

Can item response theory reduce patient burden when measuring health status in neurological disorders? Results from Rasch analysis of the SF-36 physical functioning scale (PF-10)

C Jenkinson, R Fitzpatrick, A Garratt, V Peto, S Stewart-Brown

Abstract

Background—Indices of physical function may have a hierarchy of items. In cases where this can be demonstrated it may be possible to reduce patient burden by asking them to complete only those items which relate directly to their own level of ability.

Objectives—To determine whether statistical procedures, operationalising what is known as item response theory (IRT), can be used to assess the unidimensionality of the 10 item physical functioning domain of the SF-36 in patients with Parkinson's disease and motor neuron disease, and, secondly, to determine whether it would be possible to administer subsets of items to certain patients, on the basis of their replies to other items in the scale, thereby reducing patient burden.

Methods—Rasch analysis, a form of IRT methodology, of the 10 item physical functioning domain (PF-10) in two neurological patient samples was undertaken and the results compared with results of a Rasch analysis of data gained from a population survey (the third Oxford healthy lifestyles survey).

Results—Evidence from the analyses suggests that the PF-10 does not form a perfect hierarchy on a unidimensional scale. However, certain items seem to form a hierarchy, and responses to some of them are contingent on responses to the other items.

Conclusions—Rasch analysis of the PF-10 in neurological patients has indicated that certain items of the scale are hierarchically ordered, and consequently not all respondents would need to complete them all: indeed those most severely ill would be required to complete less items than those with only limited disabilities. The implications of this are discussed.

(J Neurol Neurosurg Psychiatry 2001;71:220-224)

Keywords: SF-36, health outcomes, physical functioning, Rasch analysis, item response theory

years the place of patient self report questionnaires has come to occupy an increasingly central position.² However, a possible criticism of such measurement is that it puts a considerable burden on chronically ill patients, who may have seriously disabling conditions. Ideally, therefore, questionnaires should be simple to understand and as brief as possible, yet still satisfying the requirements of validity and reliability which are central to all measurement. The most common procedure for creating shorter form instruments is to undertake a statistical analysis of the original measure that simply reduces the number of items.^{3,4} However, although this will lead to brevity it also increases the error term in measurement, reducing accuracy and precision.⁵ Another method of reducing item burden is to request respondents to complete only the items that are of direct relevance to them. For example, someone who is unable to walk at all need not be requested to complete items on ambulation. Questionnaires developed using classic psychometric techniques request patients to complete all the items, even though some may be inappropriate for their level of ability. However, more recent psychometrics can reduce the number of items any subject may have to complete while retaining the same original item pool. If questions on a measure form a hierarchy then respondents need only complete those that assess their own level of ability. Determining a hierarchy on a questionnaire can be undertaken using item response theory (IRT) scoring procedures.^{6,7}

Often referred to as an item characteristic curve technique, IRT methodology begins with the assumption that any item will pose differing degrees of difficulty to different people in any given population. Furthermore, different items pose differing degrees of difficulty. These basic claims lead to two assumptions. Firstly, that the items constitute a hierarchical structure on a unidimensional concept, and, furthermore, that reproducibility of the item hierarchy can be achieved on different groups and across test occasions.^{8,9} If these two assumptions are satisfied then it is reasonable to assume that certain questions will be answered in a predictable manner on the basis of answers to other questions. For example, someone who answers a question such as "Can you walk at all?" in the negative is highly unlikely to affirm the statement "I can run long distances." If a hierarchy of statements can therefore be found

Health Services
Research Unit,
Department of Public
Health, University of
Oxford, Institute of
Health Sciences,
Headington, Oxford
OX3 7LF, UK
C Jenkinson
V Peto
S Stewart-Brown

Department of Public
Health
R Fitzpatrick

Centre for Health
Outcomes
Development,
Department of Public
Health
A Garratt

Pickier Institute
Europe, King's Mead
House, Oxpens Road,
Oxford OX1 1RX, UK
C Jenkinson

Correspondence to:
Dr C Jenkinson
crispin.jenkinson@
dphpc.ox.ac.uk

Received 15 September 2000
and in revised form
23 February 2001
Accepted 7 March 2001

The ultimate goal of outcomes research is to provide meaningful, accurate assessments of health, which can inform treatment decisions and regimes.¹ Within the field of neurology evaluation of the patient has been largely by means of clinical assessment, but in recent

HEALTH AND DAILY ACTIVITIES

The following questions are about activities you might do during a typical day. Does your health limit you in these activities? If so, how much?

(Please tick **one** box on each line)

	Yes, limited a lot	Yes, limited a little	No, not limited at all
(a) Vigorous activities , such as running, lifting heavy objects, participating in strenuous sports	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
(b) Moderate activities , such as moving a table, pushing a vacuum, bowling or playing golf	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
(c) Lifting or carrying groceries	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
(d) Climbing several flights of stairs	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
(e) Climbing one flight of stairs	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
(f) Bending, kneeling or stooping	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
(g) Walking more than a mile	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
(h) Walking half a mile	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
(i) Walking 100 yards	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
(j) Bathing and dressing yourself	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Figure 1 Physical function dimension of the SF-36.

then those who are most disabled by their condition need complete less items than those with less severe forms of the condition. This is a desirable solution as patient burden is considerably reduced for the most severely ill. Such measurement is sometimes referred to as “test free”, in that people can be compared to one another on a trait or ability even if they have completed different questions, or a different number of questions. Rasch analysis is the most commonly used form of IRT methodology.¹⁰

The purpose of this paper is twofold: firstly, to determine whether IRT scoring criteria, using Rasch analysis, would be appropriate for the 10 item physical functioning domain (PF-10) of the 36 item short form health survey (SF-36) in patients with Parkinson's disease and motor neuron disease by assessing the unidimensionality of the scales, model fit, and comparability with data gained from a general population sample; and secondly, to determine whether it would be possible to administer subsets of items to certain patients, on the basis of their replies to other items in the scale, thereby reducing patient burden. The PF-10 is reproduced in figure 1.

Methods

Three data sets were analysed for this paper: a normative dataset of the general population, patients who are members of the United Kingdom Parkinson's Disease Society, and a sample of patients with motor neuron disease drawn from across European countries. Recruitment into the studies is explained in full elsewhere,^{11–13} although a brief outline of the methodology of each of the surveys is outlined below.

Normative data were gained from the Oxford healthy lifestyles survey (OHLS III).

Questionnaires containing questions on life-style, as well as a copy of the SF-36, were mailed to randomly selected people in Oxfordshire, Northamptonshire, Berkshire, and Buckinghamshire. Completed questionnaires were obtained from 8889 of 13 800 people originally contacted, a response rate of 64.4%. Of those who did return questionnaires 8801 (99.0%) of respondents answered the question relating to sex, of whom 3863 (43.4%) were men and 4938 (55.6%) were women. The mean age of the sample was 41.6 years (SD 12.6; range 18 to 65).

Patient's with Parkinson's disease were recruited from a postal survey of members of local Parkinson's Disease Society branches. Four hundred and five patients who were registered with five branches of the Parkinson's Disease Society were contacted. Fifteen people were subsequently removed from the denominator as they could not be traced, were deceased, or did not have Parkinson's disease. A total of 227 questionnaires were returned, yielding a response rate of 58.2%. The mean age of this sample was 70.3 years (SD 9.0; range 40.9 to 87.7); 57.4% men; 42.6% women. The mean number of years since diagnosis was 8.6 (SD 6.7) (n=218).

Patients with motor neuron disease were recruited via the amyotrophic lateral sclerosis health profile study (ALS-HPS), a Pan European survey of motor neuron disease patient experiences and health status. Patients are recruited into the study when visiting their doctor and then return the questionnaires via the post. Five hundred and fifty one patients have been recruited into the ALS-HPS, of which there were 451 (81.85%) patient responses. The mean age of respondents was 59.91 years (SD 11.24; range 24.3–88.8). Two

Table 1 Mean (SEM) item calibrations and strata for PF-10 based on the third Oxford health and lifestyles survey (OHLS III, $n=8853$)

	<i>n</i>	Average item calibration (logit)	SEM	Infit statistic	Strata
Vigorous activities	5332	3.98	0.03	0.3	9
Climbing several flights of stairs	5339	1.19	0.03	1.6	8
Bending, kneeling, or stooping	5340	0.97	0.03	7.1	7
Walking more than a mile	5345	0.32	0.04	0.7	6
Lifting or carrying groceries	5349	0.10	0.04	1.7	5
Moderate activities	5352	-0.17	0.04	-4.6	4
Walking half a mile	5318	-0.96	0.05	-5.1	3
Climbing one flight of stairs	5319	-1.30	0.05	-4.4	2
Bathing or dressing	5355	-2.03	0.06	9.9	1
Walking 100 yards	5314	-2.11	0.06	2.2	1

hundred and fifty three (56.1%) patients were men, 197 (43.7%) women, and one did not reply to the question. The mean number of years since diagnosis was 1.39 years (SD 1.88) ($n=420$), and the mean number of years since first symptoms were noticed by the patient was 2.18 (SD 2.46) ($n=388$).

ANALYSIS PLAN

Rasch analyses were performed on the three data sets outlined above. Two claims are tested with the Rasch rating scale model: firstly, the more capable a person is in physical functioning the less likely that person is to have limitations on any given item, and, secondly, the easier the item, the more likely the person will report no limitations. The Rasch model provides item locations along a hypothesised common measurement continuum. These calibrations define the hierarchical order of the items along the continuum. Calibrations for each item are expressed in logits, which is the natural log of an odds ratio. In this instance, an odds ratio is the level of performance of an item in relation to the performance (in terms of difficulty) of the total set of items. Logits typically range from -4 to +4, with logits of greater positive magnitude representing increasing item difficulty.

Unidimensionality was assessed using the information weighted fit statistic (infit). This fit statistic is standardised such that it takes the approximate form of a t distribution. Values lying outside -2.0 and +2.0 indicate that data for that item may not fit the model.

Results

Rasch analysis of the OHLS III normative dataset is reported in table 1.

Close inspection of the results tends to suggest that the Rasch model does not provide a perfect fit for the items on the physical function

domain of the SF-36. The unidimensionality of a multi-item index for a given sample is partly determined by goodness of fit statistics, which is an index of how well the item calibration (expressed in logits) fits the data for all of the subjects in the sample, who did not score all items at the floor or, alternatively, did not score all items at the ceiling. Infit statistics are reported, which are standardised to approximate a mean of zero and an SD of 1. As noted above, high infit statistics (>2.0) may indicate that an item does not fit the model well and is not closely related to the overall construct. Low infit statistics (<-2.00) indicate that items measure redundant or overlapping content areas.¹⁴ The items bending, kneeling, and stooping and bathing and dressing have very large infit statistics indicating that they do not fit the model at all well. On the other hand moderate activities, walking half a mile, and climbing one flight of stairs have fairly large negative infit statistics indicating that one or more of the items are redundant as they are measuring overlapping areas.

The results also suggest that the hierarchical nature of the items on the PF-10 is not completely satisfied. For example, to determine the spacing of each item calibration (expressed in logits) an associated SE estimate is calculated and used to define distinct strata along the measurement continuum. The spacing of items can be described by the number of distinct strata that can be identified in the scale. Strata can be defined as a separation of at least ± 0.15 logits.¹⁵ Nine distinct strata can be found for the PF-10 on the OHLS III data, with the items bathing and dressing and walking 100 yards having very similar item calibrations. Overall, the PF-10 data from the OHLS-III indicate a hierarchy of items, but with some possible redundancy.

Table 2 Mean (SEM) calibrations and strata for PF-10 based on the Parkinson's disease sample ($n=227$)

	<i>n</i>	Average item calibration (logit)	SEM	Infit statistic	Strata	Item order in population survey
Vigorous activities	186	2.78	0.25	0.9	7	10
Walking more than a mile	187	1.35	0.19	-1.4	6	7
Climbing several flights of stairs	186	0.77	0.17	0.3	5	9
Walking half a mile	183	0.14	0.16	-0.6	4	4
Bending, kneeling or stooping	186	0.03	0.16	1.5	4	8
Lifting or carrying groceries	186	-0.26	0.15	-2.2	3	6
Moderate activities	187	-0.27	0.15	-1.7	3	5
Bathing or dressing	186	-1.12	0.14	4.8	2	2
Climbing one flight of stairs	185	-1.66	0.14	-1.0	1	3
Walking 100 yards	183	-1.75	0.14	-1.8	1	1

Table 3 Mean (SEM) calibrations and strata for PF-10 based on the ALS-HPS sample (n=446)

	n	Average item calibration (logit)	SEM	Infit statistic	Strata	Item order in population survey
Vigorous activities	278	3.31	0.18	-0.7	7	10
Walking more than a mile	277	0.65	0.13	0.0	6	7
Moderate activities	275	0.58	0.13	-1.7	6	5
Lifting or carrying groceries	278	0.42	0.13	-0.3	5	6
Climbing several flights of stairs	277	0.40	0.13	0.5	5	9
Bending, kneeling, or stooping	279	-0.29	0.13	-0.5	4	8
Walking half a mile	276	-0.48	0.13	-1.7	3	4
Bathing or dressing	281	-1.14	0.12	6.3	2	2
Climbing one flight of stairs	278	-1.66	0.13	-2.4	1	3
Walking 100 yards	272	-1.80	0.13	-1.6	1	1

Table 2 provides results of a Rasch analysis of the Parkinson's disease data. Once again the items do not form a perfect fit with the model, and once again the item bathing and dressing has a large infit statistic. However, in general the fit of items is better than that for the general population, although the number of strata are less due to limited differences between some items in terms of the infit statistics. It is particularly interesting to note that the order of items is not exactly the same as that for the OHLS III data, which further suggests that the data are not truly unidimensional. This result is also borne out for the ALS-HPS dataset, where once again the bathing and dressing item gains a large infit statistic, and the number of strata is also seven, and does not reflect the same hierarchy as either the OHLS or Parkinson's disease dataset (table 3).

Although the scale as a whole does not seem to fulfil the requirements of unidimensionality, there are none the less hierarchies of items within the scale. Thus the following items form the same hierarchy in both patient groups as well as the OHLS III dataset:

- Vigorous activities
- Walking more than a mile
- Walking half a mile
- Climbing one flight of stairs
- Walking 100 yards
- as do
- Climbing several flight of stairs
- Climbing one flight of stairs.

These items also conform to the requirements of the infit statistics, and never overlap in terms of logits (never within ± 0.2 of each other), even if they sometimes overlap with other items. It could be argued that severely ill patients who indicate severe disability on the easiest of the items in these groups do not have to complete the other item or items. For example, people who indicate that they are limited a lot in walking 100 yards need not complete the items on climbing one flight of stairs, walking half a mile, walking more than a mile, or vigorous activities, and consequently need not compete the item climbing several flight of stairs. Indeed this is borne out by the data. For example, 80 patients with Parkinson's disease indicated they were limited a lot in their ability to walk 100 yards, and at least 97.5% also indicated they were limited a lot in walking half a mile, walking a mile, or vigorous activities. Similarly, 200 patients with ALS claimed that they were limited a lot in their ability to walk

100 yards. At least 98% of this group also indicated they were limited a lot in walking half a mile, walking a mile, or vigorous activities.

Discussion

Health status measures must fulfil several requirements to be useful. Typically, the attributes that are most discussed centre on reliability, validity, sensitivity to change, and interpretability. However, data from such measures are likely to be compromised if patients find completing instruments a burden. Instruments with what may seem a modest number of items to someone in perfect health can present a considerable challenge for people who find writing and movement difficult, or, indeed, impossible. Consequently, brevity is to be sought whenever possible in the design and implementation of health status measures. Rasch analysis of data from the SF-36 physical functioning domain suggests that some items need not be administered to the most severely ill patients. Consequently, the most severely impaired are most likely to benefit from instruments the length of which is determined by Rasch methods. However, this potential advantage cannot be tested in the current study. The reduction of questionnaire length by Rasch analysis is also likely to be of considerable value when questionnaires are completed on computers. Such on line questionnaires may be completed by the patient, or with help from someone else and may be used in hospital and doctors' surgeries. Such computer programmes have already been developed in the United States, for example the dynamic health assessment system (DynaHA™) developed by QualityMetric¹⁶ uses a pool of items from widely used health surveys including general and disease specific-DynHA™ designs. Only those items relevant to a person's health state are used. By scoring all responses on a standard metric, results can be compared for those who answer different questions. The brevity of the assessment means that the DynHA™ system determines scores at a fraction of the burden of traditional health assessments. Furthermore, the developers of this system go further and claim that DynHA™ is the first system to provide results in user friendly, real time reports that are precise enough for monitoring and managing care. Such computer adaptive testing is likely to become increasingly popular in the 21st century¹⁷ as technology advances and

becomes more efficient, easy to use, and economically attractive.

The analysis presented here outlines the potential benefits of item response theory to those using and developing questionnaires for neurological patients who are likely to be severely ill. Not all questionnaires or domains in questionnaires will be appropriate for this form of analysis, as not all questionnaires are designed to cover unidimensional concepts with a hierarchy of items. However, we present some indication of the possible use of this technique. Research is currently under way as to the usefulness of this technique with disease specific questionnaires designed for use within the sphere of neurological disorders.

The studies reported here were funded by the NHS Executive—South East, the Directors of Public Health for Berkshire, Buckinghamshire, Northamptonshire, and Oxfordshire Health Authorities (for OHLS data), the Parkinson's Disease Society (for Parkinson's disease data) and Aventis Pharma (for data from the ALS-HPS). More information on all of these studies can be gained from CJ.

- 1 Jenkinson C, McGee H. *Health status measurement: a brief but critical introduction*. Oxford: Radcliffe Medical Press, 1998.
- 2 Jenkinson C, Fitzpatrick R, Jenkinson D. Health status measurement in neurology. In: Jenkinson C, Fitzpatrick R, Jenkinson D, eds. *Health status measurement in neurological disorders*. Oxford: Radcliffe Medical Press, 2000.
- 3 Jenkinson C, Fitzpatrick R, Peto V, et al. The PDQ-8: development and validation of a short-form Parkinson's disease questionnaire. *Psychology and Health* 1997;12:805–14.
- 4 Jenkinson C, Fitzpatrick R. A reduced item set for the amyotrophic lateral sclerosis assessment questionnaire: development and validation of the ALSAQ-5. *J Neurol Neurosurg Psychiatry* 2001;70:70–3.
- 5 Oppenheim AN. *Questionnaire design, interviewing and attitude measurement, new edition*. London: Pinter, 1992.
- 6 Hambleton RK, Swaminathan H, Rogers HJ. *Fundamentals of item response theory*. London: Sage, 1991.
- 7 Streiner D, Norman G. *Health measurement scales: a guide to their development and use*. 2nd ed. Oxford: Oxford University Press, 1995.
- 8 Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. *Med Care* 2000;38(suppl II): 28–42.
- 9 Andrich D. *Rasch models for measurement*. London: Sage, 1988.
- 10 Rasch G. *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press, 1980.
- 11 Jenkinson C, Stewart-Brown S, Petersen S, et al. Evaluation of the SF-36 version II in the United Kingdom. *J Epidemiol Community Health* 1999;53:46–50.
- 12 Jenkinson C, Peto V, Fitzpatrick R, et al. Self reported functioning and well being in patients with Parkinson's disease: comparison of the short form health survey (SF-36) and the Parkinson's disease questionnaire (PDQ-39). *Age Ageing* 1995;24:505–9.
- 13 Jenkinson C, Fitzpatrick R, Swash M, et al. The ALS health profile study: quality of life of ALS patients and carers in Europe. *J Neurol* 2001;415:835–40.
- 14 Haley SM, McHorney CA, Ware JE. Evaluation of the MOS SF-36 physical functioning scale (PF-10):1: unidimensionality and reproducibility of the Rasch item scale. *J Clin Epidemiol* 1994;47:671–84.
- 15 Silverstein B, Kilgore KM, Fisher WP, Harley JP, Harvey RF. Applying psychometric criteria to functional assessment in medical rehabilitation, 1: exploring unidimensionality. *Arch Phys Med Rehabil* 1991;72:631–7.
- 16 <http://www.qualitymetric.com/products/dynha>
- 17 Hambleton RK. Emergence of item response modelling in instrument development and data analysis. *Med Care* 2000; 38(suppl II): 60–5.