Supplementary Material: Online Resource 1

Article Title: Investigating the Patient Acceptable Symptom State cut-offs: longitudinal data from a community cohort using the Shoulder Pain and Disability Index

Journal: Rheumatology International

Authors: Gui Tran[1, 2], Bright Dube[1, 2], Sarah R Kingsbury[1, 2], AlanTennant[1], Philip Conaghan[1, 2, 3], Elizabeth M.A. Hensor[1, 2]

Affiliations: [1]Leeds Institute of Rheumatic and Musculoskeletal Medicine, University of Leeds, Leeds, UK

[2]NIHR Leeds Biomedical Research Centre, Leeds Teaching Hospitals NHS Trust, Leeds, UK

[3] Centre for Sport, Exercise and Osteoarthritis Research, Versus Arthritis, Nottingham, UK

CORRESPONDING AUTHOR:

Dr Elizabeth Hensor,

Leeds Institute of Rheumatic and Musculoskeletal Medicine,

2nd Floor Chapel Allerton Hospital, Chapeltown Rd,

Leeds, LS7 4SA, UK

Phone: +44 (0)113 3924883

Fax: +44 (0)113 3924991

Email: E.M.A.Hensor@leeds.ac.uk

Online Resource 1. The Rasch model and assessment of model fit

Supplementary methods

The data were tested for fit to the partial credit parametrisation of the polytomous Rasch model. The partial credit model allows the difference in the latent degree of shoulder impairment between response categories (for example between a response of 1 and a response of 2 on a 0-10 SPADI item) to differ across items (for example between 'pain at its worst' and 'pain when lying on the involved side'). This is in contrast to the rating scale model, which requires the distance between score categories to be the same for all items. A highly significant likelihood ratio test comparing the two (p<0.001) supported the use of the partial credit model.

A testlet approach was used to assess fit of the total combined score to the Rasch model, in which items relating to pain (n=5) and disability (n=8) were first summed within each subscale to give two testlets. This approach allows issues such as local dependency and reciprocal differential item functioning (DIF) among items within a testlet, which can be complex to address on an item-by-item basis, to be absorbed before testing the overall fit of the total scale to the Rasch model.

Fit to the Rasch model was assessed in terms of the total item-trait interaction Chi-squared (for which the P value was required to be >0.05), high person separation index and Cronbach's alpha (which were required to be at least 0.7), small fit residuals (absolute values of which should not exceed 2.5), unidimensionality, a lack of residual correlation among items, a lack of differential item functioning, and a paucity of patients with extreme scores.

The standard error of measurement (SEM) for the SPADI was calculated as:

$$SD \times \sqrt{(1 - Cronbach's\ \alpha)}$$

The smallest detectable between-patient difference (SDD) was calculated as: $SEM \times 1.96$

The smallest detectable within-patient change (SDC) was calculated as: $SEM \times 1.96 \times \sqrt{2}$

Rasch model fit

The SPADI was found to fit the partial credit Rasch model (total item-trait interaction Chi-square=10.95, p=0.896; person separation index=0.87; Cronbach's alpha=0.88; individual testlet fit Chi-square pain p=0.96, disability p=0.56; acceptable fit residual for pain -2.35, slightly large fit residual for disability -2.67; person location mean=0.09, SD=0.42; person fit residual mean=-0.54, SD=0.84; only 4 patients of 492 in calibration sample had extreme scores).

Comparing estimates before and after the creation of the two testlets showed that 7% of the unique variance had been discarded to create a unidimensional latent estimate for the total SPADI. There was no evidence of remaining multidimensionality (2.83% of t-tests significant at 5% level). There was no evidence of differential item functioning (DIF), i.e. patients with the same underlying (latent) degree of shoulder impairment scoring differently according to patient characteristics such as age or sex, or over time. To assess for substantive DIF we used a criterion of effect size >0.2 for the paired comparison of person estimates obtained before and after splitting an offending item for DIF i.e. allowing threshold locations for the item to differ between (for example) males and females. If the effect size were within 0.2 we concluded that the DIF was trivial in extent, even if statistically significant, and could be ignored.