British Journal of Clinical Psychology (2006), in press
© 2006 The British Psychological Society

The British Psychological Society

www.bpsjournals.co.uk

An introduction to the Rasch measurement model: An example using the Hospital Anxiety and Depression Scale (HADS)

Julie F. Pallant[1]* and Alan Tennant[2]

[1]Faculty of Life and Social Sciences, Swinburne University of Technology, Australia
[2]Academic Unit of Musculoskeletal and Rehabilitation Medicine, University of Leeds, UK

Objectives. To demonstrate the use of Rasch analysis by assessing the appropriateness of utilizing the Hospital Anxiety and Depression Scale (HADS) total score (HADS-14) as a measure of psychological distress.
Design. Cross-sectional, using Rasch analysis.
Methods. The HADS was administered to 296 patients attending an out-patient musculoskeletal rehabilitation program. Rasch analysis was conducted using RUMM2020 software to assess the overall fit of the model, the response scale used, individual item fit, differential item functioning (DIF) and person separation.
Results. Rasch analysis supported the viability of the HADS-14 as a measure of psychological distress. It showed good person separation, little disordering of the thresholds and no evidence of DIF. One anxiety item (item 11) showed some misfit to the model. The residuals patterned into the two subscales (anxiety and depression), but the person estimate derived from these two subscales was not statistically different to that derived from all items taken together, supporting the assumption of unidimensionality. A cut-point of 12 on the HADS-14 identified all cases that were classified as both anxious and depressed on the original individual HADS subscales.
Conclusions. The results of Rasch analysis support the use of the HADS-14 as a global measure of psychological distress. The study demonstrates the usefulness of Rasch analysis in assessing the psychometric properties of a scale and suggests that further use of this technique to assess the HADS-14 in other clinical groups is warranted.

In clinical psychology, as in many other specialties, evaluation of the outcome of an intervention is often undertaken with the help of rating scales, either administered or self-completed. Such scales are usually developed according to traditional psychometric standards of validity and reliability (Nunnally, 1978). In the former case construct

*Correspondence should be addressed to Dr Julie F. Pallant, Faculty of Life and Social Sciences, Swinburne University of Technology, PO Box 218, Hawthorn, Victoria 3122, Australia (e-mail: jpallant@swin.edu.au).

DOI:10.1348/014466506X96931

validity can be supported through factor analytic techniques, which confirm the presence of one or more valid unidimensional scales, a requirement for summating any set of Likert-style items (Svensson, 2001).

More recently, modern psychometric approaches have been adopted and, particularly, that associated with the Rasch measurement model (Rasch, 1960). In this approach, data collected from questionnaires (or for scales completed by clinical staff), which include items for a new (or existing) scale, that are intended to be summated into an overall score (which may be at the subscale or overall level) are tested against the expectations of this measurement model. The model is seen as a template that operationalizes the formal axioms which underpin measurement (Karabatos, 2001) and against which the data from the scale may be tested. In addition to testing for unidimensionality, other issues such as category ordering (do the categories of an item work as expected?) and item bias, or differential item functioning (DIF) (Holland & Wainer, 1993) may also be addressed within the framework of the Rasch model. Furthermore, where the data fit the model, a linear transformation of the raw ordinal score is obtained, opening up valid parametric approaches given appropriate distributions (Svensson, 2001; Wright & Stone, 1979). Thus, fitting data to the Rasch model offers an elegant approach to addressing several key methodological aspects associated with scale development and construct validation, as well as providing a linear transformation of the ordinal raw score. This paper illustrates this approach in the context of a suggestion made recently in this journal (Martin, Tweed, & Metcalfe, 2004), namely that it may be desirable to create a total summed score for the Hospital Anxiety and Depression Scale (HADS: Zigmond & Snaith, 1983) as an index of psychological distress.

Background

The Hospital Anxiety and Depression Scale

The HADS was developed over 20 years ago for the measurement of anxiety and depression (Zigmond & Snaith, 1983). It is a self-administered scale consisting of 14 polytomous items scored as two 7-item subscales for anxiety and depression, each of which has cut-points to identify caseness. Originally developed for use in a hospital setting, it is now widely used across all settings, including screening in normal populations (Crawford, Henry, Crombie, & Taylor, 2001). Since its appearance, many studies have reported on the construct validity in various clinical populations, however, certain issues relating to its underlying dimensionality have emerged (Bjelland, Dahl, Haug, & Neckelmann, 2002; Johnston, Pollard, & Hennessey, 2000).

Concern has been raised over the original conceptualization of the HADS as measuring two separate aspects of anxiety and depression. Some studies have identified a three-factor structure (Dunbar, Ford, Hunt, & Der, 2000; Friedman, Samuelian, Lancrenon, Even, & Chiarelly, 2001; Martin & Newell, 2004) while others have identified problems with individual items in specific clinical populations. In a number of studies, item 7 (*I can sit at ease and feel relaxed*) has failed to load on its anxiety subscale, but instead has loaded strongly with the depression items (Bjelland *et al.*, 2002; Mykleutun, Stordal, & Dahl, 2001). A number of authors have suggested that the HADS be used as a total score, summing all 14 items, representing psychological distress, rather than the separate dimensions of anxiety and depression (Martin *et al.*, 2004; Razavi, Delvaux, Farvacques, & Robaye, 1990). However, currently the weight of evidence appears to

support a multidimensional construct (see reviews by Bjelland *et al.*, 2002; Martin, 2005).

To date, the dimensionality of the HADS has only been assessed using traditional or classical test theory, adopting a range of techniques including exploratory and confirmatory factor analysis. However, in the medical literature on health outcome scale development these techniques are now being complemented and, in some cases, replaced by item response theory approaches, and particularly by the application of the Rasch measurement model (Banerji, Smith, & Dedrick, 1997; Küçükdeveci, Yavuzer, Elhan, Sonel, & Tennant, 2001; Tennant *et al.*, 2004; Van Alphen, Halfens, Hasman, & Imbos, 1994).

The Rasch model

The Rasch model was named after the Danish mathematician Georg Rasch (Rasch, 1960). The model shows what should be expected in responses to items if measurement (at the metric level) is to be achieved. For the Rasch model, dichotomous (Rasch, 1960) and polytomous (Andrich, 1978) versions are available. The response patterns achieved are tested against what is expected, a probabilistic form of Guttman scaling (Guttman, 1950), and a variety of fit statistics determine whether this is the case (Smith, 2000).

The model assumes that the probability of a given respondent affirming an item is a logistic function of the relative distance between the item location and the respondent location on a linear scale. In other words, the probability that a person will affirm an item is a logistic function of the difference between the person's level of, for example, anxiety (θ) and the level of anxiety expressed by the item (*b*), and only a function of that difference.

$$p_{ni} = \frac{e^{(\theta n - b_i)}}{1 + e^{(\theta n - b_i)}} \tag{1}$$

where p_{ni} is the probability that person n will affirm the item, θ is the person's level of anxiety, and b is the level of anxiety expressed by a positive response to the item. The formulae can be expressed as a logit model:

$$\ln\left(\frac{P_{ni}}{1 - P_{ni}}\right) = \theta_n - b_i \tag{2}$$

where ln is the normal log, P is the probability of person n affirming item i; θ is the person's level of anxiety, and b is the level of anxiety expressed by the item. Fitting data to the Rasch model thus places both item and person parameter estimates (note that they are independent parameters) on the same log-odds units (logit) scale, and it is this that gives the linear transformation of the raw score.

The model can be extended to the polytomous case and is known as the rating scale model (Andrich, 1978):

$$\ln\left(\frac{P_{nij}}{1 - P_{nij-1}}\right) = \theta_n - b_i - \tau_j \tag{3}$$

where, in addition to the parameters shown in (2) above, the τ represents the threshold (0.5 probability point) between adjacent categories. A further variant of this is known as the partial credit model (Masters, 1982), and it makes no assumptions about the

equidistance between thresholds across items, which is the case of the rating scale model:

$$\ln\left(\frac{P_{nij}}{1 - P_{nij-1}}\right) = \theta_n - b_{ij} \quad (4)$$

Statistics indicating fit to the model test how far the observed data match that expected by the model (see below). Note the orientation; because the model defines measurement, data are fitted to the model to see if they meet the model's expectations. This is opposite to the practice in statistical modelling where models are developed to best represent the data.

Within the framework of Rasch measurement, the scale should also work in the same way, irrespective of which group (e.g. gender) is being assessed (Holland & Wainer, 1993). For example, in the case of measuring anxiety, males and females should have the same probability of affirming an item (in the dichotomous case), *at the same level of anxiety*. Thus, the probability is conditioned on the trait. If for some reason one gender did not display the same probability of affirming the item (in the dichotomous case), then this item would be deemed to display DIF, and would violate the requirement of unidimensionality (Smith, 2000).

A further test for unidimensionality is undertaken by looking at patterns in the residuals. These are the standardized person-item differences between the observed data and what is expected by the model for every person's response to every item. This is one way of testing the model's assumption of local independence of items; after extracting the 'Rasch factor' there should be no further pattern in the data.

This paper examines the potential contribution of Rasch analysis in exploring a number of issues that have been raised concerning the HADS. This includes an assessment of the appropriateness of using all HADS items to represent the underlying dimension of psychological distress. In addition it will, from the perspective of the Rasch model, include an evaluation of the validity of the category scoring system, the fit of individual items and an assessment of the potential bias of items by gender.

Method

Participants

The sample consisted of 296 outpatients attending a 6-week musculoskeletal rehabilitation programme at Cedar Court Health South Hospital, a private rehabilitation hospital in Melbourne, Australia. There were 152 (51.4%) females, 140 (47.9%) males) and 4 cases (1.4%) sex unspecified. Patients ranged in age from 16 to 80 years (mean = 44.3, $SD = 12.47$). Of the patients, 55% reported pain in the lower back, 20% in an upper or lower limb, 15% in the cervical region and 10% in other locations.

Procedure

The HADS is administered to all patients on admission to the Cedar Court Health South Hospital Musculoskeletal Rehabilitation Program as part of routine clinical practice. Patients are informed, as part of admission procedures, that the data collected may be used for administrative and research purposes. Scores on the HADS from all patients attending the programme between 2001 and 2003 were extracted in a de-identified form from the medical records with permission from the Hospital Research Committee.

Rasch analysis

Data are fitted to the Rasch model using the RUMM2020 software (Andrich, Lyne, Sheridan, & Luo, 2003). The objective is to test how well the observed data fit the expectations of the measurement model. Three overall fit statistics are considered. Two are item–person interaction statistics transformed to approximate a z score, representing a standardized normal distribution. Therefore if the items and persons fit the model we would expect to see a mean of approximately zero and a standard deviation of 1. A third is an item–trait interaction statistic reported as a chi square, reflecting the property of invariance across the trait. A significant chi square indicates that the hierarchical ordering of the items varies across the trait, thus compromising the required property of invariance.

In addition to these overall summary fit statistics, individual person- and item-fit statistics are presented, both as residuals (a summation of individual person and item deviations) and as a chi squared statistic. In the former case residuals between ± 2.5 are deemed to indicate adequate fit to the model. In the latter case this misfit to the model can also be viewed graphically where observed model fit for groups of responders across the trait (called class intervals) can be plotted against the expected model curve (item characteristic curve, ICC). Items with good fit will show each of the group plots lying on the curve; those with plots which are steeper than the curve would be considered to be over-discriminating; those flatter than the curve, under-discriminating. The summed chi square within each group provides the overall chi square for the item, and the overall chi square for items is summed to give the item–trait interaction statistic. To take account of multiple testing Bonferroni corrections are applied to adjust the chi squared p value (Bland & Altman, 1995).

In addition to item fit, examination of person fit is important. From a practical perspective, a few respondents who deviate from model expectation may cause significant misfit at the item level. In terms of validation of a scale, this runs the risk of discarding the scale when it would be more appropriate to find out why a few respondents may be responding in a way different to everyone else. In medical outcome studies this could be due to unrecorded co-morbidity or, for example, respondents with cognitive deficits. Thus, where some respondents misfit in this way, removal from the analysis may make a significant difference to scale internal construct validity, while at the same time raising questions about the external construct validity of the scale with the particular patient group.

Comparison of the mean location score obtained for the persons with that of the value of zero set for the items, provides an indication of how well-targeted the items are for people in the sample. For a well-targeted measure (not too easy, not too hard) the mean location for the persons would also be around the value of zero. If a positive mean value for the persons was obtained this would indicate that the sample as a whole was located at a higher level (e.g. of psychological distress) than the average of the scale, while a negative value would suggest the opposite. Arguably, if many patients are at the margins, then the scale is not properly targeted, and this also has an influence on sample size (Linacre, 1994). Consequently, if a scale is well-targeted (i.e. 40–60% endorsement rates on dichotomous test items) then a sample size of 108 will give 99% confidence of the person estimate being within ± 0.5 logits (Linacre, 1994). If the scale is not well-targeted (i.e. <15 or $>85\%$ endorsement rate), then the sample size required for accurate estimation rises to 243. For polytomous scales additional criteria are required to ensure that responses are appropriately distributed across the

response categories, and this is usually indicated as a minimum of 10 categories per response option (Linacre, 1999).

In the current study a sample of 296 patients was available and thus the Rasch analysis can expect to have an appropriate degree of precision irrespective of the targeting of the group, or of the distribution across the response options of each item.

An estimate of the internal consistency reliability of the scale is also available, based on the Person Separation Index (PSI) where the estimates on the logit scale for each person are used to calculate reliability.

Sources of deviation from model expectation are examined to see if the scale construct can be improved. For a good fitting model we would expect that, for each of the items, respondents with high levels of the attribute being measured would endorse high scoring responses, while individuals with low levels of the attribute would consistently endorse low scoring responses. In Rasch analysis terms this would be indicated by an ordered set of response thresholds for each of the items. The term *threshold* refers to the point between two response categories where either response is equally probable. That is the point where, for example, the probability of scoring a 0 on the item or scoring a 1 is 50/50. For a given item the number of thresholds is always one less than the number of response options.

To investigate responses to an item the category probability curves can be inspected. For a well-fitting item you would expect that, across the whole range of the trait being measured, each response option would systematically take turns showing the highest probability of endorsement. One of the most common sources of item misfit concerns respondents' inconsistent use of these response options. This results in what is known as *disordered thresholds* – the failure of respondents to use the response categories in a manner consistent with the level of the trait being measured. Disordered thresholds occur when respondents have difficulty consistently discriminating between response options. This can occur when there are too many response options, or when the labelling of options is potentially confusing or open to misinterpretation (e.g. the use of terms *sometimes, often, frequently*). Usually, although not always, collapsing of categories where disordered thresholds occur improves overall fit to the model.

One other issue that can affect model fit is a form of item bias known as differential item functioning (DIF). This occurs when different groups within the sample (e.g. males and females), despite equal levels of the underlying characteristic being measured, respond in a different manner to an individual item. For example, men and women with equal levels of depression may respond systematically differently to an item in a depression inventory (see Lange, Thalbourne, Houran, & Lester, 2002). Two types of DIF may be identified. One is where the group shows a consistent systematic difference in their responses to an item, across the whole range of the attribute being measured, which is referred to as uniform DIF. When there is non-uniformity in the differences between the group (e.g. it varies across levels of the attribute) then this is referred to as non-uniform DIF. In the former case, when detected, the problem can be remedied by splitting the file by group and separately calibrating the item for each group. In the latter case there is little that can be done to correct the problem, and it is often necessary to remove the item from the scale.

In RUMM the presence of DIF can be detected both statistically and graphically. Analysis of variance is conducted for each item comparing scores across each level of the person factor (in this case, gender) and across different levels of trait (referred to as class intervals). Uniform DIF is indicated by a significant main effect for the person

factor (gender), while the presence of non-uniform DIF is indicated by a significant interaction effect (person factor × class interval).

Finally, when issues of threshold disordering, DIF and fit have been resolved, it is usual to undertake a principal components analysis (PCA) of the residuals, to detect any signs of multidimensionality. The absence of any meaningful pattern in the residuals will be deemed to support the assumption of local independence and consequently the unidimensionality of the scale (Smith, 2002). This is formally tested by allowing the factor loadings on the first residual to determine subsets of items and then testing, by a paired *t* test, to see if the person estimate (the logit of person 'ability' or, in this case 'psychological distress') derived from these subsets significantly differs from that derived from all items (Smith, 2002). If the person estimate is found to differ between the subset and the full scale this would indicate a breach of the assumption of local independence.

Results

Individual anxiety and depression scales

Briefly, for comparative purposes, we present the results of fitting data from the traditional anxiety and depression subscales to the Rasch model. The residual mean value for items in the anxiety subscale is .04 with a standard deviation (*SD*) of 1.87, whereas the latter would be expected to be much closer to 1, given adequate fit to the model. This deviation is supported by a significant chi squared interaction of 59.70 ($df = 28$) with $p = .00044$, showing lack of invariance of item difficulty across the scale. Thus the scale fails to fit the Rasch model, and the principal reason for this is the item *I can sit at ease and feel relaxed* (anxiety 7), which shows significant misfit to the model.

The depression subscale fares slightly better. The residual mean value for items in the depression subscale is −0.11 with a *SD* of 1.43, much closer to the expected value of 1 than was the case with the anxiety subscale. The chi squared interaction of 56.2 ($df = 28$) with $p = .001$ shows slight invariance of item difficulty across the scale, but no individual item misfits model expectation.

Consequently, there is some concern about the construct validity of the existing scales and particularly so with the anxiety subscale, where serious misfit to the model expectation occurs. How far does this misfit manifest when we consider a 14-item scale of psychological distress?

Overall fit of the 14-item scale of psychological distress

Initial inspection of the fit of the data from all 14 items to the Rasch model shows a significant item–trait interaction total chi-square, suggesting that there is some degree of misfit between the data and the model. This could be caused by misfit to model expectations of respondents or items, or both. The residual mean value for items was .09 with a *SD* of 1.53, much higher than the expected value of 1, given adequate fit to the model. This deviation is supported by a significant chi squared interaction of 95.78 ($df = 56$) with $p = .00074$, showing lack of invariance of item difficulty across the scale. The residual mean value for persons was −0.22 with a *SD* of 1.19, indicating no serious misfit among the respondents in the sample.

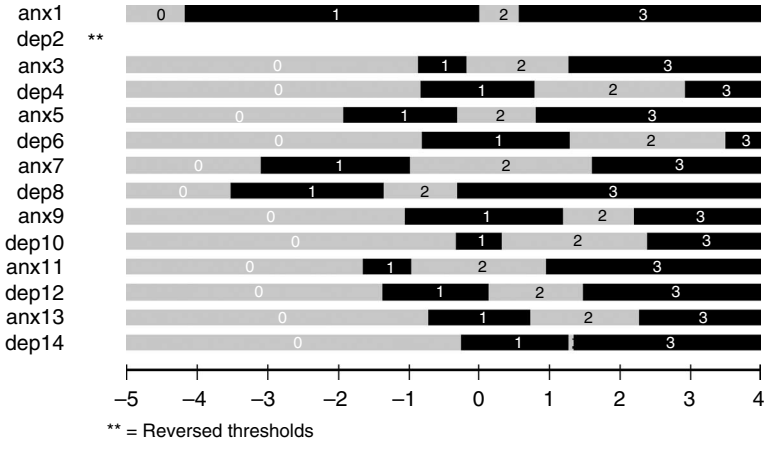

Figure 1. Threshold ordering.

Thresholds

Initially, the pattern of thresholds is examined to see if disordering maybe affecting fit. In the current example only one disordered threshold (item dep2: *I still enjoy the things I used to enjoy*) was detected (Fig. 1). Inspection of this figure shows that the threshold distances vary across items, supporting the use of the partial credit model for the analysis of this scale.

The ordering of thresholds is graphically demonstrated in the category probability curves shown in Figs 2 and 3. Figure 2 shows clearly how item thresholds for dep6 (*I feel cheerful*) are properly ordered, where each response category (0,1,2,3) systematically has a point along the ability continuum where it is the most likely response, as indicated by a peak in the curve. In contrast, Fig. 3 for item dep2 shows disordered thresholds. The point at which the lines for adjacent response categories cross in dep2 indicates that the transition between categories 2 and 3 is lower on the trait (lower psychological distress) than between categories 1 and 2, which is not how the variable is intended to work.

Consequently, scores for dep2 were recoded by collapsing the responses to the second (scored 1) and third (scored 2) response category to form three, rather than four

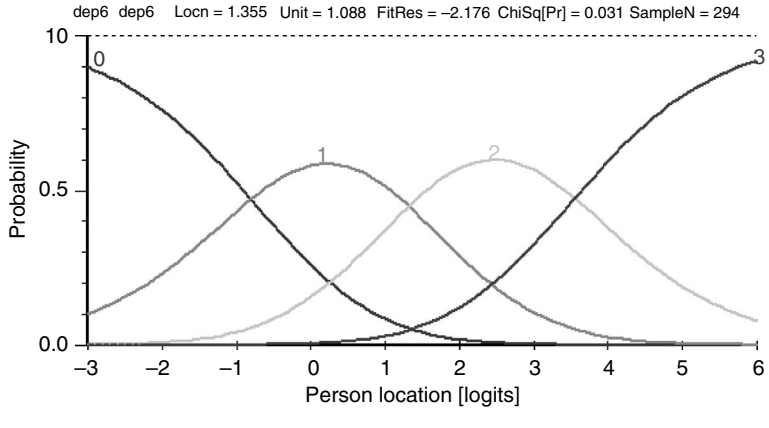

Figure 2. Category probability curve for item dep6 showing ordered thresholds.

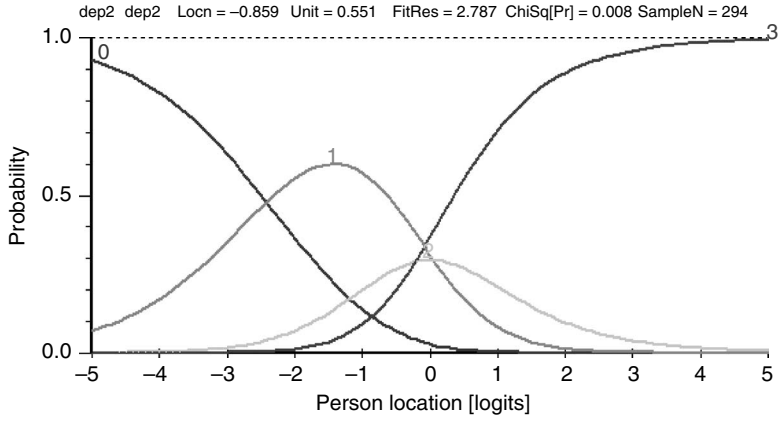

Figure 3. Category probability curve for item dep2 showing disordered thresholds.

categories (coded 0112). This resulted in an improvement in the overall model fit, with a change in the item–trait interaction probability value from .0007 to .0038. The PSI remained at .89. Alternative recoding procedures were also checked; however, no other solution improved the overall fit of the model.

Individual item fit

Following recoding of item dep2, the fit of the individual items was checked revealing two items (dep2: *I feel cheerful* and anx11: *I feel restless as if I have to be on the move*) showing misfit to model expectation (Table 1). Both items showed fit residual values above 2.5, and the probability value for item anx11 is less than the Bonferroni adjusted α value of .004, indicating significant deviation from the model.

The positive fit residual values obtained for these two items suggest low levels of discrimination. This is confirmed by inspection of the ICC for item anx11. The plot of observed group responses deviates from the model curve, and this observed response is flatter than the ICC, showing under-discrimination (Fig. 4). Thus responses from the lowest group (low levels of anxiety/depression) are above what is expected by the model and those for the highest group (high levels of anxiety/depression), are below model expectation.

Person fit

Individual person fit statistics showed that nine respondents had residuals outside the acceptable range. On removal of these persons, the chi squared interaction statistic improved further to $p = .011$; with the PSI remaining high at 0.892. However, at the individual item level item Anx11 still remained problematic with a chi squared fit p value of .002. Its removal, however, did not improve the overall fit of the scale, so it was decided to retain the item.

Differential item functioning

In the current data the possibility of gender differences in response to the HADS items was explored by analysis of DIF with a Bonferroni-adjusted p value of .002 (.05/28). None of the items showed probability values exceeding the adjusted alpha value

Table 1. Fit of the HADS items to the Rasch model after rescoring of item dep2

Item	Location	SE	Fit residual	DF	ChiSq	DF	Prob
anx1: I feel tense or 'wound' up	−1.197	0.089	−0.091	270.07	8.274	4	0.082
dep2: I still enjoy the things I used to enjoy	**−0.859**	**0.076**	**2.787**	**270.07**	**13.684**	**4**	**0.008**
anx3: I get a sort of frightened feeling as if something awful is about to happen	0.075	0.077	−0.142	270.07	7.701	4	0.103
dep4: I Can laugh and see the funny side of things	0.97	0.09	−1.408	270.07	6.357	4	0.174
anx5: Worrying thoughts go through my mind	−0.468	0.079	−1.332	270.07	10.029	4	0.039
dep6: I feel cheerful	1.331	0.097	−2.125	270.07	10.517	4	0.032
anx7: I can sit at ease and feel relaxed	−0.817	0.093	0.829	270.07	3.447	4	0.486
dep8: I feel as if I am slowed down	−1.716	0.085	1.179	270.07	3.313	4	0.506
anx9: I get a sort of frightened feeling like 'butterflies' in the stomach	0.785	0.091	−0.28	270.07	0.885	4	0.926
dep10: I have lost interest in my appearance	0.796	0.084	0.098	270.07	1.298	4	0.861
anx11: I feel restless as if I have to be on the move	**−0.549**	**0.078**	**3.168**	**270.07**	**17.365**	**4**	**0.001**
dep12: I look forward with enjoyment to things	0.092	0.081	−1.227	270.07	4.243	4	0.374
anx13: I get sudden feelings of panic	0.764	0.087	−0.715	270.07	2.946	4	0.566
dep14: I can enjoy a good book or radio or TV programme	0.794	0.085	0.514	270.07	5.724	4	0.221

Note. Misfitting items are bolded.

(Table 2). However, item anx13 (*I get sudden feelings of panic*) showed some degree of uniform DIF ($p = .009$) as shown in Fig. 5.

Inspection of the graph in Fig. 5 suggests that at equal levels of the overall attribute (anxiety/depression) males are slightly less likely than females to endorse this item. This difference is only noted for the four lowest categories (reading from the left), with no sex difference noted to this item in respondents in the highest category (representing high levels of anxiety/depression). If the difference had been more pronounced (and statistically significant) this item would have been calibrated separately for males and females, treating it as two separate scale items for purposes of providing an unbiased linear estimate of a person's level of psychological distress.

Targeting and reliability

It is important, particularly in clinical practice, that the measures used are appropriately targeted at the population being assessed. Poorly targeted measures often result in floor or ceiling effects. Figure 6 shows the distributions of persons (top half of the graph) and item thresholds (bottom half of the graph) for the HADS total score. The average mean person location value of −0.468 suggests that on the whole the scale was reasonably

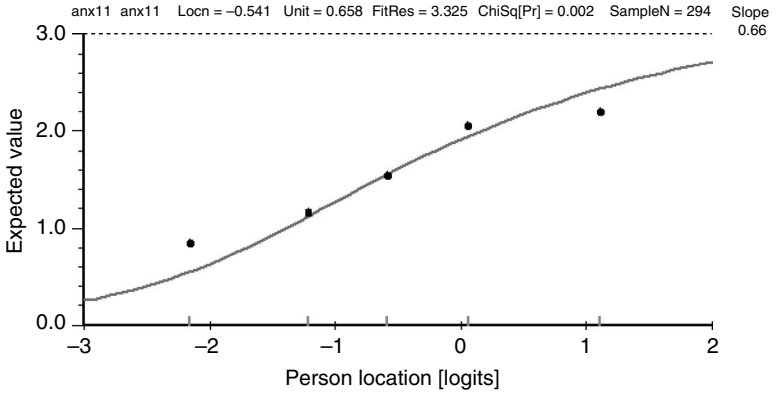

Figure 4. Item characteristic curve for item Anx11.

well-targeted for use with this group, with patients on average at a slightly lower level of psychological distress than the average of the scale items (which would be zero logits). The thresholds positioned at the extreme left of the graph are those easiest to endorse. In this case the easiest threshold is for the transition between the first response point (*not at all*) and the second response point (*sometimes*) for item dep8, *I feel as if I am slowed down* (not shown). In other words, the transition between these thresholds is the most likely to occur. At the other extreme, the hardest item to endorse is the third threshold (*most of the time*) for the item dep6, *I feel cheerful*, which is the transition least likely to occur.

With respect to reliability, the PSI statistic was 0.893, which indicates that the Total HADS also has good person separation reliability.

Table 2. Uniform and non-uniform DIF statistics for all HADS items

	Uniform DIF				Non-uniform DIF			
Item	MS	F	DF	Prob	MS	F	DF	Prob
anx1	3.99546	4.46306	1	0.035539	0.70103	0.78307	4	0.536988
dep2	0.21566	0.2156	1	0.64278	1.91853	1.91803	4	0.107615
anx3	0.09402	0.10405	1	0.747258	0.39585	0.4381	4	0.781036
dep4	0.97821	1.20968	1	0.272358	0.68714	0.84974	4	0.494746
anx5	0.22571	0.28736	1	0.592349	1.1682	1.4873	4	0.206198
dep6	0.0565	0.07849	1	0.779569	2.0754	2.88324	4	0.023007
anx7	4.10239	4.04715	1	0.045216	1.28907	1.27171	4	0.281334
dep8	0.32676	0.31307	1	0.576258	3.07251	2.94377	4	0.020836
anx9	1.69945	1.89262	1	0.170022	1.20786	1.34515	4	0.253421
dep10	0.24937	0.25181	1	0.616209	0.55706	0.56249	4	0.69008
anx11	1.4871	1.28592	1	0.257786	2.40521	2.07981	4	0.083653
dep12	0.00036	0.00042	1	0.983545	0.05385	0.06288	4	0.99268
anx13	5.67214	6.84384	1	0.009384	1.46133	1.76319	4	0.136467
dep14	2.39503	2.38467	1	0.12368	0.69173	0.68874	4	0.600285

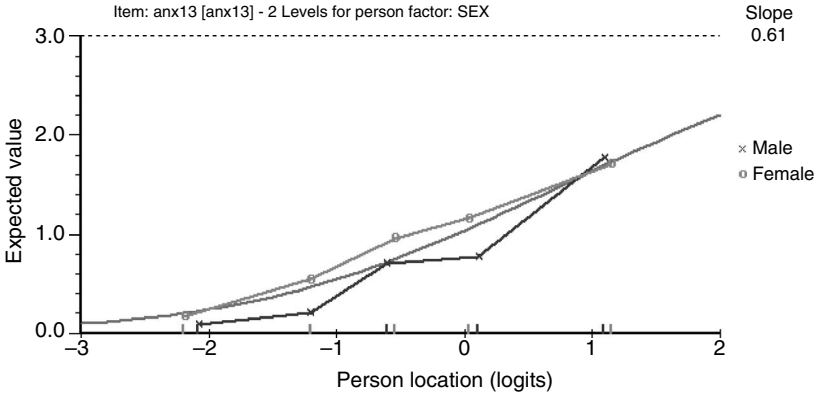

Figure 5. Differential item functioning graph of males and females for item Anx13.

Test of local independence assumption

Analysis of the pattern of residuals showed that the residuals loaded in opposite directions on the two original subscales, although Anx7 also loaded with the depression items (Table 3). These two subsets of items (defined by positive and negative loadings on the first residual component) were then separately fitted to the Rasch model and the person estimates obtained. The differences in person estimates derived from these analyses were trivial, with the difference between the depression subset and the overall scale being 0.01 logit, and between the anxiety subset and the overall scale 0.03 logit. Consequently, neither subset showed a significant difference (using a Bonferroni-adjusted p value of .025: .05/2) of person estimate to the full 14-item scale (paired t test: $p > .025$), supporting a unidimensional construct.

Determining cut-points for Total HADS

The results of the preceding Rasch analysis suggest that there is support for the use of a total 14-item HADS scale for assessing psychological distress. The raw score-logit graph for the HADS-14 (with an upper score of 41 given Dep2 was rescored) is shown in Fig. 7. To determine optimal cut-points on the Total HADS scores for use in clinical practice, an

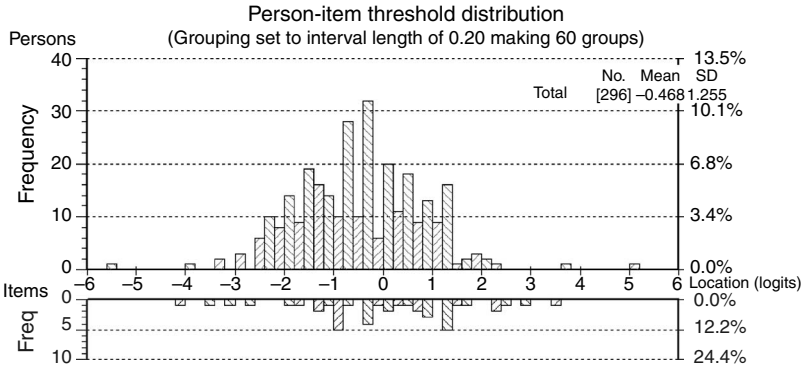

Figure 6. Person item distribution graph for Total HADS Psychological Distress Scale.

Table 3. Principal component analysis of the residuals showing first component loading

Item	Loading
anx1	0.121
dep2	**−0.533**
anx3	0.573
dep4	**−0.336**
anx5	0.249
dep6	**−0.193**
anx7	**−0.417**
dep8	**−0.329**
anx9	0.662
dep10	**−0.206**
anx11	0.267
dep12	**−0.530**
anx13	0.680
dep14	**−0.190**

Note. Negative loadings are in bold.

examination was made of the extent of agreement between the HADS-14 and the original clinical case values of the respective anxiety and depression subscales. This allowed each person to be identified as not anxious or depressed (score of less than 8 on the HADS anxiety and depression subscale), anxious only (score of 8 or more on HADS anxiety), depressed only (score of 8 or more on HADS depression) and both anxious and depressed (score of eight or more on both HADS subscales).

Using a cut point of 12 on the HADS-14 all cases that were classified as both anxious and depressed on the original HADS subscales were accurately detected. Only eight cases were identified as false positives (a score of 12+ on the Total HADS, but scores below 8 on both of the original HADS subscales) (Table 4). Figure 8 shows the distribution of scores on the Total HADS for each of the four groups classified using the original HADS subscales. The cut-point of 12 (shown as a horizontal line on the graph) clearly separates cases with no evidence of anxiety or depression from those with either anxiety or depression, or both. A Kruskal-Wallis test showed a significant difference among the groups ($\chi^2 = 237.36$, $df = 3$, $p < .001$).

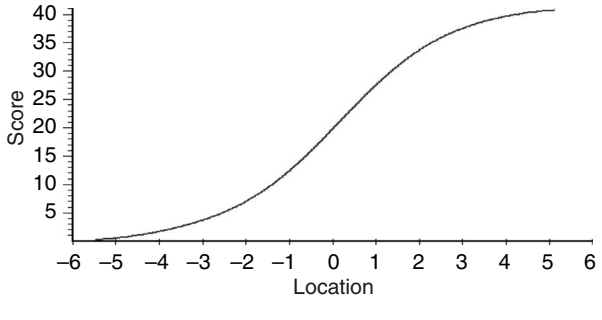

Figure 7. Raw score-logit graph of the HADS-14.

Table 4. Breakdown of cases using a cut-point of 12 on Total HADS for original HADS subscale classification groups

	Total HADS score		
Classification group using original HADS subscales	0–11	12+	Total
Not anxious or depressed[a]			
Count	82	8	90
%	91.1%	8.9%	100.0%
Anxious only[b]			
Count	5	49	54
%	9.3%	90.7%	100.0%
Depressed only[c]			
Count	2	23	25
%	8.0%	92.0%	100.0%
Both anxious and depressed[d]			
Count	0	127	127
%	0%	100.0%	100.0%
Total			
Count	89	207	296
%	30.1%	69.9%	100.0%

[a] Cases with scores below 8 on both HADS anxiety and depression subscales.
[b] Cases with scores 8 or above on HADS anxiety, but below 8 on depression subscale.
[c] Cases with scores 8 or above on HADS depression, but below 8 on anxiety subscale.
[d] Cases with scores 8 or above on both HADS anxiety and depression.

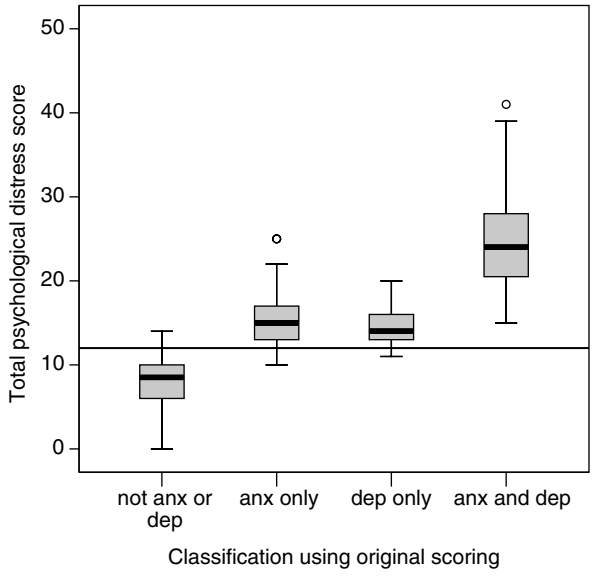

Figure 8. Boxplot of Total HADS scores for original HADS classification groups showing suggested cut-point of 12. The horizontal line represents the suggested cut-point of 12 on the Total HADS.

Discussion

Application of the Rasch measurement model in this study has supported the viability of a total 14-item HAD scale (HAD-14) for measuring psychological distress. The scale shows high reliability, with little disordering of thresholds and no evidence of differential item functioning. There is, however, some concern over item *Anxiety 11* which continues to misfit, notwithstanding the adequate fit overall. The residuals patterned into the two subscales, but the person estimate derived from these two subscales did not deviate from that derived from all items taken together. Given also that a high PSI is indicative of the power of the test of fit, then there is good evidence from this sample that a single total score of psychological distress is viable. Thus, the 14-item scale appears robust when tested against the strict assumptions of the Rasch measurement model. These results would appear to support the recommendations of a number of researchers utilizing factor analytic techniques with clinical groups. Martin *et al.* (2004) investigated the use of the HADS in patients with end-stage renal disease concluding that the 'HADS may be a too generic a measure to accurately and robustly assess independent but related domains of anxiety and depression' (p. 61). In an earlier study on cancer patients Razavi *et al.* (1990) also assessed a full-scale HADS score to screen for adjustment disorders and depressive disorders in cancer patients concluding that it was a 'simple, sensitive and specific tool' (p. 79).

Nevertheless, the issue of dimensionality of the HADS remains problematic, not least because of recent evidence which argues for a three-factor structure for the scale (Martin, 2005). How is it possible that factor analysis and Rasch analysis can apparently give different results? One issue may be the conceptual level of the construct being tested. When we briefly examined the fit of data at the subscale level, we found results similar to those reported elsewhere. The item anxiety 7 *I can sit at ease and feel relaxed* fails to fit the model when just the anxiety items are considered. This is a commonly reported problem from traditional factor analytic studies. Bjelland *et al.* (2002) consistently found, in the 19 studies they reviewed, that this item had relatively low loadings on the anxiety factor (less than .6) and relatively high loadings on the depression factor (greater than .45). It is the anxiety subscale which appears to split into two, with the depression subscale remaining the same (Martin, 2005). Again, our analysis supported the validity of the depression scale.

In this study what we were testing was the presence of a higher-order construct, namely psychological distress. Can all the items be taken together to provide a valid measure of this construct? This idea is not new, and both classical and modern psychometric theories have developed approaches to test the validity of such scales. In classical test theory a second-order factor analytic approach has been used to confirm higher-order dimensions, although its utility with just two or three factors is unclear (Baldwin *et al.*, 2005). In modern test theory, most recently, multidimensional Rasch models have been proposed, which also test for higher-order measurement constructed from valid subscales, although this approach has yet to gain widespread use (Briggs & Wilson, 2003). Thus proposing and evaluating a higher-order construct is not inconsistent with evidence showing a valid subscale structure. The real challenge to understanding dimensionality is where confirmatory factor analysis conducted on a subset of the current datafile (Pallant & Bailey, 2005) has demonstrated that a single-factor structure is inappropriate, whereas the Rasch analysis of the HADS in this study supports such a single structure. The rigorous test of the local independence assumption applied in the Rasch analysis above is equivalent to the lack of correlation of

residuals in confirmatory factor analysis. How can they give a different view of dimensionality? These differences remain unresolved, but one potential influence may be the validity of using ordinal data in a parametric procedure such as factor analysis, the assumptions of which require interval-level data (Tabachnick & Fidell, 2001). In contrast, Rasch analysis is designed for binary or ordinal data and makes no distributional assumptions. Further research, particularly mathematical simulation studies, is needed to address this issue in more depth.

Using Rasch analysis has enabled a detailed examination of the structure and operation of the HADS scale. The ordering of categories (threshold ordering) has not been examined previously, and the evidence from this work again supports a scale that is working well. The evidence that the scale appears free of DIF for gender provides important additional material to support construct validity. A cut-point of 12 on the HADS-14 for probable psychological distress seems to be able to identify those cases with an original classification of probable anxiety or depression. Thus, the Total HADS may prove useful for initial screening, avoiding the dimensionality problems inherent in the two-subscale approach. It will also have potential value in applications such as computer adaptive testing (Revicki & Cella, 1997) where the greater number of thresholds (41 in the version presented above) will give greater precision.

Some issues remain concerning the psychometric properties of the HADS. Construct validity is a continuing quest, and further evidence should be provided for the validity of the HADS-14 in different diagnostic groups. Also, DIF by age and other relevant clinical subgroups should be evaluated, particularly when diagnostic or repeated measures are pooled. Recoding for disordered thresholds remains problematic for existing scales, particularly when, as in this case, it is just one item displaying this problem. Recoding item Dep2 improved the fit of the data to the model, and thus provides a strong case for scoring the item as 0112, reducing the overall maximum score from 42 to 41. Nevertheless, it is recognized that a long history of use, as well as investment in computerized administration procedures may make the adoption of this recoding difficult.

In summary, it would appear that the HADS-14 is a viable scale for the measurement of psychological distress in patients attending musculoskeletal rehabilitation. The results of this study suggest that further use of Rasch analysis of the HADS on other clinical groups is warranted. Ideally, these studies would utilize data pooled for patients with a variety of physical conditions to allow further assessment of the presence of DIF across clinical groups.

References

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

Andrich, D., Lyne, A., Sheridan, B., & Luo, G. (2003). *RUMM 2020*. Perth: RUMM Laboratory.

Baldwin, K. A., Grinslade, M. S., Baer, L. C., Watts, P., Dinger, M. K., & McCubbin, J. (2005). Higher order factor analysis of an instrument with dichotomous data. *Research in Nursing and Health*, *5*, 431–440.

Banerji, M., Smith, R. M., & Dedrick, R. F. (1997). Dimensionality of an early childhood scale using Rasch analysis and confirmatory factor analysis. *Journal of Outcome Measurement*, *1*, 56–85.

Bjelland, I., Dahl, A. A., Haug, T. T., & Neckelmann, D. (2002). The validity of the Hospital Anxiety and Depression scale: An updated review. *Journal of Psychosomatic Research*, *52*, 69–77.

Bland, J. M., & Altman, D. G. (1995). Multiple significance tests: The Bonferroni method. *British Medical Journal*, *310*, 170.

Briggs, D. C., & Wilson, M. (2003). An introduction to multidimensional measurement using Rasch models. *Journal of Applied Measurement*, *4*, 87–100.

Crawford, J. R., Henry, J. D., Crombie, C., & Taylor, E. P. (2001). Normative data for the HADS from a large non-clinical sample. *British Journal of Clinical Psychology*, *40*, 429–434.

Dunbar, M., Ford, G., Hunt, K., & Der, G. (2000). A confirmatory factor analysis of the Hospital Anxiety and Depression scale: Comparing empirically and theoretically derived structures. *British Journal of Clinical Psychology*, *39*, 79–94.

Friedman, S., Samuelian, J. C., Lancrenon, S., Even, C., & Chiarelly, P. (2001). Three-dimensional structure of the Hospital Anxiety and Depression Scale in a large French primary care population suffering from major depression. *Psychiatry Research*, *104*, 247–257.

Guttman, L. A. (1950). The basis for Scalogram analysis. In S. A. Stouffer, L. A. Guttman, F. A. Suchman, P. F. Lazarsfeld, S. A. Star & J. A. Clausen (Eds.), *Studies in social psychology in World War II: Vol 4. Measurement and prediction* (pp. 60–90). Princeton: Princeton University Press.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hilldale, NJ: Lawrence Erlbaum Associates.

Johnston, M., Pollard, B., & Hennessey, P. (2000). Construct validation of the Hospital Anxiety and Depression scale with clinical populations. *Journal of Psychosomatic Research*, *48*, 579–584.

Karabatos, G. (2001). The Rasch model, additive conjoint measurement, and new models of probabilistic measurement theory. *Journal of Applied Measurement*, *2*, 389–423.

Küçükdeveci, A. A., Yavuzer, G., Elhan, A. H., Sonel, B., & Tennant, A. (2001). Adaptation of the Functional Independence Measure for use in Turkey. *Clinical Rehabilitation*, *15*, 311–319.

Lange, R., Thalbourne, M. A., Houran, J., & Lester, D. (2002). Depressive response sets due to gender and culture based differential item functioning. *Personality and Individual Differences*, *33*, 937–952.

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions*, *7*, 28.

Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement*, *3*, 103–122.

Martin, C. R. (2005). What does the Hospital Anxiety and Depression Scale (HADS) really measure in liaison psychiatry settings? *Current Psychiatry Reviews*, *1*, 69–73.

Martin, C. R., & Newell, R. J. (2004). Factor structure of the Hospital Anxiety and Depression Scale in individuals with facial disfigurement. *Psychology, Health and Medicine*, *9*, 327–336.

Martin, C. R., Tweed, A. E., & Metcalfe, M. S. (2004). A psychometric evaluation of the Hospital Anxiety and Depression Scale in patients diagnosed with end-stage renal disease. *British Journal of Clinical Psychology*, *43*, 51–64.

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

Mykleutun, A., Stordal, E., & Dahl, A. A. (2001). Hospital Anxiety and Depression (HAD) scale: Factor structure, item analysis and internal consistency in a large population. *British Journal of Psychiatry*, *179*, 540–544.

Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw-Hill.

Pallant, J. F., & Bailey, C. M. (2005). Assessment of the structure of the Hospital Anxiety and Depression Scale in musculoskeletal patients. *Health and Quality of Life Outcomes*, *3*, 82.

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

Razavi, D., Delvaux, N., Farvacques, C., & Robaye, E. (1990). Screening for adjustment disorders and major depressive disorders in cancer in-patients. *British Journal of Psychiatry*, *156*, 79–83.

Revicki, D. A., & Cella, D. F. (1997). Health status assessment for the twenty-first century: Item response theory, item banking and computer adaptive testing. *Quality of Life Research*, *6*, 595–600.

Smith, E. V. (2002). Detecting and evaluation the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement*, *3*, 205–231.

Smith, R. M. (2000). Fit analysis in latent trait measurement models. *Journal of Applied Measurement*, *2*, 199–218.

Svensson, E. (2001). Guidelines to statistical evaluation of data from rating scales and questionnaires. *Journal of Rehabilitation Medicine*, *33*, 47–48.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using Multivariate Statistics*. Needham Heights, MA: Allyn & Bacon.

Tennant, A., Penta, M., Tesio, L., Grimby, G., Thonnard, J. -L., Slade, A., *et al.* (2004). Assessing and adjusting for cross cultural validity of impairment and activity limitation scales through Differential Item Functioning within the framework of the Rasch model: The Pro-ESOR project. *Medical Care*, *42*(Suppl 1), 37–48.

Van Alphen, A., Halfens, R., Hasman, A., & Imbos, T. (1994). Likert or Rasch? Nothing is more applicable than a good theory. *Journal of Advanced Nursing*, *20*, 196–201.

Wright, B. D., & Stone, G. (1979). *Best test design*. Chicago: MESA Press.

Zigmond, A. S., & Snaith, R. P. (1983). The Hospital Anxiety and Depression Scale. *Acta Psychiatrica Scandinavica*, *67*, 361–370.

Received 7 October 2005; Received in revised form 29 December 2005