Balancing fidelity and practicality in short version musculoskeletal patient reported outcome measures

Philip Gabel[1], Brendan Burkett[1] and Michael Yelland[2]

[1]Centre for Healthy Activity Sport and Education (CHASE), University of the Sunshine Coast, Queensland, Australia
[2]Griffith University – Medical School, Logan Campus, Queensland, Australia

Background: Musculoskeletal patient reported outcome (PRO) measures are essential to clinical practice as they determine the status of a patient's health. While such measures are meant to improve the delivery of evidence based medicine, the design process often overlooks their clinical relevance and utility.

Discussion: The demand for greater efficiency through shorter, user friendly PRO measures is discussed. The aim is to reduce respondent, clinician and researcher burden while retaining precision instruments with validated psychometric properties. The reductive statistical methodology and techniques used to achieve these goals are highlighted. The general lack of important qualitative input from the clinician and patient in these processes is noted. This lack of input can be detrimental to the clinical practicality and cost efficiency of the final product. It may also raise problems of potential conflicts of interest. Four additional areas of practical concern with particular significance are outlined: questionnaire format, item number, time benefits and scoring methods.

Summary: It is important that developers of new and modified musculoskeletal PRO measures ensure that their instruments, whilst maximising the psychometric properties and methodological characteristics, satisfy the requirements of patient and clinician practicality while emphasising the essential principles of evidence based medicine.

Keywords: patient reported outcome measures, practicality, questionnaires, psychometrics, musculoskeletal

Background

The development and evaluation of patient reported outcome (PRO) measures designed for health status measurement has progressed since their introduction in the 1970s.[1–4] The use of PRO measures is universally accepted as a means of quantifying patients' perceptions of their health, impairment and functioning.[5–7] In the area of musculoskeletal medicine, PRO measures have progressed from generic to region specific measures, such as for the spine and upper and lower limbs.[8–10] The current progression refines what are commonly complicated and lengthy existing measures into PRO measures that are shorter and quicker to complete, yet still reflect the specificity of the condition or region under scrutiny.[3,11] However, the path used to accomplish these goals is not yet standardised as the methods employed and the conceptual purposes of different PRO measures vary widely.

In musculoskeletal medicine, PRO measures are increasingly used in clinical settings to quantify a patient's status and comply with evidence based practice.[12,13] Consequently, it is important to use standardised methodology to shorten existing measures

DOI 10.1179/174328809X452890

and develop new ones. The preferred tool is one that is valid but with greater practicality as this reduces the burden to both the patient and clinician which in turn increases compliance and reduces errors.[14] In order to produce clinically utile measures, the development and validation process of new and shortened PRO measures, including those using 'computerised adaptive testing' (CAT), must include a balance of qualitative reductive methodology and quantitative statistical and external evidence methods. Evidence based medicine is the 'integration of best research evidence with clinical expertise and patient values',[15] yet this often appears to be overlooked in the development of PRO measures. Consideration of a measure's practical clinical characteristics is essential.

The objective of this paper is to focus on inadequacies in existing methods used to develop, shorten and validate musculoskeletal PRO measures, specifically those that do not sufficiently involve qualitative input. The incorporation of expert clinician opinion and patient feedback into the design and item selection process is a critical requirement.

Discussion

Methods to develop shorter patient reported outcome measures

It is not an easy task to facilitate the clinical practicality of PRO measures. There is a need to reduce items within the existing validated measures while concurrently retaining or improving their precision and psychometric properties.[1,4,16,17] Such modifications risk the production of inappropriate measures that sacrifice validity, reliability, internal consistency and responsiveness.[3,14] However, developing new short form measures from first principles, without drawing on existing validated items, also has potential difficulties, particularly in determining the content validity and reliability of the new instrument.[17,18]

Quantitative approaches

Frequently it is only quantitative statistical approaches that are employed to shorten existing measures. These include methods such as classical test theory with a reliance on testing equi-discriminative item-total correlations, item analysis methodology such as factor analysis and item averaging.[6,11] However, these approaches may lack either an explicit ordered continuum of items or evidence of additivity of the scale data.[19] An alternative approach is based on the use of modern test theory including item response theory and Rasch modelling. These enable an examination of the scale item's hierarchical structure, one-dimensionality and additivity.[5,19,20] They also provide an indication of a possible objective response order of the items.[7,21–23]

Qualitative approaches

The qualitative item review approach is more subjective. It predominantly uses expert opinion, patient feedback and clinical consideration of the content of items prior to the use of statistical analysis and validation. This qualitative approach can include techniques such as concept-retention methodology,[11] item impact,[24,25] subjective item-pool reduction through binning and winnowing[6,18,26,27] and focus groups that can include both patients and clinicians.[22]

Practicality approaches

In the daily clinical settings with musculoskeletal injuries, time restraints are a reality. A PRO measure becomes impractical if it requires more than 2–3 minutes to complete and either more than 30 seconds or a computational aide to score. Such measures will not achieve acceptance and widespread clinical use.[8,9,28] To facilitate clinical acceptance Liang[29] proposed six critical factors that must be considered during development:

(i) the measure must have a patient self-administered format
(ii) be brief
(iii) be easily understood
(iv) be applicable across a variety of conditions
(v) a variety of levels of disease severity
(vi) if used to measure work related injury, it must be relevant to working populations.

Summary considerations

These quantitative, qualitative and practicality approaches are not mutually exclusive. However, items in shortened PRO measures are frequently chosen by only one method.[14] This problem is replicated across many health fields such as respiratory,[24] dental,[30] sexual[31] and mental[32] and also affects the areas of rehabilitation,[7] quality of life,[25] pain[2,7] and chronic illness.[33] Undoubtedly, musculoskeletal PRO measures demonstrate similar traits[11,34] and consequently, the use of either a quantitative or a qualitative item reduction approach leads to appreciably different instruments being developed.[24] A quantitative emphasis reflects a strong focus on psychometric aspects and provides items with a high statistical relevance. However, these may not have high face validity. Conversely, over reliance

on a qualitative approach will have the opposite effect. A balance between approaches is required.

Recommendations for study design and methodology

To reduce the potential for serious flaws in musculoskeletal PRO measures, five critical recommendations to standardise the shortening process were proposed by Coste in 1997[14] and have been subsequently supported in recent studies:[11,25,26]

(i) the choice of measure – carefully assess for item content, psychometric properties and level of measurement
(ii) assess whether the original instrument can be considered a 'gold standard' for the phenomenon of interest
(iii) when there is a gold standard, shorten the measure by focusing on optimal criterion validity and avoid separate tests of the part(s) and whole
(iv) when there is no gold standard, it is essential to use impartial and credible expert opinion assisted by statistical considerations, not a purely statistical consistency approach
(v) a validation study must always be performed within an appropriate independent sample including patient and clinician feedback for face validity and practicality.

Coste's approach supports a reductive or modification methodology with retention and integration of the three critical aspects of evidence based medicine – external evidence, clinical expertise and patient values.[15] All three are required, otherwise there is the risk that the determinants of external evidence alone, particularly quantitative statistical based reduction methods, will become dominant.[14]

Specific concerns for computerised adaptive testing (CAT) systems

The suitability of CAT based PRO measures, which use item response theory to enhance measurement precision and reduce respondent burden, needs consideration as these systems remove clinicians from the scoring process.[7,22,32] The item selection is typically statistically dependent[35] but also needs to undergo a rigorous qualitative process.[26] Although providing a promising and positive future,[22] the trend toward CAT based PRO measures also raises significant practical concerns for a measure that is technologically dependent. This is exacerbated in a world where many sectors and areas have neither the access nor capacity and financial resources to utilise such advances. There is also the potential for conflict of interest between research for the profit motive and the ethics of health services.[32,36] There remain several critical challenges in the areas of deliberation, debate, research completion and decision making that must be addressed before the adoption of any CAT initiatives, particularly in older persons.[2]

Practical areas of concern

The acceptability of a PRO measure to the patient and the clinician is of prime importance. This acceptability remains critical when the data banks of items that formulate existing PRO measures are modified to improve the assessment of a patient's status. This applies whether the shortening process or a CAT system is used. In either situation, the following aspects must be considered:

(i) format
(ii) item number
(iii) time benefits in completion and/or scoring
(iv) the scoring method.

Format is a critical aspect of a measure's design. Ergonomic approaches can be used to assist in preparation and simplification. This will result in PRO measures that appear inviting to both patient and clinician and are easy to complete and score without the loss of critical data. In the CAT situation, a single sentence overview or schematic of the subsequent process can ease patient apprehension and facilitate accuracy and compliance in the initial administration. An effective format for a PRO measure is the layout and scoring found in dichotomous or three-point Likert measures, such as the Roland Morris disability questionnaire[37] or the upper limb functional index.[9] In these measures, marked boxes help the respondent complete the questionnaire and at the same time simplify the clinicians' scoring task. By contrast, there is a risk of inaccurate and missing data in the completion and scoring phases[9] for measures that utilise a matrix approach, such as the back pain functional scale,[38] the upper extremity functional index[39] and the lower extremity functional scale[3,40] or measures that use multiple scoring methods within the one measure, such as the low-back SF-36 PF18.[20] If a measure is longer than a single page, as in the original version of the chiropractic designed questionnaires of the neck disability index[41] and spine functional rater index,[28] it may be at risk of invalid results due to the failure of respondents to simply turn the page.

The use of both descriptive and numeric scale markers within a measure's scaling format can simplify the response task and significantly reduce missing data. This in turn improves respondent and

clinician acceptance.[28] Clinical validation studies with clinician and respondent feedback must ensure that items are acceptable and that missing responses are kept to a minimum.[9]

Item number is critical. It is suggested that outcome measures with a format of 10-items have the highest clinical utility as they are the easiest to score.[28,41–43] However, to ensure cross-sectional reliability[44] and suitable internal consistency,[11,45] it may not be possible to follow a simplistic development approach based purely on an *a priori* decision of 10 items. Satisfactory analytical validation procedures are dependent on several factors. They must support face, content and construct validity and consider whether the constructs present are supportive of internal consistency and appropriate dimensionality as found through factor analysis.[8] The shortened versions of the SF-36 general health status measure, the SF-12 and SF-8, are perhaps the most widely known examples of PRO measure with reduced items and improved practicality that is both statistically sound and clinically accepted.[46,47] Even this significant reduction process has been considered too conservative by some authors and single-item measures for self-rated general health have been suggested to capture the same information.[48] Acceptance of a measure that is conceptually sound and practical will be compromised if it has only a small improvement in the psychometric properties. An example of this is the 18-item Roland Morris disability questionnaire which is rarely used as the original longer measure that performs the same task is preferred.[49–51] The number of items is crucial, but a balance between psychometric, practical and statistical aspects must be attained.[11,14,16,31]

Time advantages in completion and scoring must be assessed and considered to encourage clinical use. If there are only marginal improvements in completion time but scoring complexity is increased[20] then the continued burden of practical requirements for both patient and clinician will inhibit acceptance.[3] If there are variations in scoring systems or scales within the one measure,[20,46] then clinical utility and simplicity may be adversely affected.[9,11,22,46] This may apply to a measure that is scientifically sound but requires a computational aid for scoring; in this case the measure may become clinically impractical and inefficient.[9,29] Such PRO measures will serve only for research purposes where imputation is computer generated.[16] This problem can be overcome by integrated software systems,[16] including CAT systems,[4,22,52] however universal clinical technology, compatibility and access is not yet available. Hence, scoring design aspects of measures[9,28] and CAT systems[2,22,23] are critical during the development or modification phase. A balance must be established between statistical validity and practical implementation if a PRO measure is designed for the clinical setting.

Summary

The concept of shortening, modifying or combining existing musculoskeletal PRO measures through methodological processes to improve scientific properties, acceptance and wider application is a tempting and admirable objective for researchers and clinicians. Within this process the methodological and statistical requirements are often addressed and established in depth, but the practical and logistical characteristics of the final PRO measure are often overlooked. It is clear that the development of any new PRO measure must be subjected to appropriate statistical procedures to satisfy validity issues, but it must also embrace patient and clinician values as well as practicality and clinical utility. Modifications and refinements to valid and reliable PRO measures, including CAT systems, must find the balance between scientific standards, universality of access and clinician and patient requirements.

Acknowledgement

The authors wish to thank Jonathan Hill of Keele University for his assistance in editing of the final draft.

References

1 Garratt A, Schmidt L, Mackintosh A, Fitzpatrick R. Quality of life measurement: bibliographic study of patient assessed health outcome measures. *BMJ* 2002;**324**:1417.
2 McHorney CA. Ten recommendations for advancing patient-centered outcomes measurement for older persons. *Ann Int Med* 2003;**139**:403–9.
3 Stratford PW, Riddle DL. Assessing sensitivity to change: choosing the appropriate change coefficient. *Health Qual Life Outcomes* 2005;**3**(23):23.
4 Haley SM, Gandek B, Siebens H, Black-Schaffer RM, Sinclair SJ, Tao W, *et al*. Computerized adaptive testing for follow-up after discharge from inpatient rehabilitation: II. Participation outcomes. *Arch Phys Med Rehabil* 2008;**89**(2):275–83.
5 Gandek B, Sinclair SJ, Jette AM, Ware JEJ. Development and initial psychometric evaluation of the participation measure for post-acute care (PM-PAC). *Am J Phys Med Rehabil* 2007;**86**(1):57–71.
6 Kopec JA, Sayre EC, Davis AM, Badley EM, Abrahamowicz M, Sherlock L, *et al*. Assessment of health-related quality of life in arthritis: conceptualization and development of five item banks using item response theory. *Health Qual Life Outcomes* 2006;**4**:33.
7 Ware J, Gandek B, Sinclair S, Bjorner J. Item response theory and computerised adaptive testing: implications for outcome measurement in rehabilitation. *Rehabil Psychol* 2005;**50**(1):71–8.
8 Grotle M, Brox JI, Vollestad N. Functional status and disability questionnaires: what do they assess? A systematic review of back-specific outcome questionnaires. *Spine* 2005;**30**(1):130–40.

9 Gabel CP, Michener L, Burkett B, Neller A. The upper limb functional index (ULFI): development and determination of reliability, validity and responsiveness. *J Hand Ther* 2006;**19**(3): 328–49.
10 Rogers JC, Irrgang JJ. Measures of adult lower extremity function. *Arthritis Care Res* 2003;49(S5):S67–S84.
11 Beaton DE, Wright JG, Katz JN, Group UEC. Development of the QuickDASH: comparison of three item-reduction approaches. *J Bone Joint Surg Am* 2005;**87**(5):1038–46.
12 APA APA. Policy on outcome measures and treatment justification. Available from: http://apa.advsol.com.au/members/documents. Melbourne: Australian Physiotherapy Association, 2003.
13 Hart DL, Connolly JB. Pay-for-performance for physical therapy and occupational therapy: medicare part B services. Final Report. Grant No. 18-P-93066/9-01: Health & Human Services/Centers for Medicare & Medicaid Services. Knoxville, TN: FOTO – Focus on Therapeutic Outcomes, 2006.
14 Coste J, Guillemin F, Pouchot J, Fermanian J. Methodological approaches to shortening composite measurement scales. *J Clin Epidemiol* 1997;**50**(3):247–52.
15 Sackett D, Sharon E, Straus W, Richardson S, Rosenberg W, Haynes RB. Evidence-Based Medicine: How to Practice and Teach EBM. New York, NY: Churchill Livingstone, 2000.
16 Prieto L, Alonso J, Lamarca R. Classical test theory versus Rasch analysis for quality of life questionnaire reduction. *Health Qual Life Outcomes* 2003;**1**:27.
17 Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. 3rd ed. Oxford: Oxford University Press, 2003.
18 Kirshner B, Guyatt GH. A methodological framework for assessing health indices. *J Chronic Dis* 1985;**38**(1):27–36.
19 Prieto L, Thorsen H, Juul K. Development and validation of a quality of life questionnaire for patients with colostomy or ileostomy. *Health Qual Life Outcomes* 2005;**3**:62.
20 Davidson M, Keating JL, Eyres S. A low back-specific version of the SF-36 physical functioning scale. *Spine* 2004;**29**(5):586–94.
21 Fries JF, Bruce B, Bjorner J, Rose M. More relevant, precise, and efficient items for assessment of physical function and disability: moving beyond the classic instruments. *Ann Rheum Dis* 2006;**65**(Suppl 3:iii):16–21.
22 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.
23 Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. *Med Care* 2000;**38**(9 Suppl):II28–42.
24 Juniper EF, Guyatt GH, Streiner DL, King DR. Clinical impact versus factor analysis for quality of life questionnaire construction. *J Clin Epidemiol* 1997;**50**(3):233–8.
25 Moran LA, Guyatt GH, Norman GR. Establishing the minimal number of items for a responsive, valid, health-related quality of life instrument. *J Clin Epidemiol* 2001;**54**(6):571–9.
26 DeWalt DA, Rothrock N, Yount S, Stone AA, Group obotPC. Evaluation of item candidates: the PROMIS qualitative item review. *Med Care* 2007;**45**(5):S12–S21.
27 Reeve B, Hays RD, Bjorner JB, Cook KF, Crane PK, Teresi JA, *et al*. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the patient-reported outcomes measurement information system (PROMIS). *Med Care* 2007;**45**(5 Suppl 1):S22–S31.
28 Feise RJ, Menke JM. Functional rating index. A new valid and reliable instrument to measure the magnitude of clinical change in spinal conditions. *Spine* 2001;**26**(1):78–86.
29 Liang MH, Jette AM. Measuring functional ability in chronic arthritis: a critical review. *Arthritis Rheum* 1981;**24**:80–6.
30 Locker D, Allen PF. Developing short-form measures of oral health-related quality of life. *J Public Health Dent* 2002;**62**(1):13–20.
31 Badia X, Colombo JA, Lara N, Llorens MA, Olmos L, Sainz de los Terreros M, *et al*. Combination of qualitative and quantitative methods for developing a new health related quality of life measure for patients with anogenital warts. *Health Qual Life Outcomes* 2005;**3**(1):24.
32 Walter OB, Becker J, Bjorner JB, Fliege H, Klapp BF, Rose M. Development and evaluation of a computer adaptive test for 'Anxiety' (Anxiety-CAT). *Qual Life Res* 2007;**16**(Suppl 1):143–55.
33 Webster K, Cella D, Yost K. The functional assessment of chronic illness therapy (FACIT) measurement system: properties, applications, and interpretation. *Health Qual Life Outcomes* 2003;**1**:79.
34 Hart DL, Mioduski JE, Werneke MW, Stratford PW. Simulated computerized adaptive test for patients with lumbar spine impairments was efficient and produced valid measures of function. *J Clin Epidemiol* 2006;**59**(9):947–56.
35 Hays RD, Liu H, Spritzer K, Cella D. Item response theory analyses of physical functioning items in the medical outcomes study. *Med Care* 2007;45(5 Suppl 1):S32–S38.
36 Ader D. Developing the patient-reported outcomes measurement information system (PROMIS). *Med Care* 2007;**45**(5):S1–2.
37 Roland M, Fairbank JCT. The Roland-Morris disability questionnaire and the Oswestry disability questionnaire. *Spine* 2000;**25**(24):3115–24.
38 Stratford PW, Binkley JM, Riddle DL. Development and initial validation of the back pain functional scale. *Spine* 2000;**25**(16): 2095–102.
39 Stratford PW, Binkley JM, Stratford D. Development and initial validation of the upper extremity functional index. *Physiother Can* 2001;**52**:259–67, 81.
40 Binkley JM, Stratford PW, Lott SA, Riddle DL. The lower extremity functional scale (LEFS): scale development, measurement properties, and clinical application. *Phys Ther* 1999;**79**(4): 371–83.
41 Vernon H, Mior S. The neck disability index: a study of reliability and validity. *J Manipulative Physiol Ther* 1991;**14**(7):409–15.
42 Fairbank JCT, Pynsent PB. The Oswestry disability index. *Spine* 2000;**25**(22):2940–52.
43 Michener LA, Leggins BG. A review of self-report scales for the assessment of functional limitation and disability of the shoulder. *J Hand Ther* 2001;**14**:68–76.
44 Nunnally JC, Bernstein IH. Psychometric Theory. New York, NY: McGraw-Hill, 1994.
45 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.
46 Ware J. SF-36 health survey update. *Spine* 2000;**24**(24):3130–9.
47 Ware J. Patient-based assessment: tools for monitoring and improving healthcare outcomes. *Behav Healthc Tomorrow* 1996;**5**(3):87–8.
48 DeSalvo KB, Fisher WP, Tran K, Bloser N, Merrill W, Peabody J. Assessing measurement properties of two single-item general health measures. *Qual Life Res* 2006;**15**(2):191–201.
49 Chansirinukor W, Maher CG, Latimer J, Hush J. Comparison of the functional rating index and the 18-item Roland-Morris disability questionnaire: responsiveness and reliability. *Spine* 2005;**30**(1):141–5.
50 Stratford P, Binkley J. Measurement properties of the RM-18: a modified version of the Roland-Morris disability scale. *Spine* 1997;**22**:2416–21.
51 Ostelo RW, de Vet HC, Knol DL, van den Brandt PA. 24-item Roland-Morris disability questionnaire was preferred out of six functional status questionnaires for post-lumbar disc surgery. *J Clin Epidemiol* 2004;**57**(3):268–76.
52 Cook K, O'Malley KJ, Roddey TS. Dynamic assessment of health outcomes: time to let the CAT out of the bag? *Health Serv Res* 2005;**40**(5-Pt 2):1694–711.

PHILIP GABEL
Centre for Healthy Activity Sport and Education (CHASE), University of the Sunshine Coast, Queensland, Australia
E-mail: cp.gabel@bigpond.com