Kersten, P. and N. M. Kayes (2011). "Outcome measurement and the use of Rasch analysis, a statistics-free introduction." New Zealand Journal of Physiotherapy **39**(2): 92-99.

Abstract

Outcome measures, which use questions or assess performance on tasks are increasingly used in clinical practice. A key principle of such measures is that of internal construct validity. This is a characteristic that is best evaluated with Rasch analysis. This approach to analysis tends to be described in the literature in a statistical and technical manner, not easily accessible to people new to measurement research. This paper focuses on concepts and interpretation of key messages in an attempt to de-mystify Rasch analysis for the practicing clinician. The paper first explains the basic tenets of the Rasch model. This is followed by explanation of the principles of the key analytical stages involved in a Rasch analysis. The paper demonstrates that the examination of internal construct validity, using Rasch analysis, involves various qualitative and quantitative judgments. The main strength of the Rasch model lies in its theoretical and scientific underpinning. Outcome measures that fit the Rasch model are more robust than those that don't and arguably, the latter should not be used in practice or research.

Key words

Outcome measurement

Rasch analysis

Introduction

Outcome measurement is not only part of physiotherapy standards in many countries such as the UK and New Zealand (Chartered Society of Physiotherapy 2005, The Physiotherapy Board of New Zealand 2009), but now also a requirement from funders of services (ACC 2009, Department of Health 2008). This move towards more standardised measurement has been justified and promoted as arguably it allows the use of more reliable and valid data. Such data assists in "diagnosing'' the presence and severity of patients' problems, communication with patients and the team, treatment and discharge planning, the evaluation and improvement of processes of care or treatment, benchmarking against other services and informing funding priorities and health policies (ACC 2009, Chartered Society of 2011, Kayes and McPherson 2010, Laver Fawcett 2007, Lord Darzi of Denham 2008, Tyson et al 2010). However, there are also barriers to the use of outcome measures, such as how to choose between, or combine objective measures and those measuring more subjective attributes (e.g. pain, quality of life), questions whether one can measure patients' perceptions and attitudes, patients' literacy and ability to concentrate on or complete a measure, cultural appropriateness of measures in a multi-cultural society, difficulty using or interpreting outcomes data, the burden and costs of some measures, and issues concerning financial compensation (Horner and Larmer 2006, Kayes and McPherson 2010, Laver Fawcett 2007, Turner-Stokes and Rusconi 2003, Tyson et al 2010, Van Peppen et al 2008). Added to these complexities, many therapists find themselves wading through the literature in an attempt to select the most appropriate or the best outcome measure and find the literature overwhelming or not easily accessible in terms of being too technical or giving

unclear or conflicting answers. This is not helped by arguments within the literature as to which statistical or mathematical approaches should be used to develop and evaluate outcome measures. Once therapists have selected a measure or use those recommended by funders or researchers, they are left to implement them, input the data and interpret these. It is no surprise therefore to find some providers use outcome measures haphazardly or leave data in boxes as they are uncertain what to do with them. Further, there are instances when measures are being used that have not been thoroughly validated or which are inappropriate to measure the anticipated outcomes. The risks inherent in this approach are that in selecting the wrong measure (for example, one that is not reliable or not responsive), it may look like the service provided is not effective or the opposite, resulting in the possibility of allocating resources based on erroneous information.

This paper aims to provide a very specific focus on a principle of a good outcome measure, i.e. internal construct validity. This is a characteristic that is best evaluated with Rasch analysis. Here, we will demonstrate the answers that can be learned from Rasch analysis. Unlike much existing literature on this topic, this paper deliberately avoids a focus on statistics but rather focuses on concepts and interpretation of key messages in an attempt to make Rasch analysis accessible to the practicing clinician.

Key principles of a good outcome measure

Before we go on to consider what characteristics one might look for in a good outcome measure it seems appropriate to first clarify what outcome measurement intends to

achieve and introduce the different types of data they may yield. Outcome measures intend to provide quantification of a trait, which cannot be directly observed, also known as a latent trait or construct. Essentially all outcome measures aim to capture something (e.g. blood pressure, height, pain, mobility or depression) quantitatively. Different types of data are generated from outcome measures. For example, the measurement of distance with a distance measurement wheel produces ratio data: units of measurement which are equally spaced and where there is a true zero to the scale (so if someone walks 2 kilometres s/he has walked twice as far as someone who has walked 1 kilometre). Interval data is similar to ratio data, the only difference being that an interval scale does not have a true zero (e.g. measurement of temperature). Other outcome measures, which we increasingly see in clinical practice, include clinician or patient reported outcomes measures of, for example, pain, function and disability. These outcome measures typically consist of a range of items (specific questions or tasks). For example, the Barthel Index measures the level of dependency of a person through the assessment of performance on 10 tasks. Responses to items are constructed in a hierarchical order (e.g. ‘unable to do’, ‘requiring help from two people’, ‘requiring help from one person’, ‘independent’). These response options are assigned numerical values (in this example 0, 1, 2, 3). These types of data are known as ordinal data. That is, the scores may decrease or increase but they are not evenly distributed as is the case with interval/ratio data. After all, it would be silly to conclude that someone who scores 2 on an item in the Barthel Index (reflecting ‘requiring help from one person’) is twice as independent as someone who scores 1 on this question (reflecting ‘requiring

help from two people'). Yet, in practice, scores from items with ordinal data are frequently summed, as if this is completely appropriate (we'll come back to this later).

There are many excellent and accessible text books and papers, which describe the key principles of a good outcome measure (Enderby et al 2006, Hobart and Cano 2009, Laver Fawcett 2007, McDowell 2006, Streiner and Norman 2008). For starters there are a number of important things to consider when choosing a measure including the underpinning philosophical basis of the measure, whether it is fit for purpose, its feasibility and acceptability, administration cost, burden on the patient and so forth (Holmbeck and Devine 2009, Laver Fawcett 2007). Further, there is consensus that a good outcome measure should a) measure what it purports to measure (validity), b) include items that tap into the same construct (homogeneity), c) provide the same data when scored by different therapists at the same time (inter-rater reliability) or the same therapist at different time points in which no real change has occurred (intra-rater reliability), d) result in the same scores if patients complete the measure themselves at different time points (between which no real change has occurred) (test-retest reliability) and e) measure clinically meaningful change over time (when real change has occurred) (responsiveness). Importantly, a measure must have all these characteristics since a very reliable measure that is not valid provides data that isn't meaningful (albeit very accurately), a measure that is valid but not reliable provides data which is full of measurement error, and a measure which is not responsive can provide a snapshot of a patient's condition but not evaluate whether the therapy has been instrumental in achieving the desired therapeutic effect. A recent paper in this journal discussed these

different characteristics and provided an overview of the interpretation of various associated statistics (Horner and Larmer 2006). Essentially, many of these techniques draw on the statistical concepts of associations (correlation coefficients), repeated measures analysis of variance and factor analysis (Streiner and Norman 2008, p3).
In practice, once a measure has been shown to have these key properties, the scores on each item are summed to give a total score. However, as already intimated above, there are inherent problems with this approach when using ordinal data, which are largely ignored. Let's take the previous example of the Barthel Index (Mahoney and Barthel 1965). Items on this scale are summed to provide a total score ranging from 0 to 20 (where 0 is completely dependent and 20 independent). However, two patients with the same total score can have very different underlying impairments and consequent activity limitations. In addition, summing items makes two assumptions a) that the scale produces interval level data (which more often than not is not the case since most clinician or patient reported outcomes measures produce ordinal data); and b) that the scale is unidimensional (i.e. that it focuses on only one attribute or dimension). Both of these assumptions may in some cases be incorrect. For example, if a scale truly produces interval data then it would follow that a one-point change on one item would correspond to the same amount of change in another item. A one-point change on the Barthel Index could reflect an improvement in transfers (from no sitting balance to major help to sit), an improvement in continence (from incontinence to once a week accidents) or in climbing stairs (from being unable to climbing stairs to needing help), or improvements in the other seven items. Clinically, it is easy to see that these changes in a patient's status are not equivalent, yet in practice the scores are summed into a total

as if it doesn't matter. Likewise, unidimensionality assumes that every item in a scale measures the same construct. While the Barthel Index may well be a good example of a scale that aims to be unidimensional in that it purports to measure a single construct (dependence), there are a number of health outcomes of interest in rehabilitation which are complex and multidimensional. For example, health status scales invariably incorporate questions relating to both physical and mental health status, which in a Western model of health constitute two independent constructs. It therefore doesn't make sense to add these scores together to yield a single health status score. Traditional ways of exploring the properties of an outcome measure (classical test theory) don't address these issues well. A more modern approach to measurement, that of Rasch analysis, devotes great attention to these particular issues. Here we will explore the benefits of this approach.

The Rasch model

Rasch Analysis is based on the Rasch model, in which the total score summarises completely how much of a construct the person has (e.g. how much pain, how many mobility problems). The fundamental requirement of the Rasch model is that the comparison of two people is independent of which items from the total set of items in the scale they completed or were scored on. Therefore, the Rasch model expects items, sets of items and their responses to meet certain expectations. During Rasch analysis a series of tests are carried out to assess if data produced from an outcome measure fit this Rasch model. By contrast, classical test theory only provides a statistical description of the responses. Why this is important might make intuitive sense when we

examine a hypothetical example of walking ability of children with cerebral palsy. A measure evaluating this should include items which evaluate poor performance on walking ability (e.g. able to walk indoors unaided), moderate performance on walking ability (e.g. able to walk 10 meters outdoors unaided), and high performance on walking ability (e.g. able to run in the playground). Let's assume these items have a no/yes (0/1) response option. One would then assume that a child with very good walking ability would have a good chance of scoring a 'yes' to all three items. By contrast a child with very poor walking ability is likely to score 'no' on all the items. Using the total score on our hypothetical measure (in this case ranging from 0 to 3), we can evaluate the responses to the items for a group of children with varying levels of mobility and see if the responses on the items resemble what one might expect. For example, a child with moderate levels of walking ability should be more likely to have a response of 'yes', 'yes' and 'no' to the above three questions (and achieve a score of 2) than have a response pattern of 'yes', 'no', 'yes' (and also achieve a score of 2). This is exactly what the Rasch analysis aims to do for us.

The Rasch model is a probabilistic model, that expresses the probability of an item being passed by people with given levels of ability, and similarly the probability of a person passing the items in the scale (Rasch 1960/1980). Or, as in the case of a health status measure for example, the probability of an item being agreed with or being endorsed. The notion of probability is important as we assume in Rasch analysis that there is always a chance that someone passes an item when it wasn't expected (e.g. maybe the person had a very good day). This may become clearer when we examine

the following concepts that underpin the Rasch model (e.g. Andrich 1988, Bond and Fox 2001, Rasch 1960/1980):

1. Each item within a scale has its own level of difficulty (*item parameter*) on the underlying latent trait or construct. Difficulty in the health context can be seen in terms of task difficulty (i.e. how easy/difficult it is to achieve a task on the Barthel Index) or in terms of how easy/difficult it is to agree with a statement (e.g. (dis)agreeing with a statement concerning pain severity). Thus, a scale will consist of items that are easier and items that are harder to 'achieve' or 'endorse'. Using data from a group of people completing the measure, it is possible to estimate this level of item difficulty (the item parameter) and place this along an interval logarithmic scale. The statistic that is needed to work this out during the Rasch analysis is the total score each item achieves. Figure 1 displays the item difficulties on the top of the ruler. Outcome measures, which have more than two response options to individual items (e.g. no pain, mild pain, moderate pain, severe pain, very severe pain) are slightly different, since having more response options presents a hierarchy of response category difficulty within each item. For such items, the item difficulty is therefore the average difficulty of these response categories.
2. Every person has his or her own amount of the latent trait or construct (for example level of walking ability, level of dependency, amount of pain). This is also known as the *person parameter* (or person ability) and this can also be displayed on the same ruler (or x-axis – figure 1) as the item parameters. Using data from a group of people completing the measure, it is possible to estimate this level of ability (the person parameter) and place this along an interval logarithmic scale (figure 1). The statistic

that is needed to work this out during the Rasch analysis is the total score each person achieves.

3. Since the person parameters and item parameters are displayed on the same interval logarithmic scale, it follows that the likelihood that a person will pass or endorse an item is related to how much of the construct s/he has and the item's level of difficulty, or indeed difficulty of the item response options (figure 1).
4. As explained in point 1 above, item parameters are estimated independent of the distribution of abilities in the particular group of persons for whom the items are appropriate (Bond and Fox 2001, p146). Similarly, person parameters are estimated independently of the distribution of their responses to the measure's items (see point 2 above). This concept is called *specific objectivity* and is a necessary requirement for the unidimensionality of a scale.
5. Georg Rasch demonstrated that these requirements of a measure (outlined under points 1-4 above) can be summarised by a formula that specifies the probabilistic expectations of items and persons (Rasch 1960/1980). Specifically, he proved that the probability of a correct (or positive) response to an item (i.e. for it to be passed or endorsed) is a logistic function of the difference between the person and item parameter. Diagrammatically this can be seen as the distance between the person parameter and item parameter on the x-axis (figure 1). For those interested, the mathematical models underpinning the Rasch model can be found elsewhere (Andrich 1978, Andrich 1988, Rasch 1960/1980).

The Rasch model tends to be illustrated with a hypothetical item characteristic curve. Figure 2 displays the expected raw scores on an item of a Holistic Health Beliefs scale (ranging from 0 to 5) on the y-axis and the person parameter estimates (in logits) on the x-axis (Kersten et al 2011). The grey curve represents the association between the expected raw scores and the log transformed interval scores (derived from the Rasch analysis). This figure clearly illustrates that the item characteristic curve is in an s-shape. In other words, it does not behave linearly as there isn't a 1:1 ratio between expected and real (log) scores. This means the item does not produce interval level data and should be treated as ordinal data (so you can't assume a score of 4 is twice as much as a score of 2). Secondly, and importantly, the graph is monotonic, that is the probability that someone gets a higher raw score on this question increases as the score on the underlying trait increases.

In commonly used approaches, such as factor analysis, we often explore what model we can find in the data (e.g. which factors emerge during a factor analysis). By contrast, in Rasch analysis we examine if the data arising from a measure fit the Rasch model as specified above. If a scale meets the expectations of the Rasch model (i.e. fits this model), the observed raw ordinal score gained through summing the scores from all the items can be transformed into interval scale measurement (Andrich 1988). Several key questions are asked during the analysis to determine if the data fit the Rasch model and some of these will be illustrated below with graphs.

1. Do the item response categories work as intended?

In the case of measures with items that have more than two response options (polytomous data) the log-transformed item scores are generated from the patients' responses to the item response options. It is important to check therefore, that these response options indeed reflect the increasing or decreasing latent trait to be measured. For example, a person with a very low location along the trait (i.e. little belief in holistic health, figure 1 and 2) relative to the location of a given item should have a greater probability of ticking a low response option on this item. By contrast, a person with a location much higher than the item location should be more likely to score higher on the item. Thresholds are the points where the probabilities of a response of either 0 or 1, and 1 or 2 (and so forth) are equally likely. Figure 3 shows two items of the Holistic Health Beliefs measure mentioned above (Kersten et al 2011); item scores range from 0 to 5 (where a higher score equals greater holistic health beliefs). The y-axes display the probability that someone gives a particular response to the item (i.e. that s/he ticks 0, 1, 2, 3, 4, or 5) given the overall level of holistic health beliefs this person has (displayed on the x-axis, person parameter estimates). The left figure shows that as someone's holistic health beliefs increase (depicted on the x-axis), the likelihood that s/he chooses a higher score on the item increases. Indeed, we see that the locations of the thresholds points between 0 and 1, 1 and 2, and so forth increase on the x-axis as the overall latent trait increases (i.e. you can see that the peak of each item score presents itself in the anticipated order from left to right). The right graph however shows an item in which this is not the case. Here we see that in the middle of the scale the thresholds are not in the order we would expect: the first thing to spot is that the peaks for item scores 2 and

3 are low and never appear above the other peaks. This means that as the latent trait increases (on the x-axis) there isn't ever a time at which the response options 2 or 3 are most likely. Secondly, the thresholds between response category 1 and 2 lies more to the right on the x-axis than the threshold between 2 and 3. In other words the thresholds are disordered. This means that at that part of the scale the response options, which were designed to measure Holistic Health Beliefs in an increasing fashion, don't work as was intended. If we were still in the development stage of the measure this would be the point at which we could consider changing the response options (either though clarifying the labels to the response options or by reducing the number of response options). However, if we simply want to convert the ordinal data to interval level data then one option would be to combine the category options that are the cause of the disordered thresholds so that we can obtain more accurate person parameter estimates from our analysis. This is an issue that is not explored in more traditional approaches to measurement, which simply assume that the response categories will behave (be used by respondents) in the way that was intended.

2. <u>Are the items unbiased?</u>

We would expect any measurement tool to measure in an unbiased way. For example, we would not consider it acceptable if measurement of mathematical ability was biased by gender (such that boys would get higher scores than girls on a specific item even though their underlying level of mathematical ability is the same). We can all think of measures that could possibly have such bias (e.g. extended ADL measures in older populations in which men may achieve lower scores on some items simply because

they are biased to favour roles more traditionally carried out by women in that generation). Thus, the Rasch model expects that each item is invariant (unbiased) across key groups (e.g. gender or age) (Grimby 1998, Holland and Wainer 1993). If variance or bias is observed the item is said to display *Differential Item Functioning* (DIF). DIF can be uniform; that is, the bias is present consistently across the trait. Using the mathematical ability example, uniform DIF in a given item would mean that boys score higher than girls across the trait even though their underlying mathematical ability is the same. Items which display uniform DIF in a developmental stage of a scale can be removed or improved. In the data analysis stage there are techniques that can be used to deal with uniform DIF, without deleting the item. DIF can also be non-uniform, where the bias is not consistent across the trait (for example, if at the low end of the mathematical ability scale boys score lower than girls on a given item, but at the high end of the scale, boys score higher). Items which display non-uniform DIF need to be removed from the scale, both at the scale development stage and the analytical stage, due to there being no known mathematical technique which can correct for this type of bias. Both graphical interpretation of item characteristic curves (ICCs) and statistical analysis (Analysis of Variance) are used to examine presence of DIF. Figure 4 gives an example of an item which doesn't have DIF (Kersten et al 2010b), one which has uniform DIF (Kersten et al 2010b), and one with non-uniform DIF (Kersten et al 2011).

3. <u>Do the items fit the Rasch model?</u>

Each item is closely inspected to explore if it fits the Rasch model. Essentially, this is an investigation of how close the observed scores are to the expected scores for a

particular person and item. Divergence from the expected scores are termed *residuals*. Graphically this can be demonstrated by plotting the item response curve (which displays expected scores on the y-axis and people's ability on the x-axis) and the observed scores for groups of people along the trait (figure 5 showing an item that fits well and an item that does not fit well) (Kersten et al 2011). In addition to this graphical representation, a range of fit residual statistics are reported in the literature. For their interpretation please refer to key text books (Bond and Fox 2001, Sherridan and Andrich 2009, Wilson 2005). Issues that could contribute to misfitting items include the presence of disordered thresholds and DIF (which we have already discussed) and multidimensionality or local dependence (see below).

4. Are items locally independent?

A key requirement of the Rasch model is that the items should only be associated with one another (i.e. correlated) through the latent trait that the test is measuring (Lord and Novick 1968). In Rasch analysis this is termed *local independence*. This is tested by exploring the correlation between the residuals, which should be low (the cut off value used in the literature is <0.20 below the average residual correlation) (Marais and Andrich 2008). If high correlations between the residuals are observed we can deduce that item responses depend not only on the latent trait being measured but on responses to other test items (local dependency). Local dependency results in the overestimation of reliability of the measure under investigation.

5. Are people scoring the items as would be expected?

Each person is also closely inspected to explore if s/he fits the Rasch model by exploring how closely the individual's observed scores on the items relate to the expected scores. Like item fit, person fit is also evaluated using a range of statistics (see for example Bond and Fox 2001, Sherridan and Andrich 2009, Wilson 2005). There are various reasons for a lack of person fit with the Rasch model, such as cognitive impairment (or a lack of understanding of the questions), guessing, lack of concentration or fatigue.

6. Is the scale unidimensional?

Another key requirement of the Rasch model is that the scale is *unidimensional*, i.e. that it measures one latent trait only. This examines whether each item belongs to one construct (or not) by exploring the associations between items. Factor analysis provides a similar assessment of unidimensionality but Rasch analysis takes this one step further and is therefore much stricter. It checks whether there are any patterns in the residuals and if none are found it lends strength to the hypothesis that the scale truly measures one construct only. The reason such stringent requirements are placed upon unidimensionality is because the only two important parameters in the Rasch analysis are person ability and item difficulty, and these parameters are placed on the same interval ruler. If the measure is not unidimensional these parameters could not be ordered on the same latent trait. For the more statistically minded reader, the paper by Tennant and Pallant (2006) demonstrates different findings from tests exploring unidimensionality.

7. Is the spread of items along the construct good?

Spread of items along the construct requires that the scale includes a range of items in terms of their level of difficulty. The item difficulty parameter is standardized on a logarithmic interval scale, with a mean of 0 and a standard deviation (SD) of 1. It is therefore helpful if a scale has items which range from -3 to +3 (in other words ±3SD's from the mean) in terms of their level of difficulty. If a scale has many items located at the lower end of the scale and few at the upper end it suggests it is measuring only the lower end of the construct and is therefore likely to suffer from ceiling effects (i.e. when the scale doesn't measure higher levels of the construct very well). The converse would be true if the scale had a floor effect. Similarly, by looking at the spread of items we can explore how well the scale is targeted to the sample from which the data was derived. Figure 6a gives an example of a scale that demonstrates a good spread of item thresholds. This scale is relatively well targeted to the sample although there are some people in this sample who score at the top of the scale (Kersten et al 2010a).

8. Does the scale distinguish between people with different amounts of the underlying trait?

In Rasch analysis this is measured with an index, called the *Person Separation Index* (PSI) (Andrich 1988). The PSI provides information on how precisely subjects have been spread out along the measurement construct. The PSI value can range from 0 to 1 (Fisher 1992). Values of 0.70 and higher would allow for group comparisons but for individual clinical use, values should be 0.85 and above. An example can be seen in

Figure 6a and Figure 6b. Figure 6a depicts a scale (Kersten et al 2010a), which has a high PSI (0.85), suggesting the scale can distinguish between 3-4 groups of people if grouped by their social integration scores. Figure 6b displays data from a scale with a low PSI (0.69) (Kersten et al 2011) and which as a result can only distinguish between two groups of people. One can see why this is important since a scale with a low PSI is less likely to be sensitive to change than a scale with a high PSI.

Conclusions

We acknowledge that Rasch analysis is described in the research literature in a rather technical and statistical manner, often not accessible to the lay reader. This article has been written in an attempt to de-mystify some of the core principles of Rasch analysis and for that reason we have not reported on various statistical techniques employed in Rasch analysis, but rather focused on the associated output and how one might interpret that. As might be clear from the text above, the examination of fit to the Rasch model is not straight forward and involves various qualitative and quantitative judgments. Essentially, positive answers to the above eight questions indicate the data fit the Rasch model and that the scale is unidimensional. The main strength of the Rasch model lies in its theoretical and scientific underpinning, which have been shown to mathematically hold true (Andrich 1988, Bond and Fox 2001, Rasch 1960/1980). Thus, a measure that fits the Rasch model is more robust than one that doesn't and arguably, the latter should not be used in practice or research.

Key Points

- Questionnaire based outcome measures tend to be treated incorrectly as if they produce interval data when in fact the data is ordinal.
- The Rasch model is a probabilistic model used to evaluate the internal construct validity of an outcome measure.
- Rasch analysis includes a range of tests for checking how well the data arising from an outcome measure fit the Rasch model.
- Data which fit the Rasch model can be converted to an interval scale.

Acknowledgements

We thank the members of the Person Centred Research Centre, at AUT University for their valuable comments on a draft version of this paper (www.aut.ac.nz/research/research-institutes/hrrc/research-activities/personrehab)

References

ACC (2009): Guide to Outcome Measure Reporting. Accident Compensation Corporation, New Zealand.

Andrich D (1978): Rating formulation for ordered response categories. *Psychometrica* 43; 561-573.

Andrich D (1988): Rasch models for measurement series: quantitative applications in the social sciences no. 68. London: Sage Publications.

Bond TG, Fox CM (2001): Applying the Rasch model. Fundamental measurement in the human sciences. London: Lawrence Erlbaum Associates.

Chartered Society of Physiotherapy (2011): Measuring for quality improvement in physiotherapy. Chartered Society of Physiotherapy, United Kingdom http://www.csp.org.uk/topics/measuring-quality-improvement-physiotherapy (accessed May 12, 2011).

Chartered Society of Physiotherapy (2005): Core Standards of Physiotherapy Practice. Chartered Society of Physiotherapy, United Kingdom.

Department of Health (2008): Guidance on the routine collection of Patient Reported Outcome Measures (PROMs). For the NHS in England 2009/10. Department of Health, United Kingdom.

Enderby PM, John A, Petheram B (2006): Therapy Outcome Measures for Rehabilitation Professionals: Speech and Language Therapy, Physiotherapy, Occupational Therapy (2nd edn). Chichester: John Wiley & Sons Ltd.

Fisher W Jr (1992): Reliability Statistics. *Rasch Measurement Transactions* 6: 238.

Grimby G (1998): Useful reporting of DIF. *Rasch Measurement Transactions* 12: 651.

Hobart J, Cano S (2009): Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods. *Health Technology Assessment* 13.

Holland PW, Wainer H (1993): Differential Item Functioning. New Jersey: Hillsdale. Lawrence Erlbaum.

Holmbeck GN, Devine KA (2009): Editorial: an author's checklist for measure development and validation manuscripts. *Journal of Pediatric Psychology* 34: 691-696.

Horner D, Larmer PJ (2006): Health outcome measures. *New Zealand Journal of Physiotherapy* 34: 17-24.

Kayes NM, McPherson KM (2010): Measuring what matters: does 'objectivity' mean good science? *Disability & Rehabilitation* 32: 1011-1019.

Kersten P, Adams JA, Burridge J, Cooper C, Hammond A (2011): An exploration of combining unilateral hand function questions into a measure: the Michigan Hand Questionnaire. Proceedings of the New Zealand Rehabilitation Association Biennial Conference, Auckland, pp. 70.

Kersten P, Ashburn A, George S, Low J (2010a): The Subjective Index for Physical and Social Outcome (SIPSO) in Stroke: Investigation of its subscale structure. *BMC Neurology* 10.

Kersten P, White PJ, Tennant A (2010b): The Visual Analogue WOMAC 3.0 scale - Internal validity and responsiveness of the VAS version. *BMC Musculoskeletal Disorders* 11.

Kersten P, White PJ, Tennant A (2011): Construct validity of the Holistic Complementary and Alternative Medicines Questionnaire (HCAMQ) – an investigation

using modern psychometric approaches. *Evidence-Based Complementary and Alternative Medicine* eCAM Advance Access published September 30, 2009.

Laver Fawcett A (2007): Principles of Assessment and Outcome Measurement for Occupational Therapists and Physiotherapists: Theory, Skills and Application. Chichester: John Wiley & Sons Ltd.

Lord Darzi of Denham (2008): High quality care for all: NHS Next Stage Review final report. Department of Health, United Kingdom.

Lord FM, Novick MR (1968): Statistical theories of mental test scores. Reading MA: Addison-Welsley Publishing Company.

Mahoney FI, Barthel D (1965): Functional evaluation: The Barthel Index. *Maryland State Medical Journal* 14: 56-61.

Marais I, Andrich D (2008): Effects of varying magnitude and patterns of response dependence in the unidimensional Rasch model. *Journal of Applied Measurement* 9: 105-124.

McDowell I (2006): Measuring Health: A Guide to Rating Scales and Questionnaires. New York: Oxford University Press.

Rasch G (1960/1980): Probabilistic models for some intelligence and attainment tests (revised and expanded ed.). Chicago: The University of Chicago Press.

Sherridan B, Andrich D (2009): Extending the RUM2030 analysis. Perth: RUMM Laboratory.

Streiner DL, Norman GR (2008): Health Measurement Scales: a practical guide to their development and use edn). Oxford: Oxford University Press.

Tennant A, Pallant JF (2006): Unidimensionality matters! (a tale of two Smiths?). *Rasch Measurement Transactions* 20: 1048-1051.

The Physiotherapy Board of New Zealand (2009): Physiotherapy Competencies for physiotherapy practice in New Zealand. The Physiotherapy Board of New Zealand.

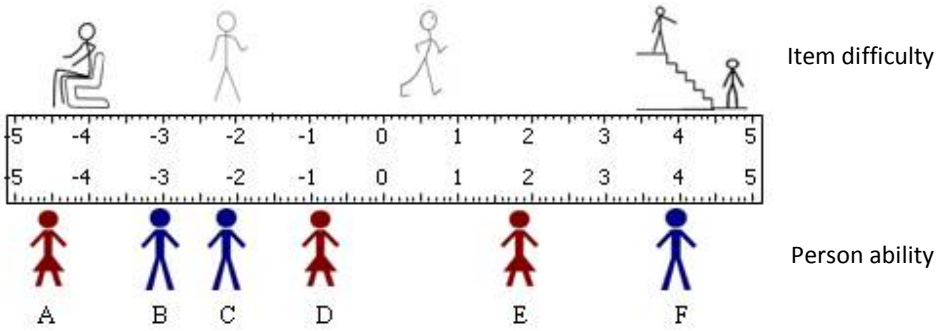
Turner-Stokes L, Rusconi S (2003): Screening for ability to complete a questionnaire: A preliminary evaluation of the AbilityQ and ShoulderQ for assessing shoulder pain in stroke patients. *Clinical Rehabilitation* 17: 150-157.

Tyson S, Greenhalgh J, Long AF, Flynn R (2010): The use of measurement tools in clinical practice: an observational study of neurorehabilitation. *Clinical Rehabilitation* 24: 74-81.

Van Peppen RPS, Maissan FJF, Van Genderen FR, Van Dolder R, Van Meeteren NLU (2008): Outcome measures in physiotherapy management of patients with stroke: a survey into self-reported use, and barriers to and facilitators for use. *Physiotherapy Research International* 13: 255-270.

Wilson M (2005): Constructing measures. An item response modelling approach. New Jersey: Lawrence Erlbaum Associates, Publishers.

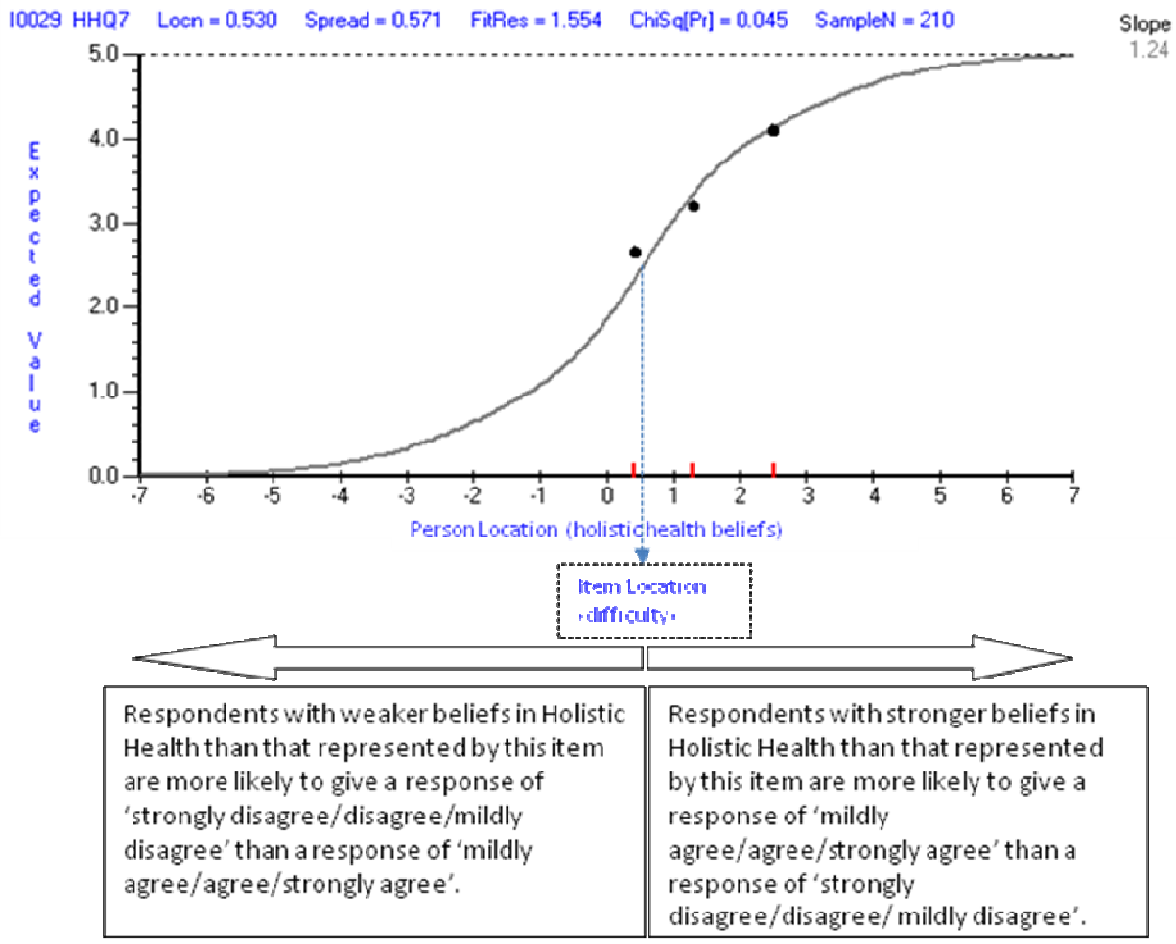
Figure 1 A visual representation of item difficulties (parameters) and person abilities (parameters)



Legend to Figure 1

Item difficulty (parameters) and person abilities (parameters) of this hypothetical mobility outcome measure have been estimated using Rasch analysis and both have been placed along the same interval scaled, logarithmic ruler. Person A achieved a low total score on the outcome measure. The diagram reveals that people with this level of mobility are unlikely to be able to sit, stand, walk or climb stairs (and therefore they are likely to fail these items). Person D achieved a higher total score on the outcome measure. People with this level of mobility are likely to pass the items which measure sitting and standing ability, but likely to fail items which measure walking or climbing stairs ability. Person F achieved high total score. People with this level of mobility are likely to pass items measuring sitting, standing, and walking ability and have a 50/50 chance of passing the stair climbing item (given their level of ability is the same as the amount of ability being measured by this item).

Figure 2 Item response curve for item 7 of the Holistic Health Beliefs scale (Kersten et al 2011).



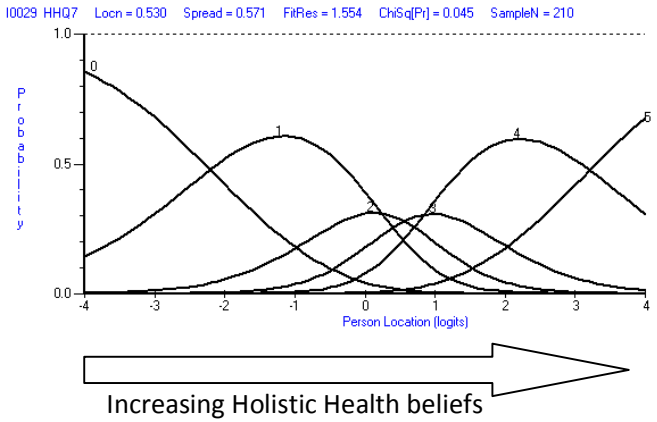
Legend to Figure 2

The y-axis displays the expected scores on Item 7 of the Holistic Health Beliefs subscale. Item 7 states: "If a person experiences a series of stressful life events they are likely to become ill" (response options 0 Strongly Disagree, 1 Disagree, 2 Mildly Disagree, 3 Mildly Agree, 4 Agree, 5 Strongly Agree). This item has an estimated difficulty of 0.53 logits. The x-axis displays person parameters in log-transformed

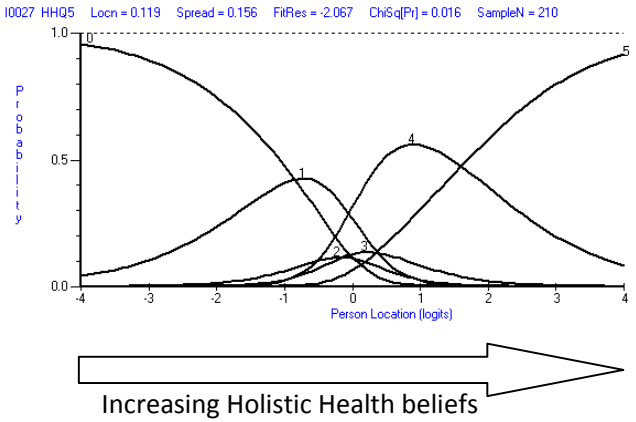
interval scores, estimated from the Rasch analysis. The three black dots represent three groups of respondents, grouped by their holistic health beliefs.

Figure 3 Example graphs of items with (a) ordered and (b) disordered thresholds (Kersten et al 2011)

(a)



(b)



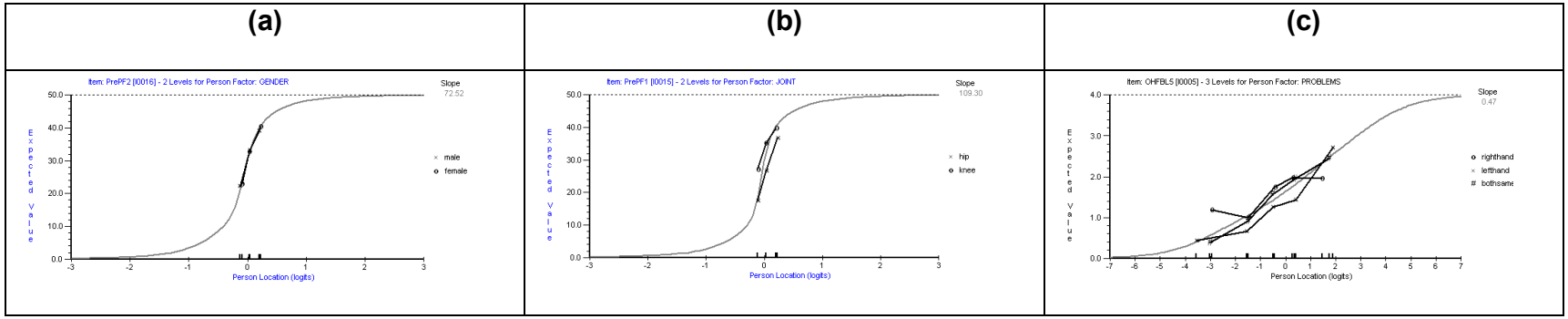
Legend to Figure 3

Each response option to an item is represented by a probability curve, which resembles the likelihood that this response option is ticked (displayed on the y-axis), given the amount of Holistic Health beliefs someone has (which is displayed on the x-axis). Response options: 0 Strongly Disagree, 1 Disagree, 2 Mildly Disagree, 3 Mildly Agree, 4 Agree, 5 Strongly Agree

a) This graph shows that as people's total scores on this measure increases (i.e. they have stronger Holistic Health beliefs) they are increasingly likely to give higher responses on this item (item 7): e.g. people with a person location of -3 are more likely to tick the 'strongly disagree' response option than the other response options, those with a score of -1 are more likely to most likely to tick the 'disagree' response option than the other response options, and so forth (note each response option has its own distinct peak that appears as the highest point along the trait, in the order one would expect).

b) This graph shows that as people's total scores on this measure increases (i.e. they have stronger Holistic Health beliefs) they are not increasingly likely to give higher responses on this item (item 5): e.g. people with a person location of -0.5 are more likely to tick the 'disagree' response option than the other response options, people with a person location of 0.5 are more likely to tick the 'agree' response option than the other response options. There is never an instance along the trait of Holistic Health beliefs when the response option 'mildly disagree' or 'mildly agree' are the most likely (not every response option has its own distinct peak that appears as the highest point along the trait, in the order one would expect; in Rasch analysis this is referred to as reversed thresholds).

Figure 4 Examples of item response curves displaying (a) no Differential Item Functioning (DIF), (b) Uniform DIF, (c) Non-Uniform DIF



Legend to Figure 4

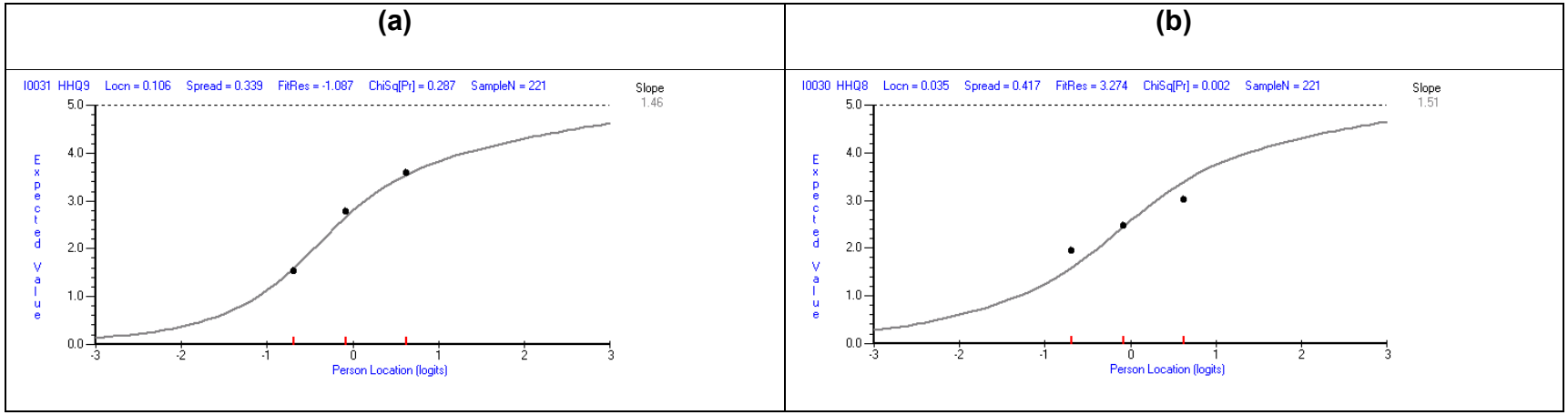
The grey curves display the association between expected scores on the item (y-axes) and the person's amount of the construct (person parameters displayed on the x-axes). Each curve is also split by different groups: i.e. (a) by gender, (b) by joint problem experienced, (c) by the hand problems experienced

a) This figure a shows an item that is unbiased, i.e. as people's Physical Functioning increases (x-axes) their expected scores on this item increases too, irrespective of gender (Kersten et al 2010b).

b) This figure shows an item that is biased uniformly across the trait, i.e. as people's Physical Functioning increases (x-axes) we can see that those with a knee problem score higher on this item than those with a hip problem (even though their overall level of Physical Functioning is the same). (Kersten et al 2010b).

c) This figure displays an item which is biased across the trait in a non-uniform fashion, i.e. when Overall Hand Function is low people with a problem of the right hand score higher than people with problems of the left hand and people with bilateral problems – further along the scale the opposite is the case (Kersten et al 2011).

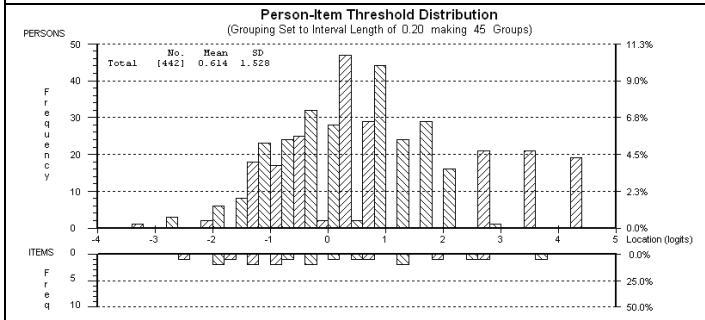
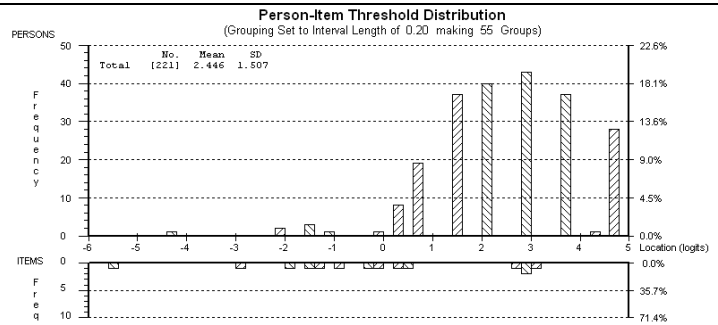
Figure 5 Examples of an item response curve of (a) an item that fits the Rasch model and (b) an item that does not fit the Rasch model



Legend to Figure 5

The y-axes show the expected score on the items, given someone's level of health beliefs (x-axes) (Kersten et al 2011). The three back dots in each of the graphs resemble the relationship between expected and observed scores for three groups of people. These lie close to the line in figure a, suggesting the item fits the Rasch model; figure b shows significant deviation from the line and his item does not fit the Rasch model (this would be tested also using statistical tests).

Figure 6 Examples of Person-Item-Threshold maps

(a)	(b)
	

Legend to Figure 6

The diagrams display the number of item thresholds on the bottom y-axis and the number of people on the top y-axis. Item and person parameters are placed along the same interval logarithmic scale (x-axis). Figure a displays item thresholds that are well spread along the construct of social integration. In addition, the item thresholds are well targeted to measure the level of social integration experienced by study participants (Kersten et al 2010a). Figure b displays findings from a Holistic Health Beliefs scale: its item thresholds are sparsely spaced along the measurement construct and not well targeted to the population studied (which displays high levels of holistic health beliefs) (Kersten et al 2011)