REPORT

More relevant, precise, and efficient items for assessment of physical function and disability: moving beyond the classic instruments

J F Fries, B Bruce, J Bjorner, M Rose

Ann Rheum Dis 2006;**65**(Suppl III):iii16–iii21. doi: 10.1136/ard.2006.059279

Objectives: Patient reported outcomes (PROs) have become standard study endpoints. However, little attention has been given to using item improvement to advance PRO performance which could improve precision, clarity, patient relevance, and information content of ''physical function/disability'' items and thus the performance of resulting instruments.

Methods: The present study included1860 physical function/disability items from 165 instruments. Item formulations were assessed by frequency of use, modified Delphi consensus, respondent judgement of clarity and importance, and item response theory (IRT). Data from 1100 rheumatoid arthritis, osteoarthritis, and normal ageing subjects, using qualitative item review, focus groups, cognitive interviews, and patient survey were used to achieve a unique item pool that was clear, reliable, sensitive to change, readily translatable, devoid of floor and ceiling limitations, contained unidimensional subdomains, and had maximal information content.

Results: A ''present tense'' time frame was used most frequently, better understood, more readily translated, and more directly estimated the latent trait of disability. Items in the ''past tense'' had 80–90% false negatives ($p<0.001$). The best items were brief, clear, and contained a single construct. Responses with four to five options were preferred by both experts and respondents. The term physical function may be preferable to the term disability because of fewer floor effects. IRT analyses of ''disability'' suggest four independent subdomains (mobility, dexterity, axial, and compound) with factor loadings of 0.81–0.99.

Conclusions: Major improvement in performance of items and instruments is possible, and may have the effect of substantially reducing sample size requirements for clinical trials.

Patient reported outcomes (PROs) of physical function/disability, such as the Disability Index of the Health Assessment Questionnaire (HAQ-DI)[1] and the PF-10 of the Medical Outcome Survey Short Form 36 (SF-36),[2] have become the most widely employed fundamental and classic outcome measures in clinical trials and observational clinical studies in rheumatology. The HAQ-DI and the PF-10 are considered to be the most sensitive of outcome measures, particularly in rheumatoid arthritis (RA) studies. They are endorsed by both the American College of Rheumatology and Outcome Measures in Rheumatology (OMERACT), and they have been used to achieve a physical function improvement ''indication'' from the US Food and Drug Administration.[3–6] Over the past 25 years, as usage of the HAQ-DI and SF-36 has grown, they have been extensively validated and are available in many languages.

These classic instruments, while adapted and modified numerous times, remain similar to their parents, although the newer sciences of item response theory (IRT) and computerised adaptive testing (CAT), widely used in educational testing, have demonstrated advantages for clinical outcome assessment.[7 8] IRT models relate characteristics of items and characteristics of individuals (such as extent of disability) to the probability of a positive response, which yields the most information for each person. CAT is a specific kind of computer based testing that asks questions emanating from larger, rather than smaller, item pools covering a wider range of difficulty, and is a more precise way to reduce questionnaire burden.[9–12] Modification of classic instruments that has generally involved attempts at shortening to reduce questionnaire burden or development of disease specific instruments are considered inadequate from the perspective of IRT and CAT technologies.

The PROMIS (Patient-Reported Outcomes Measurement Information System), part of the National Institutes of Health Roadmap (www.nihpromis.org) is charged with developing improved PROs applicable to all areas of chronic illness and involving several domains including ''physical function/disability'', which is reported here. PROMIS is the most ambitious approach yet to these issues. In simplest terms, PROMIS seeks to employ the best items in the best ways[11–14] with a focus on items that are most relevant to study endpoints in clinical trials and observational studies. Our group has the primary PROMIS responsibility for development of improved ''disability/physical function'' instruments. We approach this task from our roles as primary developers and disseminators of the two currently most widely used instruments.

Optimal instrument development requires item improvement, yet systematic approaches to advancement of improved items for physical function/disability remains to be developed.[10] Items need to have strong face validity; need to be sensitive to change, reliable, valid, and clearly understood by patients; need to include patient priorities and perceived clarity; and need to be well adapted for IRT and CAT uses. As part of PROMIS, we systematically developed a PROMIS preliminary core item bank of physical function/disability items with the goal of improving precision, and thus the value of resulting instruments. In this article we use the terms ''disability'' and ''physical function'' nearly interchangeably. The use of the two terms together indicate that disability is based on decrements from normal function to

Abbreviations: CAT, computerised adaptive testing; HAQ-DI, Disability Index of the Health Assessment Questionnaire; IRT, item response theory; PRO, patient reported outcome; PROMIS, Patient-Reported Outcomes Measurement Information System; RA, rheumatoid arthritis

severe impairment, whereas ''physical function'' may be considered a bipolar domain, ranging from functional abilities far above average to those far below. This project and the informed consent were approved by the Stanford University Administrative Panel on Human Subjects in Medical Research, and each patient gave written informed consent.

METHODS

Item selection process

To develop the item bank, we systematically identified extant physical function/disability items, assessed clarity, patient relevance, and information content. We conducted extensive literature reviews and internet searches for English language instruments that contained physical function/disability items. Items unrelated to physical function/disability, such as pain, fatigue, and quality of life, were eliminated. Each item was broken down into the following separable attributes:

- Context—for example, Because of your health …
- Stem—for example, Are you able to walk a block …
- Time frame—for example, Over the past week …
- Response categories or options—for example, Easily, with some difficulty, with much difficulty; unable to do

The 10 PF-10 items of the SF-36[15] and the 20 HAQ-DI items[4] were designated ''legacy'' or ''benchmark'' instruments since they represented instruments with the widest use and greatest acceptance.

Binning and winnowing

To identify and evaluate items for duplication and problems, all identified items were sorted into ''bins'' containing items with similar content, such as walking, dressing, or running errands. Each bin was then ''winnowed'' by eliminating items which were duplicated, narrowly applicable, confusing, unrelated to physical function/disability, containing multiple constructs, or for which there was a superior alternative.

Identification of subdomains

PROMIS has developed a hierarchy of health domains showing PROs at three or more levels.[16] The hierarchy is first separable into the primary domains of Physical, Psychological, and Social Health domains. At the second level ''Physical Health'' contains subdimensions including: Disability/Physical Function; Pain; Fatigue; and Other Symptoms. At the third level ''Disability/Physical Function'' is a multidimensional construct, where disability problems with the hands, for example, do not predict problems with walking.

The issue that arose, thus, was how many subdomains would be appropriate for measuring physical function/disability, where each subdomain is unidimensional—that is, assessed only one construct, such as walking. Review of published instruments revealed a wide variety of implicit or explicit physical function/disability hierarchies, commonly with 4–12 subdomains, but without an evident consensus. We approached this issue through expert consensus and by factor analyses, looking sequentially at different factors to find the number and nature of subdomains.

Patient input

We surveyed 1100 patients from four ARAMIS (Arthritis, Rheumatism, and Aging Medical Information System) patient cohorts, in 11 groups of 100 patients, with 25 each with RA, osteoarthritis, and two ageing cohorts with average ages of 70 and 80 years.[17–19] Each group was queried about 30 items, which included ''ringer'' items designed to be unclear, legacy items, and other items from the preliminary item pool. Imbedded comparisons tested different contexts, time periods, and response options against each other. Patient input was specifically sought for item clarity and importance. Two focus groups addressed issues of gaps and omissions in the items. Cognitive interviews by telephone explored the particulars of 50 problem items.

These procedures led to the development of the preliminary PROMIS core physical function item bank, with 204 items including the classic items, which has been undergoing field testing (summer 2006) in over 7000 subjects for IRT characteristics and development of CAT applications.

Sample size issues

''Noisy'' outcome measures require larger sample sizes than more precise ones, suggesting the possibility of major benefits from reducing the ''standard error of measurement'', a PROMIS goal. We examined the effects of varying the error terms and the effect sizes on sample size requirements for a clinical trial.[20–23]

RESULTS

Item banks and item winnowing

Of the 340 instruments identified, 165 contained 1860 physical function/disability items. There were a total of 71 bins, with the highest number of items relating to complex activities: n = 309 for walking and n = 133 for dressing and grooming. Many items were eliminated for the following reasons: redundancy (n = 562); narrow application—that is, did not apply to all (for example, ''Because of your diabetes'') (n = 444); lack of clarity (n = 123); vague or confusing (n = 206); inconsistency with the physical function/disability construct (n = 332); and other (n = 3). Confusing questions were often ''double barrelled'' where two or more dissimilar items were combined, for example, ''Do you beat your wife and kick your dog?'' Eight reviewers independently evaluated the original item bank. Differences were adjudicated by a panel of three experienced outcome assessment methodologists. This process yielded a preliminary core set of 190 items.

All retained items were rewritten, as none were identified as being ''ideal'' as originally written, due to presence of unusual response options, context, time frame, or item clarity. These item attributes were systematically studied to assist in rewriting. The original HAQ-DI and SF-10 items were not changed, and ''improved'' versions were developed as well.

Item characteristics

In a CAT application items are sequentially presented one at a time and optimally are kept similar in format. Thus we needed to achieve consensus on rules for such item nuances as formatting, uses of present or past tense, response options, and other features. Table 1 summarises results of examining the ordering of the ''response options'' from most severe to least severe or vice versa and the preferred number of response options. Response options ranged from 2 to 101 options, with analogue scales for the 11 and 101 point options. Most used a scale of difficulty, beginning with ''Without difficulty'' at one end and ending at ''Unable to do'' at the other. Listing the most negative response ''Unable to do'' to the right (or to the top on a vertical list) was far more frequent among the over 1500 items analysed; this does not necessarily mean that this is the ideal ordering, but it suggests that this format will be most familiar to both investigators and subjects. A related analysis found that the number of ''response options'' preferred by survey scientists as evidenced by their choice of response option sets in their own instruments was four or five. Reviewers agreed with these judgements on other grounds as well, citing on the

Table 1 Number of response options and preferred direction of response

No. of response options	Example	Preferred by patients (% (rank))	Preferred by authors (% (rank))	Most negative response on right	Most negative response on left/top
2	Y/N (0/1)	20 (2)	15 (3)	171	16
3	0,1,2	13 (4)	12 (4)	144	62
4	0,1,2,3	9 (6)	25 (2)	290	70
5	0,1,2,3,4	27 (1)	31 (1)	367	201
11	0–10	7 (7)	4 (7)	50	0
101	0–100	11 (5)	6 (6)	73	0

questionnaire burden of having too many options, the imprecision of having too few, the observation that scales act almost like continuous scales after four or five options, and that questionnaire burden is increased when subjects have to consider very fine distinctions.

Time frames

The time frames varied by item and by instrument. The most frequently used time frame, found in 52% (n = 975) of items, was the present, taking the form of ''no time frame given'' or the words ''now'' or ''today''. This was followed by: ''past week'' (21% of items); ''past month'' (15%); ''past two weeks'' (8%); ''past two days'' (3%). In cognitive debriefing interviews, patients liked the simplicity and flexibility of the present tense question.

Item clarity and importance

Table 2 shows the results of patient input regarding item clarity. There was a wide range in the percentage of respondents who found specific items unclear, from zero to over 70%. In general, items judged most clear used the present tense ''Are you able to'', avoided limiting context ''Because of your arthritis'', and contained a single construct ''Walk a block on level ground''. In contrast, items ranked as most important were the most rudimentary. For example, ''Dressing'' was rated much more important than ''Walking or jogging two miles''. Capability items (that is, ''Are you able to …?'', ''Can you …?'') were most commonly used as compared with past tense performance items (''Over the past week did you …?''). Performance items were rated unclear about twice as frequently as capability items.

HAQ-DI legacy items were found to be unclear by less than 6% of respondents, and this rate was reduced by 0.5 per cent when rewritten HAQ-DI items were tested. Legacy PF-10 items were unclear by about 12%, and the rewritten PF-10 items reduced the rate to less than 11%. Thus, there were clear differences between instruments with a moderate improvement in clarity after revision.

Contributors to clarity

Effects on clarity by varying item stems or response option sets were assessed. Table 3 shows a pooled analysis of four items (Bath and dress yourself; Climb several flights of stairs; Open a new milk carton; and Run errands and shop) ordered by decreasing clarity scores where the item stems and response option sets were held identical across the four items. Twelve minor variations in items were assessed, with the percentage of respondents rating the item ''unclear'' ranging from less than 8% to over 19%. Although it is difficult to account fully for multiple comparisons in such an analysis, the top two formulations were each statistically better than the best of the rest ($p<0.01$). In addition, the range from best to worst was a major 553 standard error units, easily surviving a Bonferroni or other statistical correction. Again, clarity was enhanced by use of the present tense and more conversational response option sets.

Table 2 Clarity scores (per cent unclear)

	N	Score (standard error)
Capability Items (Are you able to …?)	219	8.9 (0.43)
Performance Items (Did you …?)	48	15.8 (1.23)
HAQ-DI (Legacy)	20	5.8 (0.56)
HAQ-DI (Rewritten)	20	5.4 (0.68)
PF-10 (Legacy)	10	12.5 (1.60)
PF-10 (Rewritten)	10	10.8 (1.06)

Sample size issues

Table 4 shows the effects on sample size requirements for a clinical trial from utilization of more precise items. Increased clarity acts to decrease measurement error terms and to increase study power. The table shows one of many computations based on a population with a mean of 50 on a 0–100 unit scale, and the standard deviation (SD) set at 10 units, with varying assumptions. In all cases, effects were a function of the treatment effect and the standard error of measurement. For a true treatment effect of 5 (0.5 SD), for example, the number of subjects required for each arm of the trial with a measurement error of 8 is 143, and with a measurement error of 4 it is reduced to 83. Such reductions, of 25–40%, are projected to be easily achievable by use of item improvement, refinement with IRT analyses, and implementation with CAT. Much greater reductions will sometimes be achievable; a number of common rheumatological trial endpoints, including the sedimentation rate and the swollen joint count, have measurement error terms greater than 12 units.

Subdomains of physical function

Figure 1 shows confirmatory factor analysis data using a panel of about 7700 patients with a preliminary core physical function data set, as an empirical test of the postulated subdomain hierarchy within the physical function domain.[21] A four factor model gave the best fit, with factor loadings as shown. The first factor ''blindly'' identified upper extremity items. The second factor identified bending and twisting actions of the neck and back. The third identified walking and climbing items, and the fourth identified a group of more complex activities often termed ''instrumental activities of daily living''. Thus, factor analysis confirms the hypothesised minimal number of subdomains within physical function/disability. Of additional interest, in these populations of RA and osteoarthritis of the knee or hip, there was substantial differential item functioning by RA or osteoarthritis disease state; perhaps not clinically surprising but suggesting that differential item functioning across disease states is likely to be common.

DISCUSSION

Improved outcome assessment by PRO can substantially improve clinical research and make the research process more efficient. Clinical trials may require fewer subjects, and greater assurances may be given that the perspectives of the patient are included. The legacy instruments serve as the standard from which improvement may be measured. The goal is to construct better instruments by using better items in better ways. Better items may be obtained by parsing large item banks for items best understood and considered important to patients, reduction in floor and ceiling effects, improving and standardising time frame, context, and response options, and rewriting item stems to further

Table 3 Item clarity: analysis of pooled items with differing context and response options

	Respondents (n)	Per cent unclear	Standard error
Are you able to [bath and dress yourself, climb several flights of stairs, open a new milk carton, run errands and shop] (*Without any difficulty; With a little difficulty; With some difficulty; With much difficulty; Unable to do*)	258	7.75	0.017
Does your health now limit you in […] (*Not at all; Very little; Somewhat; Quite a lot; Cannot do*)	270	7.78	0.016
Over the past week, are you able to […] (*Without any difficulty; With some difficulty; With much difficulty; Unable to do*)	258	8.53	0.017
How difficult is it for you to […] (*Impossible; Very difficult; Difficult; Slightly difficult; Easy; Very easy*)	258	8.91	0.018
Due to my health […] (*Impossible; Very difficult; Difficult; Slightly difficult; Easy; Very easy*)	270	10.74	0.019
How much difficulty do you have […] (*None; A little; Some; Quite a lot; Cannot do*)	258	11.24	0.020
How easy is it for you to […] (*Very easy; Easy; Slightly difficult; Difficult; Very difficult; Impossible*)	258	11.63	0.020
How much does your health limit you in […] (*Not at all; Very little; Somewhat; Quite a lot; Cannot do*)	202	12.87	0.024
How much does your health now limit you in […] (*Not at all; Very little; Somewhat; Quite a lot; Cannot do*)	338	15.68	0.020
For me, [bathing and dressing myself, climbing several flights of stairs, opening a new milk carton, running errands and shopping] is… (*Very easy; Easy; Slightly difficult; Difficult; Very difficult; Impossible*)	258	15.12	0.022
Does your health now limit you in […]? If so how much? (*Yes, limited a lot; Yes, Limited a little; No, not limited at all*)	136	17.65	0.033
How much are you limited in […] (*Not at all; Very little; Somewhat; Quite a lot; Cannot do*)	258	19.38	0.025

improve nuances of item structure and wording. Further, IRT can quantitatively measure the information content achieved by each item, reject poorly performing items, and provide a means to select the best items for each patient; CAT applications can then administer the best items in the best ways.

We sought to define the ''latent trait'' (so-named because it cannot be directly observed but must be inferred from its attributes)[7] of physical function/disability through use of a modified Delphi approach to achieve expert consensus, examination of historical constructs, and assessment of patient views on importance and relevance. The consensus is that the latent trait is best termed ''physical function'', and that it consists of the ability to perform ''activities of daily living'' and ''instrumental activities of daily living''. Physical function, closely related to ''disability'', is a bipolar construct bounded by ''very easy'' and ''unable'', thus allowing a scale to pick up changes in already high levels of functioning. New therapies, it was felt, may improve function from ''nearly normal'' to ''better than average'', and these effects should be able to be estimated by improved instruments, thus removing ''gaps'' in coverage.

Table 4 Required sample size in each group for various levels of measurement precision

Expected treatment effect	Required sample size in each group						
	SEM = 12	SEM = 10	SEM = 8	SEM = 6	SEM = 4	SEM = 2	SEM = 0
	$\rho^2 = 0.36$	$\rho^2 = 0.45$	$\rho^2 = 0.56$	$\rho^2 = 0.69$	$\rho^2 = 0.84$	$\rho^2 = 0.95$	$\rho^2 = 1$
2	1513	1168	885	665	508	414	383
3	673	520	394	297	227	185	171
4	379	293	222	167	128	105	97
5	243	188	143	108	83	68	63
6	169	131	100	75	58	47	44
7	125	97	74	56	43	35	33
8	96	74	57	43	33	27	25
9	76	59	45	34	27	22	20
10	62	48	37	28	22	18	17
11	51	40	31	23	18	15	14
12	43	34	26	20	16	13	12

SEM, Standard error of the mean.

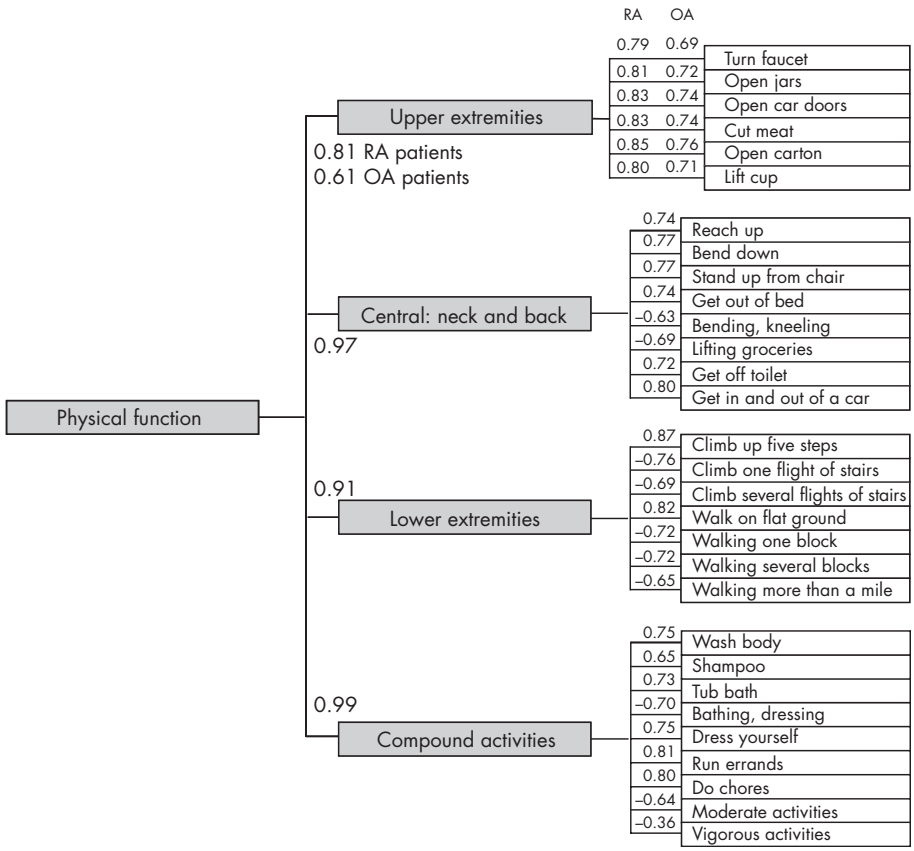

Figure 1 Hierarchical confirmatory factor analysis. A four factor model of physical function items provided a better fit than a one, two, or three factor model and recognised the four clusters shown, with high item loadings. Inspection of these clusters shows that they can be appropriately titled upper extremity, lower extremity, trunk, and compound activities, in agreement with the PROMIS postulated subdomains of ''physical function''. OA, osteoarthritis; RA, rheumatoid arthritis.

The latent trait ''able to do'' proved to be superior to ''performance of an activity'' as an endpoint for clinical trials. It better represented the desired outcome, had far fewer false positives, was clearer and preferred by patients, was more easily translatable (present tense), and had successful precedents as a clinical trial endpoint. (As a thought exercise, imagine your own responses to the questions: ''Are you able to run or jog two miles?'', and ''Over the past week did you run or jog two miles?'' Many will answer yes to the first and no to the second, and the yes is the closer approximation of the ''Are you able'' latent trait.) The consensus, however, also indicated that validation studies, where observed performance is compared with self-report, should be performed.

Limitations of these efforts are evident. The process of improving items is inevitably arbitrary to an extent, and qualitative judgements always leave room for future improvements. Assumptions, definitions, and standards are required, and all may not agree with the PROMIS consensus. However, the assumptions are documented in testable form, and studies to quantitatively compare old measures with new, including randomised controlled trials of static versus dynamic instruments for ability to detect treatment effects are underway.

''Noisy'' clinical trial endpoints are a threat to research validity and create inefficiencies. In this process it has become apparent that more reliable instruments increase efficiency and can reduce sample size requirements by 25-40%, and can reduce trial costs by a similar amount. This is increasingly recognised as a major part of the research efficiency goals of the National Institutes of Health Roadmap projects.

The new improved measures necessarily will outperform the old. Starting with improved items with demonstrated superiority in multiple areas will help with credibility, precision, and ability to detect changes. Use of CAT to dynamically administer the best items to each patient has been documented to obtain more precise estimates with any given number of questions. Remaining issues requiring careful study include valid use across a broad range of disease conditions and populations, the best methods of electronic administration, patient acceptability, and how to accelerate adoption by the Food and Drug Administration, industry, and academic specialty groups.

Authors' affiliations

J F Fries, B Bruce, Stanford University, Palo Alto, CA, USA
J Bjorner, M Rose, Health Assessment Lab, Waltham, MA, USA

This work was supported by an award from the National Institutes of Health to the PROMIS Roadmap Program, Stanford University Primary Research Site (AR52158).

Competing interests: none declared

Correspondence to: Dr J F Fries, Stanford University, 1000 Welch Road, Suite 203, Palo Alto, CA 94304; jff@stanford.edu

REFERENCES

1 **Fries JF**, Spitz P, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. *Arthritis Rheum* 1980;**23**:137–45.
2 **Brook RH**, Ware JE Jr, Davies-Avery A, Stewart AL, Donald CA, Rogers WH, *et al.* Overview of adult health measures fielded in Rand's health insurance study. *Med Care* 1979;**17**(7 suppl):iii–x, 1–131.
3 **Fries JF**. Toward an understanding of patient outcome measurement. *Arthritis Rheum* 1983;**26**:697–704.
4 **Bruce B**, Fries J. The Stanford health assessment questionnaire (HAQ): a review of its history, issues, progress, and documentation. *J Rheumatol* 2003;**30**:167–78.
5 *SF-36 Physical and Mental Health Summary Scales: A Manual for Users of Version I*, 2nd edn., Lincoln, RI: QualityMetric, 2001.
6 **Fries JF**, Spitz PW, Young DY. The dimensions of health outcomes: the health assessment questionnaire, disability and pain scales. *J Rheumatol* 1982;**9**:789–93.
7 **Fries JF**, Ramey DR. Platonic outcomes. *J Rheumatol* 1993;**20**:415–17.

8 **Ware JE Jr**, Kosinski M, Bjorner JB. Item banking and the improvement of health status measures. *Quality of Life* 2004;**2**:2–5.
9 **Ware JE Jr**, Kosinski M, Bjorner JB, Bayliss MS, Batenhorst A, Dahlof CG, *et al*. Applications of computerized adaptive testing (CAT) to the assessment of headache impact. *Qual Life Res* 2003;**12**:935–52.
10 **Cella D**, Lai J, AIB investigators. CORE item banking program: past, present, future. *Quality of Life* 2004;**2**:5–8.
11 **McHorney CA**, Cohen AS. Equating health status measures with item response theory: illustrations with functional status items. *Med Care* 2000;**38**(9 suppl):II43–59.
12 **Cella D**, Chang CH. A discussion of item response theory and its applications in health status assessment. *Med Care* 2000;**38**(9 suppl):II66–72.
13 **Fisher WP Jr**, Eubanks RL, Marier RL. Equating the MOS SF36 and the LSU HSI Physical Functioning Scales. *J Outcome Meas* 1997;**1**:329–62.
14 **Ware JE Jr**, Bjorner JB, Kosinski M. Practical implications of item response theory and computerized adaptive testing: a brief summary of ongoing studies of widely used headache impact scales. *Med Care* 2000;**38**(9 suppl):II73–82.
15 **Ware JE**, Snow KK, Kosinski M, Gandek B. *SF-36 Health Survey, Manual and Interpretation Guide*. Boston, MA: The Health Institute, New England Medical Center, 1993.
16 **Fries JF**, Bruce B, Cella D. The promise of PROMIS: using item response theory to improve assessment of patient-reported outcomes. *Clin Exp Rheumatol* 2005;**23**(5 suppl 39):S53–7.
17 **Bruce B**, Fries JF. The Arthritis, Rheumatism and Aging Medical Information System (ARAMIS): still young at 30 years. *Clin Exp Rheumatol* 2005;**23**(5 suppl 39):S163–7.
18 **Hubert HB**, Bloch DA, Oehlert JW, Fries JF. Lifestyle habits and compression of morbidity. *J Gerontol A Biol Sci Med Sci* 2002;**57**:M347–51.
19 **Bruce B**, Fries JF, Lubeck DP. Aerobic exercise and its impact on musculoskeletal pain in older adults: a 14 year prospective, longitudinal study. *Arthritis Res Ther* 2005;**7**:R1263–70.
20 **Kraemer HC**. To increase power in randomized clinical trials without increasing sample size. *Psychopharmacol Bull* 1991;**27**:217–24.
21 **Holman R**. How does item selection procedures affect power and sample size when using an item bank to measure health status? *Quality of Life* 2004;**2**:9–11.
22 **Kraemer HC**. *How Many Subjects? Statistical Power Analysis in Research*. Newbury Park, CA: Sage Publications, 1987.
23 **Raczek AE**, Ware JE, Bjorner JB, Gandek B, Haley SM, Aaronson NK, *et al*. Comparison of Rasch and summated rating scales constructed from SF-36 physical functioning items in seven countries: results from the IQOLA Project. International Quality of Life Assessment. *J Clin Epidemiol* 1998;**51**:1203–14.