CLINICAL AND EXPERIMENTAL
OPTOMETRY

RESEARCH PAPER

Multiplicative rating scales do not enable measurement of vision-related quality of life

Clin Exp Optom 2011; 94: 1: 52–62 DOI:10.1111/j.1444-0938.2010.00554.x

Vijaya K Gothwal[*†] BOpt MAppSci PhD
Thomas A Wright[*] BPsyc(Hons)
Ecosse L Lamoureux[§‖¶] BEd GradDip MAppSci PhD
Konrad Pesudovs[*] BScOptom PhD PGDipAdvClinOptom FACO FAAO FCLSA
* NHMRC Centre for Clinical Eye Research, Discipline of Ophthalmology and Discipline of Optometry and Vision Science, Flinders Medical Centre and Flinders University of South Australia, Bedford Park, South Australia, Australia
† Meera and L B Deshpande Centre for Sight Enhancement, Vision Rehabilitation Centres, L V Prasad Eye Institute, Hyderabad, India
§ Centre for Eye Research Australia, Department of Ophthalmology, University of Melbourne, Victoria, Australia
‖ Vision CRC, Sydney, Australia
¶ Singapore Eye Research Institute, Singapore National Eye Centre, Singapore
E-mail: Konrad.Pesudovs@flinders.edu.au

Submitted: 2 November 2009
Revised: 13 May 2010
Revised: 2 July 2010
Accepted for publication: 17 August 2010

Purpose: Many questionnaires for the measurement of visual impairment exist. One, the Houston Vision Assessment Test (HVAT), takes a different approach: the patient is asked to rate overall impairment and the proportion attributed to vision, then through multiplication the visual and non-visual (physical) impairments are calculated. The purpose of this study was to determine whether the scores derived from this approach can be considered to be measures.

Methods: The participants were 193 cataract patients awaiting surgery (mean age 74.1 ± 9.8 years, 54 per cent female and 53.6 per cent were awaiting first eye surgery), who self-administered the HVAT, which consists of 10 questions, whereby impairment on each activity and the proportion attributable to vision is rated. Therefore, total, visual and physical impairments are calculated. For each question, multiplying the impairment (five response categories) by the proportion due to eyesight (five categories) gives 10 possible levels of visual impairment. Assessment of the multiplicative rating scales included frequency of category use and hierarchical ordering of response categories using category thresholds. Summary statistics of Rasch analysis were generated for the rating scale and overall questionnaire performance.

Results: In the multiplicative scale, higher response categories were under-utilised and thresholds were disordered, indicating that the categories did not function as intended. Some of the dysfunction arose from disordered thresholds in the 'proportion due to eyesight scale', but repairing this gave little improvement to the multiplicative scale. The ill-defined nature of the disordered categories precluded further repair by combining categories. Measurement precision, as indicated by person separation reliability, was poor (0.70).

Conclusion: Rasch analysis demonstrated that the categories of the multiplied rating scale of the HVAT were not ordered, as the user would expect; this precludes measurement. This provides evidence against the use of multiplicative rating scales in quality-of-life questionnaires. It would be better to use a single rating scale for each construct of interest.

Key words: cataract, psychometrics, quality of life, questionnaires, visual disorders

Patient-reported outcomes (that is, questionnaires) are increasingly being used in the assessment of outcomes in health care and sound psychometric properties improve the chances that correct decisions are made from questionnaires.[1–4] Commonly, questionnaires assess a single attribute such as difficulty (for example: how much difficulty do you have walking up stairs?) or intensity (for example: how severe is your pain?).[5] There are questionnaires that assess more than one attribute,[6] for example, intensity and frequency of symptoms are evaluated together in the Functional Assessment of Chronic Illness Therapy (FACIT).[7] The value of using more than one attribute has been questioned, with increase in respondent burden being cited as a potential problem.[8] In ophthalmology, several questionnaires ask patients to rate two attributes, for example importance and difficulty in performing a task, and thereby incorporate a two-part rating scale.[9,10] Examples include the Activity Inventory (AI),[11,12] weighted version of the Melbourne Low-Vision Activities of Daily Living Index (MLVAI)[9] and the Visual Disability Questionnaire (VDQ).[10] The MLVAI weights the individual scores from the importance of and difficulty with each question to generate a composite score (importance–difficulty representing the disability impact) for the question. For example, an importance score of 4 is multiplied by a difficulty score of 4 to provide a composite disability impact score of 16 for a particular question. Conversely, in the AI and VDQ, importance and difficulty are treated as separate variables and the individual rating scales have been shown to perform well.[10–12]

The Houston Vision Assessment Test (HVAT) published in 2000 has been proposed as an instrument to guide the decision-making process between the patient and the ophthalmologist when considering cataract surgery.[13] The HVAT uses a two-part rating scale with the hypothesis that performing an activity requires a non-visual (for example, physical or cognitive) and a visual component, so both non-visual and visual impairments can be estimated and together these form the total impairment. In each question, patients estimate their total impairment on an activity (part A) and the proportion of impairment attributable to vision (part B). The values representing the choices made in parts A and B are multiplied together to give the visual impairment. The difference between total impairment and visual impairment is considered to be the physical impairment. Values for total, visual and non-visual impairments can be averaged across all answered questions to give an overall score, which is expressed as a percentage. Using rudimentary psychometric analysis, the HVAT has high internally consistent reliability (Cronbach's alpha = 0.96 preoperatively) and validity.[13] The scoring approach attempts to use Classical Test Theory (CTT) to provide the measurement. The CTT approach has other well-recognised shortcomings.[14–18] A major limitation, as it relates to HVAT, is the use of ordinal values (or numbers) as scores assigned to different response categories, with the underlying assumption of equal distances between response categories along the variable measured. This assumption is commonly invalid because numbers or scores only indicate an ordering relationship and do not represent counts of equal units of measurement, due to which the scores do not share the interval properties of measures. Consequently, raw scores cannot be considered as measures.[13,17,19] The issue becomes further complicated with the use of a two-part multiplicative rating scale and specifically the philosophy that underpins the development of the HVAT (that is, to disentangle visual from non-visual impairment). Therefore, the validity of such rating scales may be questioned,[20] and the ordering and spacing of response categories are easily tested with Rasch analysis.

Owing to several advantages, Rasch analysis is increasingly being used in the assessment of health care outcomes. Specifically, Rasch analysis allows improved assessment of the functioning of rating scales (or response categories) and enables re-engineering, if required. Furthermore, by converting ordinal level data (that is, responses on a Likert-type rating scale) to interval-level data, Rasch analysis provides a truly linear scale.[21,22] Such conversion of the scoring is beneficial, as it reduces noise and justifies the use of parametric statistical analysis. An important feature of Rasch analysis is the management of missing data, which commonly occurs, for example, for driving items, in cataract populations. In Rasch models, missing data are accounted for scientifically, as Rasch analysis computes an estimate from the available data.[23] Therefore, missing data are not problematic, other than the small loss of precision that occurs when values are estimated from fewer items.

Despite its publication a decade ago, the HVAT has been rarely used. While reasons for the lack of its use are unknown, the HVAT remains within the realm of questionnaires developed for use in the cataract population. Therefore, its psychometric performance should be tested and we are specifically interested to test the validity of the rating scale approach. Therefore, the objective of the present study was to investigate the functioning of the rating scale of the HVAT using Rasch analysis, specifically to determine if raw scores from the HVAT can be considered to represent measures (that is, a linear scale). We also aimed to repair any dysfunction in the rating scales, if possible, so as to maximise the measurement properties of the questionnaire.[24]

METHODS

The Houston Vision Assessment Test

This test (the HVAT) consists of 10 items, which are listed in Table 1.[14] The questions involve a two-part answer. Part A asks participants to rate the amount of total impairment ranging from 'not at all limited'—0, 'slightly limited'—1, 'somewhat limited'—2, 'moderately limited'—3, and 'severely limited'—4. An additional response category marked by * is used to account for a participant who does not perform the given activity and therefore constitutes missing data. Part B then asks the participants to rate the proportion of

Item no.	Item description
1A	To what extent is your cooking impaired?
2A	To what extent is your driving at night impaired by oncoming headlights?
3A	To what extent is your driving during the day impaired?
4A	To what extent is your housework impaired?
5A	To what extent are your leisure activities impaired?
6A	To what extent are your outdoor activities impaired?
7A	To what extent is your reading impaired?
8A	To what extent is your taking medicine impaired?
9A	To what extent is your watching TV impaired?
10A	To what extent is your writing impaired?
11	How certain do you feel about all the answers you gave?

Part 'B' for each of the items: If there are limitations, how much is because of eyesight? Response options for this include 'I have no visual or other limitations', 'none due to eye sight', 'some due to eye sight', 'half due to eyesight', 'most due to eyesight' and all due to eyesight'.

Table 1. Item content of the Houston Vision Assessment Test

the limitation due to eyesight. There are six response categories ranging from zero to five, with zero being assigned if a participant does not have any visual or physical limitation. Given that part B asks the participants to report how much of the limitation is due to visual problems, category zero was considered as missing data and therefore we rescored the remaining five categories from zero to 4 (from 'none due to eyesight' (zero per cent, scores 0)—0, 'some due to eyesight' (25 per cent, scores 0.25)—1, 'half due to eyesight' (50 per cent, scores 0.50)—2, 'most due to eyesight' (75 per cent, scores 0.75)—3 and 'all due to eyesight' (100 per cent, scores 1.0)—4 for each activity. The proportion due to eyesight is used to calculate how much of the total impairment is caused by vision (that is, the amount of visual impairment). Thus, visual impairment is determined by multiplying part A by the part B rating. For example, if, for one question, a participant chose the response option 'severely limited', which is scored as 4 in part A, and chose 'all due to eyesight', which is scored as 100 per cent (represented by 1.0) for part B, the visual impairment for this question would be the score in part A multiplied by the weighted score in part B or 4 multiplied by 1.0 = 4.0. Non-visual impairment is calculated in the same way except that part A is multiplied by 1—the weighted score in part B. Thereby the percentage non-visual impairment will be the residual non-visual proportion of total impairment. Overall values for total, visual and non-visual impairments are calculated by averaging across answered questions and converting scores into percentages. Therefore, the overall percentage visual impairment by the formula: (sum of individual visual impairment scores / (number of pairs completed × 4.0)) × 100, to yield a range of 1–100. A higher score indicates greater impairment.

Participants

Participants were 208 patients of the Flinders Eye Centre, Adelaide, South Australia, who were currently on the waiting list (average waiting period was three to four months) for cataract surgery. All participants were aged 18 years or older, English speaking and were able to provide written consent. Responses from 15 participants (seven per cent) were deleted, as they were uncertain of their responses, thereby leaving 193 participants in the study. Table 2 summarises their characteristics. This study sample appears to be representative of the elderly cataract population in Australia.[25]

Participants were mailed the HVAT along with the demographic data form for self-administration and return via a self-addressed envelope. Ethical approval was obtained from the Flinders Clinical Research Ethics committee. All patients who agreed provided informed consent. The study was conducted in accordance with the Declaration of Helsinki.

Clinical assessment

For demographic purposes, we report clinical data that were collected prior to cataract extraction. Visual acuity was measured using computerised testing based on logMAR principles with a screen illumination of 150 cd/m^2. All assessments were performed binocularly as binocular acuity was considered representative of real-world ability.[26,27]

Rasch analysis

We examined the visual impairment scale as originally proposed by the developers (that is, a multiplicative scale of parts A and B). If this was found to be dysfunctional, we examined each part individually (that is, part A and part B were analysed separately) to try to find the cause and potential remedial strategy for the problem.

Rasch analysis[28] was performed with the Andrich rating scale model[29] using Winsteps software (version 3.68).[30] The Rasch model and the process of Rasch analysis is described in detail elsewhere.[10,31–33]

The main variables of the Rasch model are 'person ability' and 'item difficulty', both of which are estimated in logits (that is, log-odd units). In Rasch analysis, the probability of a person choosing a response category for an item (or activity) depends on the 'person ability', as well as the 'item difficulty'. If the person's ability in performing an activity is lower than the required visual ability for that activity, then the probability of the person rating the task in the highest (that is, worse) scoring category is high. Conversely, if the person's ability is higher than the required

Characteristic	n (%) or mean ± standard deviation
Age (years)	74.1 ± 9.8
Gender	
Male	88 (45.6%)
Habitual binocular visual Acuity	
logMAR, Snellen	
	0.22 ± 0.21 (6/9.5^{-1}), range -0.26 to 1.00 (6/3^{-2} to 6/60)
Awaiting second-eye surgery	82 (42.5%)
Ocular co-morbidity*	
Present	90 (46.6%)
Systemic co-morbidity#	
Present	135 (69.9%)

* Includes glaucoma, diabetic retinopathy, age-related macular degeneration et cetera.
# Includes diabetes, hypertension, angina et cetera.
logMAR = logarithm of minimum angle of resolution

Table 2. Characteristics of participants who completed the Houston Vision Assessment Test (n = 193)

visual ability, then the probability of the person rating the task in the lowest (that is, better) scoring category is high. A relatively easy item and a less able person (that is, with greater visual impairment) will have positive logits for the HVAT.[34]

A sequence of analyses was performed. First, the use of response categories by the participants was examined. Response categories for items should form a continuum of less to more. That is, endorsing a lower category should represent being less visually impaired (in the case of HVAT) than endorsing a higher category. Specifically, the location of thresholds was investigated.[35,36] Thresholds are boundaries between contiguous response categories. More appropriately, they correspond to the location along the latent trait or construct (visual impairment in this case), where two adjacent response categories have an equal probability of selection and thus the boundaries between the response categories are delineated.[37] These thresholds should demonstrate an ordered behaviour (discussed in detail below). Lack of order in the response categories (as would be evident by disordered thresholds) suggests a lack of common understanding of the use of these categories between the designer and the participants. Consequently, the item fit and placement are affected. Such problems can be resolved, albeit post hoc, by combining categories and then reanalysing the data to determine the optimal number of response categories for the questionnaire (HVAT in this case).[38] If the disordered thresholds could not be remedied for the HVAT, further aspects of Rasch analysis, such as item reduction (that is, delete misfitting or otherwise problematic items so as to retain the most eligible or fitting items[34]), person–item map (that is, the placement of items and people on a common scale, for example, from most to least difficult items and most to least visually disabled people), item difficulty and person ability parameters and unidimensionality (that is, whether all the items in HVAT measure a single construct, an important prerequisite for the use of a total or summary score[39]) were not pursued.[22] Nevertheless, given the bi-dimensional nature of the rating scale of the HVAT, a lack of unidimensionality can be expected.

The presence of functioning response categories (that is, ordered thresholds) is the fundamental requirement of a questionnaire. In the absence of this, the reporting of further attributes of Rasch analysis described above is unacceptable.

The investigation of the functioning of the rating scale of the HVAT in the Rasch model included an assessment of the following parameters:

CATEGORY USE STATISTICS

There are five response categories (assigned values 0, 1, 2, 3, 4) each in part A and B of the HVAT. Each category of part A is multiplied by that of part B resulting in a multiplicative scale (A by B) or (0,1,2,3,4) by (0,1,2,3,4) and thereby contains 10 levels. These raw values of 10 levels would be (0,1,2,3,4,6,8,9,12,16) but are assigned (0,1,2,3,4,5,6,7,8,9) in Rasch analysis because the software would otherwise interpret the raw category labelling to contain intermediate categories between the labels (that is, 17 categories from zero to 16).

Category frequency and average measure represent the use of categories by the participants.[40] As the name implies, category frequency indicates how many participants used a particular response category. In the Andrich model, category responses are effectively averaged across all items. Linacre[41] recommends that there be at least 10 responses in each category. Categories used less frequently are often unnecessary or redundant.[40]

Based on his/her visual impairment, a participant (or person) chooses a particular response category and this represents the observed response. For example, the observed response will be 1 for a participant who chose 'slightly limited' on part A of a question on the HVAT. The average measure for a particular category is defined as the average of the difference between the person's ability and the item difficulty across all observed responses in that category.[40] For example, if, for category 1, the average measure recorded was -2.0, then this value can be interpreted as the average ability estimate for all participants who used a response category of 1 for any item on the HVAT. Participants with greater or more severe visual impairment are likely to choose higher-ranked

response categories. Therefore, the average measures are expected to increase monotonically with the response categories. Lack of monotonic increase in the average measures indicates that the categories do not function in the intended order.

FIT STATISTICS

As the Rasch model is probabilistic and not deterministic, some failure of the model to predict the observed responses can be expected. This amount of discrepancy is represented by the mean-square (MnSq) fit statistics and there are two types of fit statistics, infit and outfit.[42] While each of these identifies how well a category is used, the infit statistic is less sensitive to distortion from outliers and is therefore considered more informative of the two.[40,41] We report the infit MnSq and the expected value is 1.00. An infit MnSq greater than 1.00 indicates the presence of noise. For example, a value of 1.10 would indicate 10 per cent more noise in the response than expected. Generally, an infit MnSq up to 1.30 (30 per cent more variance than expected) is considered acceptable. The fit of each response category to the model was tested.

THRESHOLDS

Similar to the category average value, threshold locations should advance monotonically.[38,43] Thresholds that do not increase monotonically are considered disordered.[29] The distance between the threshold estimates is also critical. Ideally the distance should be at least 1.4 logits, in order to eliminate the likelihood that thresholds would be disordered in different populations.[41] Threshold disordering suggests that the response scale is not working adequately to order participants with distinct levels of ability. A common solution used for disordered thresholds is to collapse adjacent categories, specifically when categories are under-utilised (as indicated by low category frequencies) or used inconsistently (as indicated by disordered thresholds).[24,40] Consequently, each rating category would represent a distinct level of ability, compared with the adjacent category. When collapsing categories, the examiner should be guided by the person separation statistic.

Of these various indices, category use statistics and thresholds are the most commonly used diagnostic statistics to investigate functioning of the rating scale; however, these statistics should not be used in isolation.[40] In addition, Rasch analysis provides the person separation reliability, which is used to evaluate the extent to which the items in the questionnaire can distinguish between participants regarding the level of the measured construct (visual impairment in the present case).[35,44,45] A value of 0.80 or greater is acceptable indicating that three strata or groups of participants can be differentiated.[46]

For the HVAT to be considered as a measure of visual impairment, the multiplicative scale categories should demonstrate average measures that increase monotonically, with ordered thresholds, sufficient distance between thresholds, satisfactory fit statistics and the instrument should have sufficient person separation to differentiate the participants.

RESULTS

Missing data ranged from 8.3 per cent ('writing') to 38.9 per cent ('night driving'). We present the analyses related to a multiplicative scale of parts A and B first, and then individually (that is, part A and part B analysed separately).

Category use statistics

Tables 3 and 4 show the category use statistics for the 10-level (multiplicative scale) and the five-level separate rating scale for each of parts A and B. Except for categories 0 and 1, the remaining eight categories were significantly under-utilised (columns 2 and 3), leading to a skewed distribution of categories in the multiplicative scale (Table 3). Nevertheless, a monotonic increase in the average measures is evident.

In comparison, the number of responses in most of the categories is high when each of the rating scales was analysed individually (Table 4). Furthermore, the distribution across categories (columns 2 and 3) is regular in part B but skewed in part A, as with the multiplicative scale. Coinciding with this, there was a monotonic increase in the average measures for parts A and B.

Fit statistics

In Table 3, categories 0, 6, 7 and 9 have infit MnSq values substantially high, signifying that more noise than expected by the Rasch model was present (Table 3).

In comparison, the fit statistics for the individual rating scales for parts A and B (Table 4) show infit MnSq within acceptable limits for all categories except category four for part A and categories zero and four (that is, the extreme categories) for part B, signifying that these categories contain more noise than expected by the Rasch model.

Thresholds

Although the category average measures appear to advance with increasing categories, the separation between the estimates was very small for the multiplicative scale (Table 3). For example, only 0.07 logits separated category three from four. This conceals a problem with category use, which is exposed when observing the response category thresholds in the category probability curves. Figure 1 illustrates the category probability curves (CPC) that participants with a given level of visual impairment will select as a response category. The CPC plots visual impairment as a continuum on the *x*-axis against the probability of endorsing each response category on the *y*-axis. Each curve corresponds to one response category. Thresholds correspond to projection, on the x-axis, of intersections between successive category probability curves.[47] For an optimally functioning rating scale, each category should be the most likely used category along the width of scale (x-axis) and should appear like a range of hills. Figure 1 shows that the estimates of the thresholds, which define the categories, do not form distinct regions of the continuum. Instead, there is crowding, suggesting that these 10 levels are not distinctly defined for the multiplicative rating scale of the HVAT.

Category label	Category count	Category %	Average measure	MnSq fit statistics		Threshold calibration
				Infit	Outfit	
From 'not at all' to 'severely' limited, all with 'none due to eyesight'	664	50	-1.68	1.43	1.10	
Slightly limited and Some due to eyesight	283	21	-1.08	0.98	0.62	-0.97
Somewhat limited and Some due to eyesight/Slightly limited and Half due to eyesight	83	6	-0.75	0.69	0.53	-0.46
Moderately limited and Some due to eyesight/Slightly limited and Most due to eyesight	42	3	-0.54	0.89	0.51	-0.34
Severely limited and Some due to eyesight/Somewhat limited and Half due to eyesight	69	5	-0.40	1.22	0.63	-0.27
Somewhat limited and Most due to eyesight/Moderately limited and Half due to eyesight	49	4	-0.19	0.82	0.57	-0.14
Severely limited and Half due to eyesight/Somewhat limited and All due to eyesight	27	2	-0.06	1.38	1.99	-0.02
Moderately limited and Most due to eyesight	29	2	-0.03	1.27	1.38	0.06
Severely limited and Most due to eyesight/Moderately limited and All due to eyesight	56	4	0.19	1.23	1.19	0.17
Severely limited and All due to eyesight	34	3	0.38	2.46	2.87	0.54

MnSq = mean-square
Category count does not add up to 1930 because of missing data

Table 3. Category use statistics for 10-level Houston Vision Assessment Test rating scale

In comparison, category thresholds were ordered for part A when the two parts (that is, A and B) were analysed separately (Figure 2). Such ordered thresholds indicate consistency of the participants' use of the categories with the philosophy of the developers. Furthermore, the end categories ('slightly limited' and 'severely limited' in part A and 'none due to eyesight' and 'all due to eyesight' in part B) were well spaced from the adjacent categories in each of the rating scales.

Category label	Category count	Category %	Average measure	MnSq fit statistics		Threshold calibration
				Infit	Outfit	
Part A						
Not at all limited	679	47	-3.43	0.97	0.96	
Slightly limited	435	30	-1.49	1.03	0.65	-2.15
Somewhat limited	147	10	-0.35	1.01	1.03	-0.24
Moderately limited	115	8	0.69	0.91	0.89	0.57
Severely limited	75	5	1.70	1.92	2.21	1.82
Part B						
None due to eyesight	175	19	-3.07	1.36	1.02	
Some due to eyesight	456	50	-1.27	0.87	0.78	-3.00
Half due to eyesight	68	7	0.11	0.82	0.70	0.32
Most due to eyesight	106	12	1.09	0.86	0.75	0.72
All due to eyesight	102	11	2.25	1.30	1.47	1.91

MnSq = mean-square
Category count does not add up to 1930 because of missing data for both the parts A and B

Table 4. Category use statistics for the 5-level HVAT (parts A and B) rating scale

Examination of the category probability curve for part B in Figure 3 shows disordered thresholds (that is, category 2 'half due to eyesight' does not have a range along the scale where it is the most likely category to be selected). As the evident cause of response category dysfunction, categories 2 and 3 were combined and the following rescaling was adopted: 0 = 0; 1 = 1; 2 = 2; 3 = 2; 4 = 3. The four resulting categories represent the following options: 0 = 'none due to eyesight', 1 = 'some due to eyesight', 2 = 'half or most due to eyesight', and 3 = 'all due to eyesight' (Table 5). Following category collapsing, the curves were ordered with each of the categories emerging as the most likely used category (Figure 4).

Person separation

The person separation reliability was below the acceptable level for the multiplicative rating scale (0.70, while the minimum acceptable level is 0.80). For the individual rating scales, person separation reliability was acceptable (0.84) for part A but sub-optimal (0.73) for part B.

Given the category collapse required in the part B scale, we assessed the product of the rating scale obtained by multiplying the revised rating scale (part B) by part A. The four new categories formed for part B were multiplied by the original categories of part A resulting in a nine-level rating scale. The characteristics of this rating scale were similar to the original multiplicative scale. Person separation reliability was still below the acceptable level (0.74) and there was a lot of disordering of category thresholds. The potential to collapse categories to repair the multiplicative scale could not be undertaken because it is unclear which categories should be combined. As the critical aspect of Rasch analysis, that is, the rating scale, could not be fixed, further investigation of the psychometric properties of the HVAT using Rasch analysis was not undertaken.

DISCUSSION

Although the optimal number of rating categories for questionnaires has been explored, to the authors' knowledge, no research has tested the functioning of a multiplicative rating scale, as used in the HVAT, using Rasch analysis. This is appropriate because any rating scale must contain response options, which progress in an ordered manner along the scale under test; Rasch analysis easily examines this. When subjected to Rasch analysis the multiplicative rating scale of the HVAT failed to meet the criteria for a functioning rating scale. This is important because appropriately designed

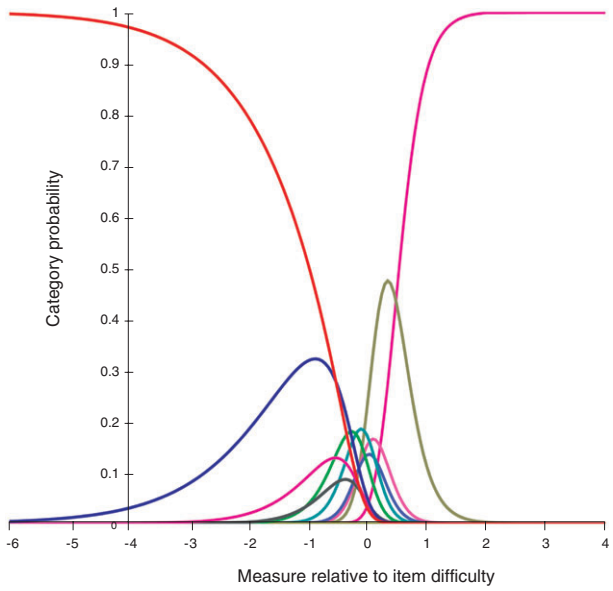

Figure 1. Category probability curves for all intermediate categories (red represents from 'Not at all to Severely limited, all with None due to eyesight', blue represents 'Slightly limited and Some due to eyesight', pink represents 'Somewhat limited and Some due to eyesight/Slightly limited and Half due to eyesight', black represents 'Moderately limited and Some due to eyesight/Slightly limited and Most due to eyesight', green represents 'Severely limited and Some due to eyesight/Somewhat limited and Half due to eyesight', aqua represents 'Somewhat limited and Most due to eyesight/Moderately limited and Half due to eyesight', light blue represents 'Severely limited and Half due to eyesight/Somewhat limited and All due to eyesight', light pink represents 'Moderately limited and Most due to eyesight', brown represents 'Severely limited and Most due to eyesight/Moderately limited and All due to eyesight', purple represents 'Severely limited and All due to eyesight'). There are disordered thresholds for all except the two extreme categories. None of the intermediate categories emerged as the most likely category to be chosen for a given part of the scale. Thus the probability of observing these categories is lower than the probability of observing the extreme categories, whatever the participant's location along the scale.

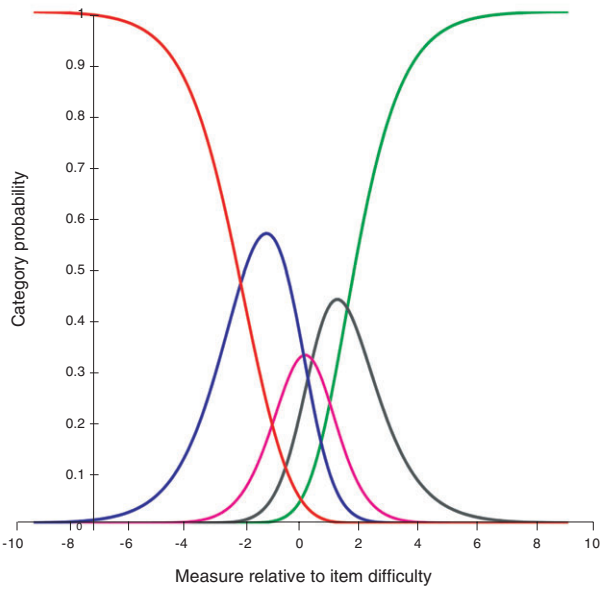

Figure 2. Category probability curves demonstrating ordered thresholds for part A of the rating scale showing that it functions, although response category three is not the most likely category to be chosen for most of the visual impairment scale. Red represents 'not at all limited', blue represents 'slightly limited', pink represents 'somewhat limited', grey represents 'moderately limited' and green represents 'severely limited'.

rating scales form the basis for data collection and flaws at this level are carried through all stages of data analysis. Disordered thresholds suggested that the HVAT was unable to distinguish participants' abilities to the degree suggested by the rating scale. Disordering of thresholds was supported by the category probability curve and the close location (less than 1.4 logits) of the thresholds. This called for collapsing of adjacent categories and data re-analysis. While categories could be collapsed for part B, the ability of HVAT to differentiate participants (that is, person separation reliability) remained low. Due to overcrowding of the categories in the multiplicative rating scales, no logical combining of categories was evident. Therefore, repair was not possible.

The Rasch methodology used in the present study provided several useful indicators to investigate rating scale categorisation. Fundamental concerns regarding a questionnaire include whether or not it has a validly functioning rating scale and whether the number of response categories is optimal.[21,48] Although polytomous response formats do offer more information than dichotomous response

Category label	Category count	Category %	Average measure	MnSq fit statistics		Threshold calibration
				Infit	Outfit	
None due to eyesight	175	19	-3.67	1.18	1.09	
Some due to eyesight	456	50	-1.43	0.83	0.74	-3.41
Half/Most due to eyesight	174	19	1.17	0.76	0.71	0.60
All due to eyesight	102	11	3.18	1.40	1.60	2.81

MnSq = mean-square
Category count does not add up to 1930 because of missing data

Table 5. Category statistics for the four-level HVAT (part B) rating scale

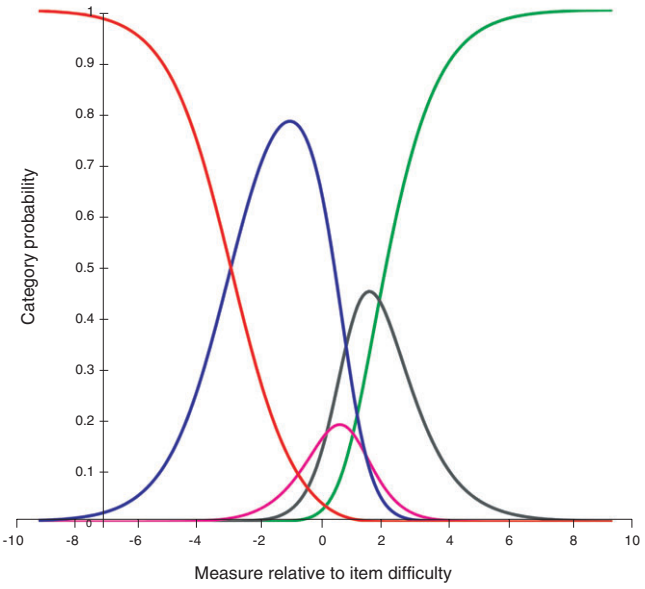

Figure 3. Category probability curves (red represents 'none due to eyesight', blue represents 'some due to eyesight', pink represents 'half due to eyesight', grey represents 'most due to eyesight' and green represents 'all due to eyesight') showing disordered thresholds. The response category 'half due to eyesight' does not have a range along the scale, where it is the most likely category to be selected. Therefore, it is less likely to be endorsed by the participants and is used interchangeably with the category 'most due to eyesight'.

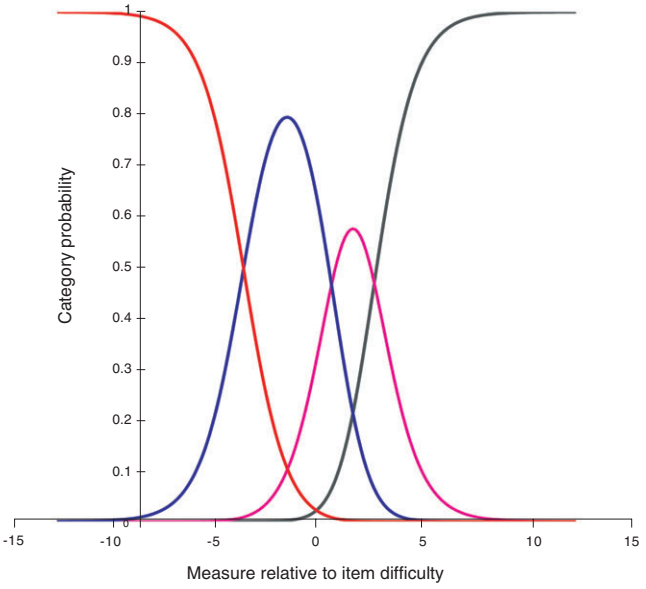

Figure 4. Following the collapse of categories in Figure 3, category probability curve thresholds become ordered for part B of the rating scale. Red represents 'none due to eyesight', blue represents 'some due to eyesight', pink represents 'half or most due to eyesight' and grey represents 'all due to eyesight'.

options, beyond a certain number the options can be confusing. Indeed, it has been shown that measurement precision tends not to be assisted by more than five categories.[49,50] In current practice, most questionnaires in health care, including those in optometry and ophthalmology, generally contain four or five Likert-style categories that assess a single attribute. With higher numbers of response categories, the likelihood of response categories being indistinguishable increases and this compromises the validity of the ratings. The HVAT has essentially five categories for each of parts A and B for every item and both parts address the same item, albeit in different aspects. Consequently, there are 10 categories for each pair of items and this would be too many for the respondents to distinguish, so the rating scale dysfunction perhaps was expected.

The fundamental problem with the large number of categories is significant under-utilisation of some categories. Stable threshold estimations require a sufficient number (at least 10) of responses per category.[41] Increasing the number of categories in the anticipation of finer discrimination between responses appears to be counter-productive.[40] Under-utilisation of categories appears to be a frequently encountered problem in visual impairment questionnaires that use many rating categories.[46] In our assessment of the various combinations of rating scale formats of the HVAT, the responses were skewed to the lower categories (more than 60 per cent of responses were in the lowest two categories) indicating little impairment. This is most likely due to the participants having quite good visual acuity, as is evident from Table 2; however, this does not indicate that the participants did not have difficulty with day-to-day activities because they were drawn from the cataract surgery waiting list, where the key indicator for cataract surgery is the presence of visual impairment arising from cataract.[51–53] In previous studies,[25,54–57] these participants have been reported to have a visual impairment. Furthermore, while most of our participants used the lower end of the response categories, we also had participants, who used the higher response categories. As mentioned above, our aim was not a comprehensive evaluation of the HVAT, rather we were interested in examining the design of its rating scale using Rasch analysis. For this purpose, our data appear satisfactory.

This study examining the functioning of the rating scales of the HVAT indicated the invalidity of these scales. This calls into question previous studies that have used the HVAT.[13] Results of the present study also suggest that if the aim is to measure both visual and physical impairments, it would be more appropriate to have a separate questionnaire, or at least a separate question, to assess each concept. This approach would treat each concept as a separate variable, as has been done in the use of AI[11,12] and VDQ.[13] The results of this study must also be considered to call into question other instruments that use multiplicative rating scales, such as the MacDQoL, the RetDQoL and the ADDQoL.[58–60]

Compared with some of the legacy questionnaires that were developed using CTT and were subsequently re-examined using Rasch analysis,[46,61] the recent questionnaires have been developed using Rasch analysis to circumvent the need for later re-validation.[62–64] Among the limitations of CTT are the use of values of internal consistency to select the final set of items in a questionnaire, which might lead to retaining redundant items and thereby reducing the ability of the questionnaire to differentiate between the strata of participants.[65] This might explain, in part, the reason for the poor performance of the HVAT in terms of inadequate person separation reliability.

In conclusion, this analysis using HVAT as an illustrative example demonstrates that rating scale design is an important consideration when choosing questionnaires for health-care research. The HVAT does not meet the criteria for optimal functioning of the rating scale and it is hoped that researchers who are new to the field of questionnaire research will consider the issues that need to be avoided as well as considered in designing rating scales for their questionnaires. The problem could be avoided simply by choosing a rating scale that directly asks patients to rate their visual impairment. Several good Rasch-scaled questionnaires exist for this purpose.[55,57,66]

ACKNOWLEDGEMENTS

The authors would like to thank all participants who volunteered to participate in the study. We thank the consultant staff of the Discipline of Ophthalmology at Flinders Medical Centre for providing access to cataract surgery patients for research purposes.

GRANTS AND FINANCIAL SUPPORT

This study was supported in part by the National Health and Medical Research Council (Canberra, Australia) Centre of Clinical Research Excellence Grant 264620. Konrad Pesudovs is supported by the National Health and Medical Research Council (Canberra, Australia) Career Development Award 426765.

CONFLICT OF INTEREST

The authors have no personal financial interest in the development, production or sale of any device discussed herein.

REFERENCES

1. Fayers PM. Applying item response theory and computer adaptive testing: the challenges for health outcomes assessment. *Qual Life Res* 2007; 16 (Suppl 1): 187–194.
2. Lin JH, Wang WC, Sheu CF, Lo SK, Hsueh IP, Hsieh CL. A Rasch analysis of a self-perceived change in quality of life scale in patients with mild stroke. *Qual Life Res* 2005; 14: 2259–2263.
3. Revicki DA, Cella DF. Health status assessment for the twenty-first century: item response theory, item banking and computer adaptive testing. *Qual Life Res* 1997; 6: 595–600.
4. Massof RW, Rubin GS. Visual function assessment questionnaires. *Surv Ophthalmol* 2001; 45: 531–548.
5. Williams LS, Weinberger M, Harris LE, Clark DO, Biller J. Development of a stroke-specific quality of life scale. *Stroke* 1999; 30: 1362–1369.
6. Wright JG, Young NL. The patient-specific index: asking patients what they want. *J Bone Joint Surg Am* 1997; 79: 974–983.
7. Cella D. The Functional Assessment of Cancer Therapy-Anemia (FACT-An) Scale: a new tool for the assessment of outcomes in cancer anemia and fatigue. *Semin Hematol* 1997; 34: 13–19.
8. Chang CH, Cella D, Clarke S, Heinemann AW, Von Roenn JH, Harvey R. Should symptoms be scaled for intensity, frequency, or both? *Palliat Support Care* 2003; 1: 51–60.
9. Haymes SA, Johnston AW, Heyes AD. A weighted version of the Melbourne Low-Vision ADL Index: a measure of disability impact. *Optom Vis Sci* 2001; 78: 565–579.

10. Marella M, Gothwal VK, Pesudovs K, Lamoureux E. Validation of the visual disability questionnaire (VDQ) in India. *Optom Vis Sci* 2009; 86: E826–E835.
11. Massof RW, Ahmadian L, Grover LL, Deremeik JT, Goldstein JE, Rainey C, Epstein C et al. The Activity Inventory: an adaptive visual function questionnaire. *Optom Vis Sci* 2007; 84: 763–774.
12. Massof RW, Hsu CT, Baker FH, Barnett GD, Park WL, Deremeik JT, Rainey C et al. Visual disability variables. I: the importance and difficulty of activity goals for a sample of low-vision patients. *Arch Phys Med Rehabil* 2005; 86: 946–953.
13. Prager TC, Chuang AZ, Slater CH, Glasser JH, Ruiz RS. The Houston Vision Assessment Test (HVAT): an assessment of validity. The Cataract Outcome Study Group. *Ophthalmic Epidemiol* 2000; 7: 87–102.
14. Novich MR. The axioms and principal results of classical test theory. *J Math Psychol* 1966; 3: 1–18.
15. Allen MJ. Introduction to Measurement Theory. California: Brooks/Cole, 1979.
16. Crocker L, Algina J. Introduction to Classical and Modern Test Theory. Forth Worth, Texas: Harcourt, Brace, Jovanovich, 1986.
17. Massof RW. The measurement of vision disability. *Optom Vis Sci* 2002; 79: 516–552.
18. Hambleton RK, Swaminathan H, Rogers HJ. Fundamentals of Item Response Theory. Newbury Park: Sage, 1991.
19. Towensend JT, Ashby FG. Measurement scales and statistics: the misconceptions misconceived. *Psychol Bull* 1984; 96: 394–401.
20. Wright BD, Linacre JM. Observations are always ordinal; measurements, however, must be interval. *Arch Phys Med Rehabil* 1989; 70: 857–860.
21. Pesudovs K, Noble BA. Improving subjective scaling of pain using Rasch analysis. *J Pain* 2005; 6: 630–636.
22. Pesudovs K, Burr JM, Elliott DB. The development, assessment and selection of questionnaires. *Optom Vis Sci* 2007; 84: 664–675.
23. Conrad KJ, Smith EV Jr. International conference on objective measurement: applications of Rasch analysis in health care. *Med Care* 2004; 42: I1–I6.
24. Linacre JM. Investigating rating scale category utility. *J Outcome Meas* 1999; 3: 103–122.
25. Kirkwood BJ, Pesudovs K, Latimer P, Coster DJ. The efficacy of a nurse-led preoperative cataract assessment and postoperative care clinic. *Med J Aust* 2006; 184: 278–281.
26. Rubin GS, Bandeen-Roche K, Huang GH, Munoz B, Schein OD, Fried LP, West SK. The association of multiple visual impairments with self-reported visual disability: SEE project. *Invest Ophthalmol Vis Sci* 2001; 42: 64–72.
27. Elliott DB, Hurst MA, Weatherill J. Comparing clinical tests of visual function in cataract with the patient's perceived visual disability. *Eye (Lond)* 1990; 4: 712–717.
28. Rasch G. Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen, Denmark: Institute of Educational Research, 1960.
29. Andrich DA. A rating scale formulation for ordered response categories. *Psychometrika* 1978; 43: 561–573.
30. Linacre JM. WINSTEPS Rasch measurement computer program. Available at *winsteps.com*, 2008.
31. Gothwal VK, Wright T, Lamoureux EL, Pesudovs K. Psychometric properties of visual functioning index using Rasch analysis. *Acta Ophthalmol Scand* 2009, June 26. [Epub ahead of print].
32. Gothwal VK, Wright TA, Lamoureux EL, Pesudovs K. Rasch analysis of the quality of life and vision function questionnaire. *Optom Vis Sci* 2009; 86: E836–E844.
33. Haley SM, Ni P, Hambleton RK, Slavin MD, Jette AM. Computer adaptive testing improved accuracy and precision of scores over random item selection in a physical functioning item bank. *J Clin Epidemiol* 2006; 59: 1174–1182.
34. Wright BD, Stone MH. Best Test Design. Chicago, IL: MESA Press, 1979.
35. Wright BD, Masters GN. Rating Scale Analysis. Chicago: MESA Press, 1982.
36. Wright BD, Panchapakesan N. A procedure for sample-free item analysis. *Edu Psychol Meas* 1969; 29: 23–48.
37. Hawthorne G, Densley K, Pallant JF, Mortimer D, Segal L. Deriving utility scores from the SF-36 health instrument using Rasch analysis. *Qual Life Res* 2008; 17: 1183–1193.
38. Andrich D, de Jong J, Sheridan B. Diagnostic opportunities with the Rasch model for ordered response categories. In: Rost J, Kangehiene R, eds. Application of Latent Trait and Latent Class Models in the Social Sciences, 1997. p 59–70.
39. Tennant A, McKenna SP, Hagell P. Application of Rasch analysis in the development and application of quality of life instruments. *Value Health* 2004; 7 (Suppl 1): S22–S26.
40. Bond TG, Fox CM. Applying the Rasch Model: Fundamental Measurement in the Human Sciences. London: Lawrence Erlbaum Associates, 2001.
41. Linacre JM. Optimizing rating scale category effectiveness. *J Appl Meas* 2002; 3: 85–106.
42. Linacre JM. Categorical misfit statistics. *Rasch Meas Trans* 1995; 9: 450–451.
43. Linacre JM. Investigating rating scale category utility. *J Outcome Meas* 1999; 3: 103–122.
44. Fisher WP. Reliability statistics. *Rasch Meas Trans* 1992; 6: 238.
45. Wright BD. Reliability and separation. *Rasch Meas Trans* 1996; 8: 472.
46. Pesudovs K, Garamendi E, Keeves JP, Elliott DB. The Activities of Daily Vision Scale for cataract surgery outcomes: re-evaluating validity with Rasch analysis. *Invest Ophthalmol Vis Sci* 2003; 44: 2892–2899.
47. Zhu W. Should total scores from a rating scale be used directly? *Res Q Exerc Sport* 1996; 67: 363–372.
48. Likert R. A technique for the measurement of attitudes. *Arch Psycho* 1932; 140: 1–55.
49. Lissitz RW, Green SB. Effect of the number of scale points in reliability: a Monte-Carlo approach. *J Appl Psychol* 1975; 60: 10–13.
50. Jenkins GD Jr, Taber TD. A Monte-Carlo study of factors affecting three indices of composite scale reliability. *J Appl Psychol* 1977; 62: 392–398.
51. Goggin M, Pesudovs K. Assessment of surgically induced astigmatism: toward an international standard II. *J Cataract Refract Surg* 1998; 24: 1552–1553.
52. Pesudovs K, Coster DJ. Cataract surgery reduces subjective visual disability. *Aust N Z J Ophthalmol* 1997; 25 (Suppl 1): S3–S5.
53. Lamoureux EL, Hooper CY, Lim L, Pallant JF, Hunt N, Keeffe JE, Guymer RH. Impact of cataract surgery on quality of life in patients with early age-related macular degeneration. *Optom Vis Sci* 2007; 84: 683–688.
54. Gothwal VK, Wright TA, Lamoureux EL, Pesudovs K. Rasch Analysis of Visual Function and Quality of Life Questionnaires. *Optom Vis Sci* 2009; 86: 1160–1168.
55. Gothwal VK, Wright TA, Lamoureux EL, Pesudovs K. Cataract symptom scale: clarifying measurement. *Br J Ophthalmol* 2009; 93: 1652–1656.
56. Gothwal VK, Wright T, Lamoureux EL, Pesudovs K. Activities of Daily Vision Scale: What do the subscales measure? *Invest Ophthalmol Vis Sci* 2010; 51: 694–700.
57. Gothwal VK, Wright TA, Lamoureux EL, Pesudovs K. Using Rasch analysis to revisit the validity of the Cataract TyPE Spec instrument for measuring cataract surgery outcomes. *J Cataract Refract Surg* 2009; 35: 1509–1517.
58. Bradley C, Todd C, Gorton T, Symonds E, Martin A, Plowright R. The development of an individualized questionnaire measure of perceived impact of diabetes on quality of life: the ADDQoL. *Qual Life Res* 1999; 8: 79–91.
59. Brose LS, Bradley C. Psychometric development of the individualized retinopathy-dependent quality of life questionnaire (RetDQoL). *Value Health* 2010; 13: 119–127.
60. Mitchell J, Wolffsohn JS, Woodcock A, Anderson SJ, McMillan CV, Ffytche T, Rubinstein M et al. Psychometric evaluation of the MacDQoL individualised measure of the impact of macular degeneration on quality of life. *Health Qual Life Outcomes* 2005; 3: 25.
61. Lamoureux EL, Pallant JF, Pesudovs K, Hassell JB, Keeffe JE. The Impact of Vision Impairment Questionnaire: an evaluation of its measurement properties using Rasch analysis. *Invest Ophthalmol Vis Sci* 2006; 47: 4732–4741.
62. Gothwal VK, Lovie-Kitchin JE, Nutheti R. The development of the LV Prasad-Functional Vision Questionnaire: a measure of functional vision performance of visually impaired children. *Invest Ophthalmol Vis Sci* 2003; 44: 4131–4139.
63. Pesudovs K, Garamendi E, Elliott DB. The Quality of Life Impact of Refractive Correction (QIRC) Questionnaire: development and validation. *Optom Vis Sci* 2004; 81: 769–777.
64. Pesudovs K, Garamendi E, Elliott DB. The Contact Lens Impact on Quality of Life (CLIQ) Questionnaire: development and validation. *Invest Ophthalmol Vis Sci* 2006; 47: 2789–2796.
65. Prieto L, Alonso J, Lamarca R. Classical Test Theory versus Rasch analysis for quality of life questionnaire reduction. *Health Qual Life Outcomes* 2003; 1: 27.
66. Lundstrom M, Pesudovs K. Catquest-9SF patient outcomes questionnaire: nine-item short-form Rasch-scaled revision of the Catquest questionnaire. *J Cataract Refract Surg* 2009; 35: 504–513.

Corresponding author:
Professor Konrad Pesudovs
NHMRC Centre for Clinical Eye Research
Department of Ophthalmology
Flinders Medical Centre
Bedford Park SA 5042
AUSTRALIA
E-mail: Konrad.Pesudovs@flinders.edu.au