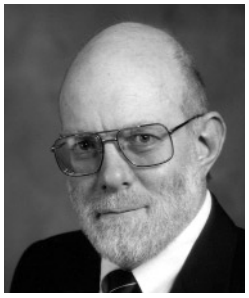
Editorial

The Psychometrics of Functional Status Questionnaires: Room for Improvement



The article by Eyres, *et al*[1] in this issue of *The Journal* should provoke 2 questions of interest to rheumatologists: (1) How good are our questionnaires? and (2) What do questionnaire results mean? If you don't think these are important questions, consider that all trials of new disease modifying antirheumatic drugs and tumor necrosis factor agents are built around questionnaire results, and that claims for improved function and quality of life are derived from these results. Maybe questionnaires like the Health Assessment Questionnaire (HAQ) or the Medical Outcome Study Short Form-36 (SF-36) seem simple to you. But if they do, let me ask you a few questions about them, a short pop quiz, to see if you have been doing your homework.

Describe the clinical findings and problems of a patient with rheumatoid arthritis (RA) whose SF-36 physical component score[2] is 30.4. How bad is a HAQ score[3] of 1.25? Is a HAQ score of 1.00 twice as bad as a score of 0.50? Is a Modified HAQ score[4] of 1 equal to a HAQ score of 1.25? To what extent do the HAQ, MHAQ, or SF-36 physical function scores capture the full range of disability in RA? Is the HAQ or the MHAQ better? Or the SF-36? Does it matter?

If you found that you could not answer these questions, you're not alone. Although widely used, questionnaires and their interpretation represent a pleasant mixture of the familiar, the impenetrable, and the incomprehensible.

To start with, there are 2 types of questionnaires: those that are multidimensional and those that are unidimensional. A unidimensional questionnaire is one that measures only one concept (dimension, construct), e.g., disability (or its inverse, function). Multidimensional questionnaires assess more than one domain, for example, disability and depression. The scores from multidimensional questionnaires can be reported as a summary of all or some of the dimensions (as in the summary SF-36 physical component scale) or as separate scores for each dimension (as in SF-36 scores for physical function, vitality, etc.). It always sounds as if you're getting a better questionnaire when you go for the multidimensional questionnaire. However, that frequently is not the case, for the unidimensional parts may be shortened to make a multidimensional questionnaire that is of acceptable length.

ENTER PSYCHOMETRICS

Using disability as an example, a good questionnaire should be scaled to identify the full range of disability, from slight functional loss to severe functional loss. It should also be accurately and linearly scaled so that any value on the scale (e.g., a HAQ score of 1.25) can be understood in terms of an absolute level of disability. This concept of a unidimensional, linear scale that captures a full range of the measured construct underlies (or should underlie) all questionnaire development. The article from the Leeds group[1] uses one tool to explore these psychometric issues, Rasch analysis. Rasch analysis has had some previous use in rheumatology[5–12], and is gaining wide recognition as a helpful tool for understanding current questionnaires and for modifying and developing new questionnaires[13-25]. At different levels of complexity, a number of texts and articles are available to help understand the Rasch method[13-31]. Rasch analysis can be key in transforming raw scores into those that represent linear, absolute measures of disability.

UNIDIMENSIONALITY

The most useful questionnaires are unidimensional, for they can provide single, linear measurement of constructs such as disability or, for example, depression or fatigue. Multidimensional summary scores are not interpretable unless there is some reference population or standard against which they can be compared. What makes the summary SF-36 physical and mental component scores so useful — that they can be used across different diseases as a *summary* measure of overall health status — is exactly what makes them of little practical use in rheumatology measurement and care,

See Measuring disability in ankylosing spondylitis: comparison of BASFI with revised Leeds Disability Questionnaire, *page 979*

where *specific* details are required. Therefore, (Question 1, above) an SF-36 score of 30.4 is uninterruptible practically, unless you think it is possible to add the incommensurable components of disability and affect. Multidimensional questionnaires, however, can often be broken down into component unidimensional parts. For practical interpretation, the Fibromyalgia Impact Questionnaire[6,32] and SF-36 function scores are more useful and interpretable than the larger multidimensional score.

In the design of a unidimensional questionnaire care must be taken not to include items from other dimensions. If we want to measure disability, we should not be measuring depression or athleticism. For example, in a disability/function questionnaire, activities such as running, jogging, or walking several miles tap into a dimension of athleticism rather than function, and there are people who simply don't run, jog, or walk several miles because they don't enjoy it. It's not easy to avoid this type of problem because it is often only in vigorous physical activities that the full range of function can be assessed. Questionnaires that have this problem include the SF-36 and the Multidimensional HAQ[17,33,34].

Allied to the issue of unidimensionality is differential item functioning (DIF), in which tasks are perceived as being more difficult in one group than in the other. For example, regardless of overall physical ability, jogging is perceived to be more difficult by older persons and vacuuming more difficult by women (who usually do more of it). If at all possible, questionnaires should have the least DIF possible, or for certain items of functioning it will be impossible to tell whether the results are an artifact of the questionnaire or a characteristic of the persons under study. DIF has one more important effect: it distorts the linear measurement scale for the groups. Rasch analysis provides an easy tool for the detection of DIF.

Non-unidimensional items and those with DIF add "noise," or error, to the measurement process. Another important source of error comes from items that are not clearly understood by respondents or not usually performed by them. "Taking a bath" (HAQ and WOMAC) is the most notorious of these items[6,9], but shampooing one's hair and performing sports are others.

Why is all of this important? Questionnaires with items that are not clearly understood or not within the experience of respondents — those with DIF, and those that are non-unidimensional — have reduced reliability and measurement accuracy. Simply, they don't perform as well as they could or they should.

LINEAR MEASURES

Figure 1 (based on 2229 patients with rheumatoid arthritis from the US National Data Bank for Rheumatic Diseases) displays the simultaneous distribution of HAQ scores (above) and the difficulties of the 8 HAQ categories (below) from a Rasch analysis. "Difficulty" can be thought of simply as a numeric score that indicates how difficult an item is to perform[27]. Because each HAQ category represents "none, mild, moderate, severe" (scaled 0–3), there are actually 24 divisions, each representing a different level of difficulty. The abilities of patients and the difficulties of the items are displayed on the same scale. Scored from –6 to +6, each division is perfectly linear. More functionally capable patients and more difficult items are at the left of the scale, and more impaired patients and easier items are at the right part of the scale. The mean difficulty of the items is 0 and the mean of HAQ scores is at –1.8 on the logit scale. Don't be scared of logits. Although they have precise mathematical meaning, they can be thought of simply as even divisions along the continuum of disability. Each HAQ score on the histogram below represents a 0.125 increment (0, 0.125,

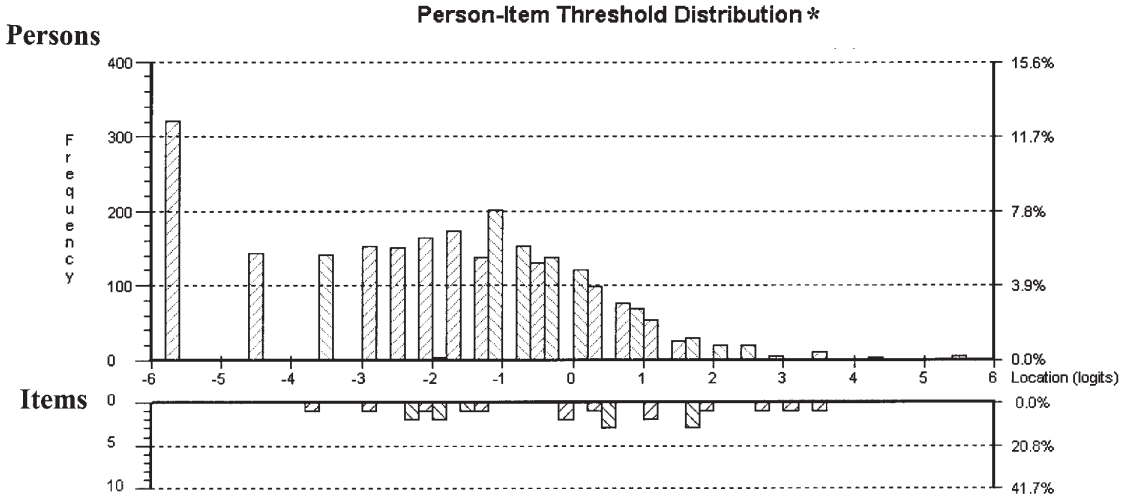


Figure 1. Person-item threshold analysis from 2229 RA patients in the National Data Bank for Rheumatic Diseases. *Grouping set to interval length of 0.20, making 60 groups.

0.250, 0.375, and so on). Notice that an increase in HAQ from 0 to 0.375 is a linear distance of 3 (logits). However, the same increase in HAQ score (from a HAQ of 1 to a HAQ of 1.375) is less than 1 logit unit. Practically, this means that changes in the HAQ score in the range from 1 to 2 represent much less change in function than changes in the HAQ score in the range from 0 to 1.

Does this have any practical meaning? You bet it does. A 0.25 or 0.5 change in the HAQ score at a HAQ score between 1 and 2 represents much less improvement than a similar change at a HAQ score below 1. What about the American College of Rheumatology 20% improvement criteria[35,36]? A 20% change can be seen to be less meaningful in the middle of the scale, where the mean HAQ score has been in almost all recent clinical trials. What about a "clinically meaningful" or "clinically important" change[37,38]? A HAQ score change of 0.25 is said to be clinically meaningful[37,38]. But how can that be, if such a change represents a different amount of change in function depending upon where the 1 is in the scale?

Figure 1 also indicates other deficiencies in the HAQ. It cannot reliably detect differences in patients below a HAQ score of 0.24 (–3.8 logits), and there is a big gap in the scale between approximately 0 and –1.

How can such problems be solved? By the proper choice of questionnaire items, it is possible to construct a questionnaire that will yield a linear, evenly spaced scale that addresses the wide range of functional abilities and impairments. Questionnaires and questionnaire development should be subject to the same rigorous rules of testing and validation that we apply to laboratory testing. The time is past for validation by correlation.

But psychometrics is not everything. One can build a perfect questionnaire only to find that it is useless in clinic practice and randomized trials. Questionnaires must be not only reliable and sensible, but also sensitive to change. Rasch analysis is a way to begin, but it is only that. A sensitive questionnaire that satisfies the Rasch model will always be better than one that does not. For all of its psychometric defects, the HAQ is a good, sensitive questionnaire, the best we have to date, and one that has stood the test of time.

FREDERICK WOLFE, MD,
National Data Bank for Rheumatic Diseases,
Arthritis Research Center Foundation,
and University of Kansas School of Medicine,
1035 N. Emporia, Suite 230,
Wichita, Kansas 67214, USA.
E-mail: fwolfe@arthritis-research.org

Address correspondence to Dr. Wolfe.

REFERENCES

1. Eyres S, Tennant A, Kay L, Waxman R, Helliwell PS. Measuring disability in ankylosing spondylitis: comparison of BASFI with revised Leeds Disability Questionnaire. J Rheumatol 2002;29:979-86.
2. Ware JE, Sherbourne CD. The MOS 36-Item Short-Form Health Survey (SF-36). 1. Conceptual framework and item selection. Med Care 1992;30:473-83.
3. Fries JF, Spitz PW, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. Arthritis Rheum 1980;23:137-45.
4. Pincus T, Summey JA, Soraci SA Jr, Wallston KA, Hummon NP. Assessment of patient satisfaction in activities of daily living using a modified Stanford Health Assessment Questionnaire. Arthritis Rheum 1983;26:1346-53.
5. Wolfe F. Which HAQ is best? A comparison of the HAQ, MHAQ and RA-HAQ, a difficult 8 item HAQ (DHAQ), and a rescored 20 item HAQ (HAQ20): Analyses in 2491 rheumatoid arthritis patients following leflunomide initiation. J Rheumatol 2001;28:982-9.
6. Wolfe F, Hawley DJ, Goldenberg DL, Russell IJ, Buskila D, Neumann L. The assessment of functional impairment in fibromyalgia: Rasch analyses of 5 functional scales and the development of the FM Health Assessment Questionnaire. J Rheumatol 2000;27:1989-99.
7. Wolfe F, van der Heijde DM, Larsen A. Assessing radiographic status of rheumatoid arthritis: introduction of a short erosion scale. J Rheumatol 2000;27:2090-9.
8. Ryser L, Wright BD, Aeschlimann A, Mariacher-Gehler S, Stucki G. A new look at the Western Ontario and McMaster Universities Osteoarthritis Index using Rasch analysis. Arthritis Care Res 1999;12:331-5.
9. Wolfe F, Kong SX. Rasch analysis of the Western Ontario MacMaster questionnaire (WOMAC) in 2205 patients with osteoarthritis, rheumatoid arthritis, and fibromyalgia. Ann Rheum Dis 1999;58:563-8.
10. Nordenskiold U, Grimby G, Hedberg M, Wright B, Linacre JM. The structure of an instrument for assessing the effects of assistive devices and altered working methods in women with rheumatoid arthritis. Arthritis Care Res 1996;9:358-67.
11. Stucki G, Daltroy L, Katz JN, Johannesson M, Liang MH. Interpretation of change scores in ordinal clinical scales and health status measures: the whole may not equal the sum of the parts. J Clin Epidemiol 1996;49:711-7.
12. Tennant A, Hillman M, Fear J, Pickering A, Chamberlain MA. Are we making the most of the Stanford Health Assessment Questionnaire? Br J Rheumatol 1996;35:574-8.
13. Sheehan TJ, DeChello LM, Garcia R, Fifield J, Rothfield N, Reisine S. Measuring disability: application of the Rasch model to activities of daily living (ADL/IADL). J Outcome Meas 2000;4:681-705.
14. Raczek AE, Ware JE, Bjorner JB, 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.
15. Prieto L, Alonso J, Ferrer M, Anto JM. Are results of the SF-36 health survey and the Nottingham Health Profile similar? A comparison in COPD patients. Quality of Life in COPD Study Group. J Clin Epidemiol 1997;50:463-73.
16. Prieto L, Alonso J, Lamarca R, Wright BD. Rasch measurement for reducing the items of the Nottingham Health Profile. J Outcome Meas 1998;2:285-301.
17. McHorney CA, Haley SM, Ware JE Jr. Evaluation of the MOS SF-36 Physical Functioning Scale (PF-10): II. Comparison of relative precision using Likert and Rasch scoring methods. J Clin Epidemiol 1997;50:451-61.
18. McHorney CA. Generic health measurement: Past accomplishments and a measurement paradigm for the 21st century. Ann Intern Med 1997;127:743-50.
19. MacKnight C, Rockwood K. Rasch analysis of the hierarchical assessment of balance and mobility (HABAM). J Clin Epidemiol

2000;53:1242-7.
20. Hopman-Rock M, Van Buuren S, Kleijn-De Vrankrijker M. Polytomous Rasch analysis as a tool for revision of the severity of disability code of the ICIDH. Disabil Rehabil 2000;22:363-71.
21. Granger CV, Deutsch A, Linn RT. Rasch analysis of the Functional Independence Measure (FIM) Mastery Test. Arch Phys Med Rehabil 1998;79:52-7.
22. Fisher WP Jr. Foundations for health status metrology: the stability of MOS SF-36 PF-10 calibrations across samples. J La State Med Soc 1999;151:566-78.
23. Custers JW, Hoijtink H, van der Net J, Helders PJ. Cultural differences in functional status measurement: analyses of person fit according to the Rasch model. Qual Life Res 2000;9:571-8.
24. Cook KF, Rabeneck L, Campbell CJ, Wray NP. Evaluation of a multidimensional measure of dyspepsia-related health for use in a randomized clinical trial. J Clin Epidemiol 1999;52:381-92.
25. Andresen EM. Criteria for assessing the tools of disability outcomes research. Arch Phys Med Rehabil 2000;81 Suppl 2: S15-20.
26. Andrich D. Rasch models for measurement. Newbury Park, CA: Sage Publications; 1988.
27. Bond TG, Fox CM. Applying the Rasch model: fundamental measurement in the human sciences. Mahwah, NJ: Erlbaum; 2001.
28. Fischer GH, Molenaar IW, editors. Rasch models: foundations, recent developments, and applications. New York: Springer; 1995.
29. Linacre JM, Wright BD. A user's guide to Bigsteps: Rasch model computer program. Version 2.8. Chicago: Mesa; 1997.
30. Linacre JM. Understanding Rasch measurement: estimation methods for Rasch measures. J Outcome Meas 1999;3:382-405.
31. Wright BD, Masters GN. Rating scale analysis: Rasch measurement. Chicago: Mesa; 1982.
32. Burckhardt CS, Clark SR, Bennett RM. The Fibromyalgia Impact Questionnaire: development and validation. J Rheumatol 1991;18:728-33.
33. Haley SM, McHorney CA, Ware JE Jr. Evaluation of the MOS SF-36 physical functioning scale (PF-10): I. Unidimensionality and reproducibility of the Rasch item scale. J Clin Epidemiol 1994;47:671-84.
34. Pincus T, Swearingen C, Wolfe F. Toward a multidimensional Health Assessment Questionnaire (MDHAQ) — Assessment of advanced activities of daily living and psychological status in the patient-friendly Health Assessment Questionnaire format. Arthritis Rheum 1999;42:2220-30.
35. Felson DT, Anderson JJ, Lange MLM, Wells G, LaValley MP. Should improvement in rheumatoid arthritis clinical trials be defined as fifty percent or seventy percent improvement in core set measures, rather than twenty percent? Arthritis Rheum 1998;41:1564-70.
36. Felson DT, Anderson JJ, Boers M, et al. American College of Rheumatology preliminary definition of improvement in rheumatoid arthritis. Arthritis Rheum 1995;38:727-35.
37. Redelmeier DA, Lorig K. Assessing the clinical importance of symptomatic improvements — an illustration in rheumatology. Arch Intern Med 1993;153:1337-42.
38. Wells GA, Tugwell P, Kraag GR, Baker PR, Groh J, Redelmeier DA. Minimum important difference between patients with rheumatoid arthritis: the patient's perspective. J Rheumatol 1993;20:557-60.