis of variance. DIF analysis per practical session and per student group showed similar plots. The finding of an absence of DIF between questionnaires 1 and 2 is an important one, because it means that the items functioned in the same manner irrespective of the fact that perceptions and expectations are probed in questionnaire 1 and actual experiences in questionnaire 2. This finding corroborates our decision to combine the two data sets for the purpose of instrument refinement.

Operation of response categories

The RUMM2030 software has a function for the graphical display of item category probability curves (CPC) as derived from analysis of the actual data. For each item the probability of selection of individual categories is plotted along the continuum of person locations on the x-axis and the probability (likelihood of choosing that category) on the y-axis. The CPC gives a visual indication of whether

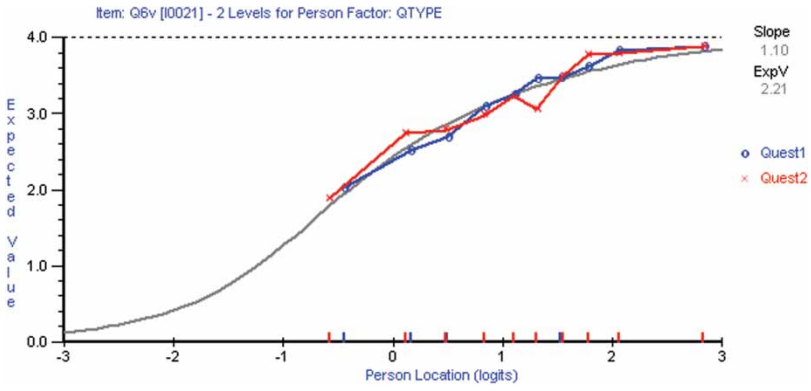

Figure 1: Item characteristic curve (ICC) to illustrate absence of DIF for Item 21 between data obtained from the first and second questionnaires

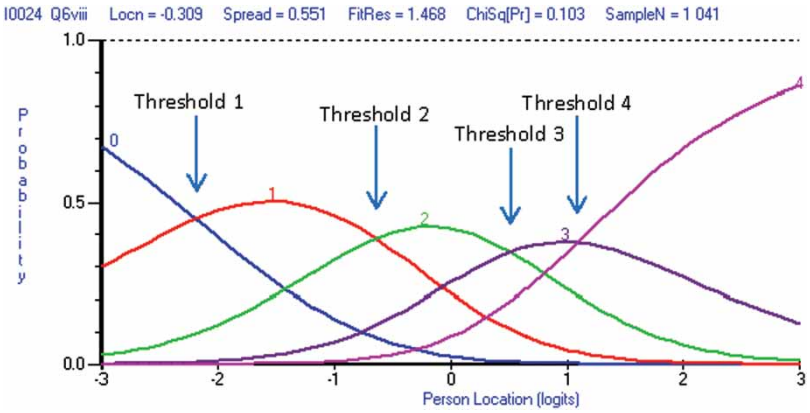

Figure 2a: Category probability curve (CPC) of Item 24

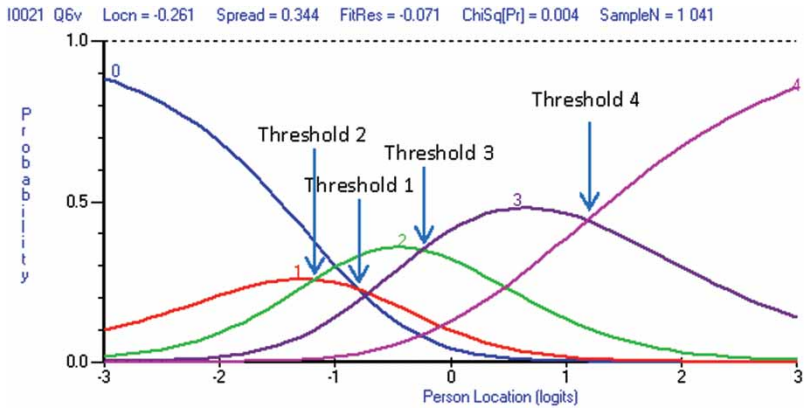

Figure 2b: Category probability curve of Item 21

the response categories on the Likert scale are functioning as intended. In these curves the categories are labelled 0 to 4 reflecting the scoring convention in RUMM2030 for a Likert scale of 5 levels. Where the curves for two categories intersect, there is an equal probability that the respondent will choose either one of the two associated options. For the sake of precision, the locations of these points of intersection for adjacent categories, called *thresholds*, are estimated and reported, not the positions of maximum probability of each category.

Figure 2a illustrates the category probability curve of Item 24 where all the categories function as intended, i.e. for each of the response categories there is a region of person locations where that category is most likely category to be chosen and the order in which this maximal chance of categories occurs matches the order of the categories in the test item. However, this matched pattern is not the case for Item 21. In the CPC of Item 21 (Figure 2b) the curve for the second category, labelled 1, lies below the other curves in such a way that there is no region where this category ever has the maximum probability of being chosen. For an item in which the categories function as expected, such as Item 24, the category thresholds occur in numerical order along the x-axis. In contrast, the order of thresholds is reversed in places when categories are not functioning as expected, as for Item 21 (Figures 2a and 2b). Disordered thresholds in an item constitute an anomaly which is likely to contribute to item misfit.

Table 4: Location of thresholds between response categories

	Item location (logits)	Threshold 1 (logits)	Threshold 2 (logits)	Threshold 3 (logits)	Threshold 4 (logits)
Item 17	–0.393	–0.767	–0.894	0.088	1.573
Item 18	–0.083	–1.034	–0.807	–0.080	1.920
Item 19	0.533	–1.042	–1.268	0.579	1.732
Item 20	0.177	–0.940	–0.564	0.377	1.128
Item 21	–0.218	–0.531	–0.920	0.014	1.437
Item 22	0.352	–1.469	–1.560	0.909	2.120
Item 23	–0.059	–1.295	–0.520	0.167	1.648
Item 24	–0.309	–0.767	–0.338	0.822	1.400

The CPCs for four of the eight items (Items 17, 19, 21 and 22) showed a similar pattern to that for Item 21 in Figure 2b, with no region on the x-axis where the second category has the maximum probability of being chosen and with the same disorder (2, 1, 3) in the sequence in which the first three thresholds occur (Table 4). Collectively, this repeated disorder is interpreted as evidence that the 'Disagree slightly' option on the Likert scale did not function as intended and may not be useful. A better item fit may result from combining the first two ordered options, 'Strongly disagree' and 'Disagree slightly' into a single first category labelled 'Disagree'. With only 7% of responses projecting a negative opinion of demonstrator effectiveness it is unlikely that these responses differentiated clearly between statements of varying intensity, 'Strongly disagree' and 'Disagree slightly'. However, judged by the CPCs for each of the items, the middle category, 'Neutral', seemed to be functioning as expected. We tested this hypothesis empirically by rescoring all of the items, firstly by combining categories 1 and 2 (scored 0 0 1 2 3) and secondly by combining categories 2 and 3 (scored 0 1 1 2 3). Since neither option resulted in a major reduction of misfit we decided to resort to qualitative means of addressing this problem. This finding answers the second research question by showing that all of the response categories attracted responses, but not all of the categories were useful.

With regard to the last research question (Do the psychological distances between response categories differ for distinct items?) careful inspection of the data reported in Table 4 provides some indication of the spacing between response categories. However, this information must be obtained indirectly from the locations of thresholds between categories. Also, there are only four items with ordered categories for which the threshold positions are meaningful for such an interpretation, i.e. Items 18, 20, 23 and 24. It is clear that the spacing between thresholds is unique to each of these four items. For example, the spacing between thresholds for Item 18 is 0.227, 0.727 and 2.000 logits, respectively, compared to 0.429, 1.160 and 0.578, respectively, for Item 24. It is important to realise that the choice of the PCM allowed these results to emerge from the data as a reflection of collective decision-making of all of the respondents. No external restrictions were placed on the data as would have been the case in the RSM or TTT. The raw data were transformed by an iterative process to generate measures of endorsability of each of the item categories.

Qualitative analysis

The items were subsequently analysed qualitatively in an attempt to understand the reasons for the anomalies associated with response category 2 and misfitting Items 17, 18 and 20. Item 20 may be improved by removing the word 'often', since intensity is already captured in the choice of categories. Furthermore, the use of the term 'slightly' in the second and fourth response categories is unconventional and may also be problematic. The reasons for misfit of Items 17 and 18 are still unclear at this point, but these items will not be rejected because they represent the voice of the students. It is likely that the issue of Rasch model fit for these two items can be addressed more effectively once the other issues have been dealt with.

Finally, the following aspects of the analysis should be emphasised: Firstly, during Rasch analysis empirical evidence is analysed to locate the origin of misfit, thereby pointing to possible areas of

improvement. Quantitative analysis must then be complemented by qualitative analysis of item wording which will inform further refinement. This combination of quantitative and qualitative analysis creates hypotheses that must be confirmed experimentally in future applications of the instrument. Secondly, the multistep process of analysis demonstrated in this article is aimed at diagnosing inadequacies of the current instrument in order to improve fit of the data from a revised instrument to the Rasch model. However, the ultimate goal of this exercise is not to achieve perfect fit, but to enable authentic measurement for which fit is a prerequisite.

Summary and conclusions

In this paper, we have presented and discussed several issues associated with the design, refinement and scoring of Likert scale questionnaires in standard practice which could threaten the validity of inferences that are made based on raw score data. Response scales may include response categories that are redundant or not functioning as expected, and the accepted practice of rejection of items with poor item characteristics may compromise construct validity while artificially enhancing the reliability index. Perhaps the most serious complication associated with standard practice in processing Likert scale data is the fact that raw score data do not represent real measurement and can only legitimately be used for the ranking of respondents. Likert scale data are often scored by assigning integer numbers to categories in a manner that suggests measurement when that is clearly not achieved. The Rasch model, based on modern test theory, has proven potential to address these concerns. Using an example of data obtained from a questionnaire under development, we have shown that Rasch analysis can be used to identify Likert scale categories that are not functioning as intended, to estimate the real psychological distances between response categories and to validate the legitimacy of the reliability index as indicator of internal consistency.

Our departure point was the assumption that survey instrument construction is informed by robust theory or by contextual evidence as in our case, and that every item is carefully crafted to capture unique information that is both relevant and required to fully describe the construct in question. This implies that a concerted effort has to be made to improve rather than reject misfitting items. We have demonstrated the process of analysing empirical evidence on item functioning and on the basic assumptions of the Rasch model in an attempt to locate the source of misfit so that improvements can be made during subsequent rounds of implementation. We have also shown that instrument refinement is a holistic approach incorporating careful consideration of both quantitative and qualitative evidence.

The Rasch analysis process aids in dealing with subtle threats to construct validity and internal consistency of the instrument by checking that the reliability statistic is not artificially inflated and by confirming the assumption of local independence empirically. This approach represents a rigorous process with the ultimate aim of making authentic measurement possible to a level of precision that approaches that which is routinely achieved in the natural sciences. It is not possible to make a precise cut with a blunt knife. Similarly, authentic measurement in the social sciences is not possible if we do not develop instruments of superior quality. The Rasch model is a powerful tool for instrument refinement with proven potential to improve the rigour of qualitative research by eliciting a valid and reliable quantitative dimension.

Acknowledgements

This project was funded by the National Research Foundation and the Centre for Teaching and Learning (Stellenbosch University).

References

Andrich, D. (1978). Rating formulation for ordered categories. *Psychometrika*, *43*, 561–573.

Andrich, D. (1982). An index of person separation in latent trait theory, the traditional KR.20 index, and the Guttman scale response pattern. *Education Research and Perspectives*, *9*, 95–104.

Andrich, D., De Jong, J., & Sheridan, B. (1997). Diagnostic opportunities with the Rasch model for ordered response categories, In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 58–68). Münster, Germany: Waxman Verlag.

Blalock, C., Lichtenstein, M., Owen, S., Pruski, L., Marshall, C., & Toepperwein, M. (2008). In pursuit of validity: A comprehensive review of science attitude instruments 1935-2005. *International Journal of Science Education*, *30*(7), 961–977.

Boone, W., & Rogan, J. (2005). Rigour in quantitative analysis: The promise of Rasch analysis techniques. *African Journal of Research in Mathematics, Science and Technology Education*, *9*(1), 25–38.

Boone, W., & Scantlebury, K. (2006). The role of Rasch analysis when conducting science education research utilizing multiple-choice tests. *Science Education*, *90*(2), 253–269.

Busch, M. (1993). Using Likerst scales in L2 research: A researcher comments. *TESO Quarterly*, *27*(4), 733–736.

Calingha, R., & Bond, T. (2006). Research in Mathematics education and Rasch measurement. *Mathematics Education Research Journal*, *18*(2), 1–10.

Cronbach, L. J. (1946). Response sets and test validity. *Educational and Psychological Measurement*, *6*(4), 475–494.

Frisbie, D. A. (1988). Reliability of scores from teacher-made tests. *Educational Measurement: Issues and Practice*, *7*(1), 25–35.

Henderleiter, J., & Pringle, D. (1999). Chemical education research – effects of context-based laboratory experiments on attitudes of analytical chemistry students. *Journal of Chemical Education*, *76*(1), 100–106.

Hofstein, A., & Lunetta, V.N. (2004). The laboratory in science education: Foundations for the twenty-first century. *Science Education*, *88*, 28–54.

Huntley, B., Engelbrecht, J., & Harding, A. (2009). Can multiple choice questions be successfully used as an assessment format in undergraduate Mathematics? *Pythagoras*, *69*, 3–16.

Johnstone, A. H. (2000). Teaching of Chemistry – logical or psychological? *Chemistry Education: Research and Practice in Europe*, *1*(1), 9–15.

Julie, C., Holtman, L., & Mbekwa, M. (2011). Rasch modelling of Mathematics and Science teachers' preferences of real-life situations to be used in Mathematical literacy. *Pythagoras*, *32*(1), 30–38.

Likert, R. (1932). A technique for the measurement of attitude scales. *Archives of Psychology*, *22*(140), 5–53.

Long, C. (2009). From whole number to real number: Applying Rasch measurement to investigate threshold concepts. *Pythagoras*, *70*, 32–42.

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

Masters, G. N. (1988). Item discrimination: When more is worse. *Journal of Educational Measurement*, *25*, 15–29.

McKelvie, S. (1978). Graphic rating scales – How many categories? *British Journal of Psychology*, *69*, 185–202.

McMillan, J. H., & Schumacher, S. (2010). *Research in education, evidence-based inquiry* (7th ed.). Upper Saddle River, NJ: Pearson.

Neumann, I., Neumann, K., & Nehm, R. (2011). Evaluating instrument quality in science education: Rasch-based analyses of a nature of science test. *International Journal of Science Education*, *33*(10), 1373–1405.

Potgieter, M., Davidowitz, B., & Venter, E. (2008). Assessment of preparedness of first-year chemistry students: Development and application of an instrument for diagnostic and placement purposes. *African Journal of Research in Mathematics, Science and Technology Education*, *12*, 1–18.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Chicago: University of Chicago Press.

Smith, R., & Plackner, C. (2009). The Family Approach to assessing fit in Rasch measurement. *Journal of Applied Measurement*, *10*(4), 424–437.

Tennant, A., McKenna, S. P., & Hagell, P. (2004). Application of Rasch analysis in the development and application of quality of life instruments. *Value in Health*, 7(supplement 1), S22–S26.

Tennant, A., & Conaghan, P. (2007). The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis Care & Research*, *57*(8), 1358–1362.

Van de Grift, W., & Van der Wal, M. (2010). Measuring the development of professional competence among teachers. Retrieved 17 August 2012 from http://www.icsei.net/icsei2011/Full%20Papers/0127_A.pdf.