An evaluation of the structural validity of the Shoulder Pain and Disability Index (SPADI) using the Rasch model

Christina Jerosch-Herold[1]
Rachel Chester[1]
Lee Shepstone[2]
Joshua I. Vincent[3, 4]
Joy C. MacDermid[3, 5]

[1] School of Health Sciences, Faculty of Medicine and Health Sciences, University of East Anglia, Norwich NR4 7TJ, Norfolk, United Kingdom. ORCID ID 0000-0003-0525-1282

[2] Norwich Medical School, Faculty of Medicine and Health Sciences, University of East Anglia, Norwich NR4 7TJ, Norfolk, United Kingdom.

[3] School of Rehabilitation Sciences, Faculty of Health Sciences, McMaster University, Hamilton Ontario L8S 4L8, Canada

[4]Lifemark Physiotherapy, London, Ontario N6C 4Y7, Canada

[5] School of Physical Therapy, Faculty of Health Sciences, Western University, London, Ontario N6A 3K7, Canada

Corresponding author:
Christina Jerosch-Herold, School of Health Sciences, Faculty of Medicine and Health Sciences, University of East Anglia, Norwich NR4 7TJ, UK, email: c.jerosch-herold@uea.ac.uk, Tel +44 (0)1603 593316

Funding:
CJH and RC were funded by the National Institute for Health Research (NIHR Senior Research Fellowship and NIHR Clinical Doctoral Research Fellowship, respectively). The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The views and opinions expressed therein are those of the authors and do not necessarily reflect those of the NIHR, NHS or the Department of Health. The authors certify that they have no affiliations with or financial involvement in any organisation or entity with a direct financial interest in the subject matter or materials discussed in the article.

Compliance with ethical standards

Conflict of Interest: the authors declare that they have no conflicts of interest

Ethical Approval:
This paper is based on a secondary analysis of data. The original study was approved by

the National Research Ethics Service, East of England - Norfolk, UK, July 2011 (reference 11/EE/0212). All procedures performed in the study involving human participants were in accordance with the ethical standards of the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

Abstract

Purpose:
The Shoulder Pain and Disability Index (SPADI) has been extensively evaluated for its psychometric properties using classic test theory (CTT). The purpose of this study was to evaluate its structural validity using Rasch model analysis.

Methods:
Responses to the SPADI from 1030 patients referred for physiotherapy with shoulder pain and enrolled in a prospective cohort study were available for Rasch model analysis. Overall fit, individual person and item fit, response format, dependence, unidimensionality, targeting, reliability and differential item functioning (DIF) were examined.

Results:
The SPADI pain subscale initially demonstrated a misfit due to DIF by age and gender. After iterative analysis it showed good fit to the Rasch model with acceptable targeting and unidimensionality (overall fit (chi-square statistic 57.2, p=0.1); mean item fit residual 0.19 (1.5) and mean person fit residual 0.44 (1.1); person separation index (PSI) of 0.83). The disability subscale however shows significant misfit due to uniform DIF even after iterative analyses were used to explore different solutions to the sources of misfit (overall fit (chi-square statistic 57.2, p=0.1); mean item fit residual -0.54 (1.26) and mean person fit residual -0.38 (1.0); PSI 0.84).

Conclusions:
Rasch Model analysis of the SPADI has identified some strengths and limitations not previously observed using CTT methods. The SPADI should be treated as two separate subscales. The SPADI is a widely used outcome measure in clinical practice and research, however the scores derived from it must be interpreted with caution. The pain subscale fits the Rasch model expectations well. The disability subscale does not fit the Rasch model and its current format does not meet the criteria for true interval-level measurement required for use as a primary endpoint in clinical trials. Clinicians should therefore exercise caution when interpreting score changes on the disability subscale and attempt to compare their scores to age and sex stratified data.

INTRODUCTION

Musculoskeletal shoulder disorders are common in the general population (1, 2). Outcomes of interventions for shoulder disorders can be measured through clinician-derived assessment as well as patient-reported outcome measures (PROM). A systematic review of outcomes reporting in shoulder disorders identified pain, range of motion and function as the most commonly assessed domains (3). Recently developed consensus on a preliminary core outcome set for shoulder disorders has identified pain and physical function/activity as two core outcome domains (4). There are several PROMS which capture pain and physical function/activity. One such instrument is the Shoulder Pain and Disability Index (SPADI) (5). First developed by Roach to capture pain and activity limitations in shoulder disorders, the SPADI is made up of a 5-item pain subscale and 8-item disability subscale. The original SPADI used a visual analogue scale which was subsequently converted to a numerical rating scale (5) (6) where each item is scored on an 11-point ordinal rating scale ranging from 0 (no pain/no difficulty) to 10 (worst imaginable pain/so difficult that it requires help). There is some discrepancy in the literature regarding the calculation of the total score with some advocating summing all 13 items and dividing by 13 and others calculating an average subscale score for the five pain and eight disability items separately which is then averaged (5) (7). The latter method gives equal weight to each subscale.

The SPADI is short, easy to complete and score and has been widely adopted and recommended for clinical practice and research (6, 8, 9). It has been studied in multiple contexts for its validity, reliability and responsiveness (6-13); using classical test theory (CTT). However, recent developments in the field of psychometrics and the scientific requirements for use of PROMs in clinical trials have highlighted that ordinal rating scales often do not meet the criteria for true interval-level measurement (14, 15). Rasch analysis (16) is a relatively new approach based on item-response theory (IRT), which is increasingly used, both in the development of new and the testing of existing PROMs. Rasch analysis uses a 1 parameter logistic regression model, also called the Rasch model, to test the fit between the observed data (the patients' responses on a scale) and the expected responses from the Rasch model. If the data fit the Rasch model then the data can be used to transform ordinal level scales into true interval-level measures (17). It is particularly useful to assess the structural validity of a questionnaire which is an aspect of construct validity relevant in multi-item health-related PROMs. It is defined as the degree to which the scores from a scale are an adequate reflection of the dimensionality of the construct being measured (18). Rasch analysis can identify several strengths and weaknesses in a rating scale. These include: whether the scores produce true interval-level measures (19); whether the items in the scale measure a single construct (unidimensionality); whether items are locally independent; whether items map onto a hierarchical order of increasing difficulty and finally whether the scale is invariant, meaning that responses are reflective of the latent trait and not dependent on any other factors such as age or gender (20). To date only one study has applied Rasch model analysis to the SPADI disability subscale (21) and not to the pain subscale. Using BIGSTEPS software the overall fit calibration of four shoulder function scales was assessed including the 8-item SPADI disability subscale ~~only~~. The authors did not examine response thresholds, item dependence or response bias.

The purpose of this study therefore was two-fold: i) to analyse the SPADI as a full scale; and ii) to analyse its subscales separately with regards to evaluating the structural validity, in particular how well the scale or its subscales are targeted to patients with musculoskeletal shoulder pain, their response format, whether they shows response bias, their dimensionality, and to explore whether scale modifications are needed to improve its fit to the Rasch model.

METHODS

The data for the Rasch analysis came from a large, prospective, multi-centre cohort study of patients referred to physiotherapy for treatment of shoulder pain at 11 primary or secondary care providers in the East of England. The full protocol and results have been described elsewhere (22, 23). Patients were asked to complete a questionnaire booklet which included the SPADI. This analysis is based on the initial SPADI score collected at baseline and prior to treatment when shoulder-related pain and disability are most likely to be present. Patients were included in the study if they had musculoskeletal shoulder pain of any duration, were aged 18 years or older and had not undergone surgery in the previous 5 years for shoulder pain. Patients with fractures, dislocation, radiculopathy and other underlying systemic conditions causing referred shoulder pain were excluded.

Rasch analysis

All analyses were performed in RUMM2030[1] software for Windows 7. As the SPADI is a polytomous scale (uses two or more ordered response categories) two models can be used; the partial credit model and the rating scale model (24). We followed recommendations by Lundgren Nilsson and Tennant for Rasch analysis of polytomous scales (25) and chose a partial credit model based on a highly statistically significant Likelihood-ratio-Test ($p<0.001$). This assumption was also re-tested at subscale level.

Test of fit:

Overall fit with the Rasch model was assessed by examining the extent to which the patients' responses correspond to the expectations of the Rasch model. The total item-trait chi-square statistic was used to assess overall fit where a statistically significant result $p<0.05$ indicates misfit (26). Individual item and person standardised fit residuals were also examined. These summarise the difference between observed and expected values (item-person interaction) Individual Item and person fit residuals localised within ±2.5 logits were considered as fitting the model (26).

Threshold order:

[1] RUMM Laboratory Pty Ltd, Perth

Category probability curves were used to examine how well the response categories were used by respondents and to identify disordered thresholds, which can be a source of misfit of items (26). A common reason for disordered thresholds is that participants cannot consistently discriminate between the available response options (25). Potential solutions include collapsing adjacent categories by rescoring the responses with fewer categories and to examine if this improves the overall fit of the model (26).

Unidimensionality:

To test the assumption of unidimensionality, a prerequisite to summing items into a total score, the principal components of the residuals were examined. Items with positive and negative loadings >0.3 on the first component are used to make up two subsets. Independent t-tests are then used to compare the person estimates on each subset. If the number of significant t-tests exceeds 5% the scale is deemed to exhibit multidimensionality (27).

Local dependency:

Dependency occurs when items either duplicate each other (redundancy) or they share some other underlying trait which may also contribute to multidimensionality (25). Residual correlations between any two items greater than 0.2 were used to indicate where any two items may be locally dependent (25). Local dependence can artificially inflate reliability (28) A possible solution to dependency is to create so-called 'testlets' where two or more locally dependent items are combined generating a summary score or 'super item' (29).

Differential item functioning (DIF):

Finally we examined whether responses differ by some other factor or variable called differential item functioning (DIF) (30). Where differences are consistent across groups, for example gender, this is termed uniform DIF and can be adjusted for by splitting items thus generating separate location estimates for men and women. Non-uniform DIF arises from random differences and cannot be resolved. We examined the item characteristic curves (ICC) visually for DIF by two person factors: gender (men; women) and age (≤ 59 years; 60 and above) and used an ANOVA test to assess statistically significant uniform and non-uniform DIF. P-values were adjusted for multiple comparisons (31). The cut-off for age was determined a priori. Median age was 58 years and a cut off at ≤ 59 years and 60 years or over produced similar sized subgroups.

Targeting:

In clinical practice it is important that the outcome measure used is appropriate for the population, referred to as targeting. The person-item threshold distribution was examined graphically and assessed statistically using summary statistics for item and person fit. Ideal mean values should be close to 0 with a SD not exceeding 1 (28). Distribution of responses across the available categories were also examined for any evidence of floor or ceiling effects.

Reliability index:

A scale's ability to discriminate between respondents is expressed as a Person Separation Index (PSI) and is used as an alternative to Cronbach alpha (26). A PSI of 0.7 is considered the lowest level acceptable and indicates that the scale can statistically discriminate between at least two groups. A higher PSI indicates greater reliability with values greater than 0.8 indicating that the scale can discriminate between at least three groups and 0.9 between four or more groups (32).

RESULTS

Baseline SPADI data were available on 1030 patients with musculoskeletal shoulder pain (mean total SPADI score= 48, SD=22) who had not been treated surgically. The mean duration of shoulder symptoms was 14 months (SD=28 months) and mean pain at rest was 3 points on a 0-10 numerical rating scale. Mean age was 57 years (SD=15) and 54% of patients were women. Mean body mass index was 27 (SD=5) and 13% were smokers. Less than 2% reported being off work due to the shoulder problem, although 12% had taken some time off work in the previous year. Fifty-eight percent of participants were in some form of employment or education and 36% of participants were retired.

Rasch analysis of full SPADI:

The initial analysis of the 13-item SPADI revealed significant misfit to the RASCH model (χ^2 = 301.7; $p<0.00001$) (Table 1, analysis stage 1). Dimensionality of all 13-items together was assessed by Principal Component Analysis (PCA) of the residuals. High negative loadings (>0.3) were seen for four out of the five items in the pain subscale and high positive loadings (>0.3) for six of the eight items on the disability subscale. T-tests between the two subsets of negatively and positively loaded items identified 10.1% of t-tests to be significant meaning it does not meet the assumption of a unidimensional scale. Given the existing evidence from exploratory factor analysis of the full SPADI which indicates that it is bidimensional (5, 6, 12, 33), all subsequent analyses were done by subscales with P1 to P5 making up the pain subscale and items D1 to D8 the disability subscale.

Rasch analysis of the SPADI-Pain subscale:

Initial analysis of the 5-item pain subscale revealed overall good fit to the Rasch Model. The total item-trait chi-square statistic was not significant (see table 1, analysis 2) and response thresholds for all five items were ordered. Pain at worst (P2) showed some misfit (fit residual +2.74). On closer inspection individual person fit statistics revealed 9 persons for whom fit residuals could not be calculated. They all endorsed the highest category (score of 10) for all 5 items. A further 8 persons had fit residuals +2.5 on 2 items. Given the large sample size these 17 persons were deleted from the analysis which improved overall fit (chi-square statistic 57.2, p=0.1) and individual item fit statistics (Table 2).

PCA of the residuals and equating t-tests as described previously confirmed a unidimensional subscale and there was no local dependence (any correlations >0.2) between items.

We examined the pain subscale for differential item functioning and found statistically significant uniform DIF by age for P1 'pain at worst' ($p<0.001$) and by gender for P5 'pain when pushing with involved arm' ($p<0.001$). Figure 1 shows the item characteristic curves for these two items. For P1 'pain at worst' people aged ≤59 years had slightly higher pain (mean 7.37) than people aged 60 or over (mean 7.02). For P5 'pain when pushing with the involved arm' women report higher pain (mean 5.36) than men (mean 4.32).

Mean item fit residual was 0.19 (SD=1.5) and mean person fit residual was 0.44 (SD=1.1). All items had fit residuals within the ±2.5 threshold.

Targeting of the items to persons was good (see Figure 2) however it provides limited information at the extremes of the sample distribution (highlighted by different width arrows) especially for those with higher pain.

Reliability was high with a person separation index (PSI) of 0.83 indicating that the scale can discriminate statistically between three or more groups.

Rasch analysis of SPADI-Disability subscale:

Initial analysis of the 8-item disability subscale showed considerable misfit to the Rasch model (see Table 1, analysis 4). Sources of misfit were explored by examining response thresholds, item-to-item correlations for dependence, and differential item functioning.

Response thresholds were disordered for three items: 'washing your hair' (D1), 'putting on a shirt that buttons at front' (D4) and 'putting on trousers' (D5) indicating that patients cannot adequately discriminate between the 11 response options. Three items had fit residuals outside the ±2.5 threshold and a significant chi-square probability ($p<0.0001$). These were D1 'washing hair (-2.56); D3 'putting on a jumper' (fit residual -3.61) and D7 'carry a heavy object' (fit residual +7.2). As described for the pain subscale, 'misfitting' persons were identified and 16 deleted from the analysis (Table 1, analysis 5).

A number of rescoring options were explored to obtain ordered thresholds. Using a rescore of 00112233445 for all 8 items achieved an ordered threshold map (see Figure 3).
There was no dependence between items (correlations >0.2) and unidimensionality was confirmed by t-tests between positive and negative loading items with only 2.96% significant below the 5% level. However three items D1, D4 and D7 still had fit residuals outside the ±2.5 threshold.

Significant uniform DIF was also observed for item D7 'carrying a heavy object' by gender and age. Deleting this item improved overall fit (see table 1, analysis 7) with only one item D3 'putting on undershirt of jumper' showing misfit (-3.1). Finally we explored the option of deleting item D3 given that a negative fit residual usually indicates redundancy (34). This 6-item version showed reasonable fit to the Rasch model with item and person fit residuals within acceptable thresholds. However significant uniform DIF is still evident for D1 and D4 by gender and D5 by age. Washing hair and putting on a shirt were slightly more difficult for women than men (see Figure 5a to c) and putting on trousers was more difficult for those aged 60 or over. The mean and standard deviation of the fit residuals however lies outside the threshold for ideal values and there are fewer items available at the extremes (see Figure 4) suggesting a ceiling and floor effect. The 6-item

version retains a high Person-Separation Index (PSI=0.84) meaning it can discriminate between at least 3 groups.

DISCUSSION

This study provides strong support for the structural validity of the pain subscale of the SPADI but only moderate support for the disability subscale with modifications based on the Rasch model. The SPADI in its traditional format using an 11-point rating scale for all five pain and eight disability questions which are summed into a single score does not meet the expectations for interval- level measurement and shows significant misfit with the Rasch model.

Firstly, combining the 13 items into a total score does not meet the assumption of a unidimensional scale. This concurs with findings from several factor analyses that have indicated that the SPADI has at least two dimensions (9, 12). Recently, data on the SPADI and Oxford Shoulder Score collected in a large randomized controlled trial comparing shoulder surgery with rest and exercise were analysed using exploratory and confirmatory factor analysis (8). The fit of both one- and two-factor hypothesised models were assessed and the authors conclude that both a single factor 13-item structure and two-factor pain and disability structure are supported and that the SPADI pain and disability subscales are suitable as primary endpoints. Our findings using Rasch analysis do not support the notion of a single underlying factor which concur with several other studies (5, 6, 33) that identified at least 2 dimensions with most of the pain items loading on the first factor and the majority of the disability items on the second factor. Therefore the SPADI constructs of pain and disability should be treated as two separate subscales.

The pain subscale on its own shows good fit to the Rasch model. Applying the concept of an 11-point numerical rating scale for pain (0 = no pain and 10 =worst pain imaginable) appears to work well as response threshold were ordered throughout and there was a good distribution of responses across the available response categories. However, two of the five items show response bias by age or gender. Since the impact of sex and gender on pain is complex and not fully understood, it is unclear why these differences exist (35, 36). It is important to also consider the clinical significance i.e. the magnitude of differences. Given that the minimal clinically important difference on a 11 point numerical rating scale for pain is at least 2 points (37) then a difference of 0.35 points could be considered negligible. However, uniform DIF by gender on the item 'pain when pushing with involved arm' is just over 1 point. The implication for practice is that this item may need to be scored separately by gender.

The pain subscale demonstrates good reliability with an ability to distinguish statistically between at least three or more groups of severity and the items are well targeted to persons but limited information is provided at the extremes especially those with high pain.

On the other hand, the disability subscale made up of eight items showed significant misfit to the Rasch model. Firstly, patients found it difficult to distinguish between the 11 response categories on at least 3 items. Collapsing the 11-point scale for all eight items into a six point scale resulted in ordered thresholds. We applied the same scoring algorithm to all eight items

(00112233445) as this makes it easier within a busy clinical setting. It is also worth noting that the anchors for the scale range from the descriptors of 'no difficulty' to 'so difficult it requires help' and may be part of the problem in optimizing scaling. It could be argued that requiring help is not as severe as being unable to carry out the activity and may also be dependent on the availability of help. Other disability measures often use 'unable to do' as a final anchor. Whilst an 11-point numerical rating scale appears to work well for pain severity this number of categories may be too many when rating function, although it may be less of a problem if the anchor was changed. It has been suggested that between 5 to 7 Likert-type adjectival responses are less burdensome for patients (38) and may be more appropriate for rating function. On the other hand, the 0-10 scale is familiar to patients, and supports simple scoring.

Even after rescoring three items still showed significant misfit This was most marked for item D7 'difficulty carrying a heavy object' which has a high positive fit residual indicating that it under discriminates. It also shows response bias by gender with a 1 point difference between men and women equating to 17% on the rescored 6-point ordinal scale. Our findings concur with gender differences in strength-based items observed in other upper limb PROMs, for example the Simple Shoulder Test (39) and the Patient-Rated Wrist and Hand Evaluation (40). It has been suggested that results of males and females should be considered in a disaggregated analysis to ensure that results are equally valid for both sexes (41). The findings in this study and others that report gender-bias in musculoskeletal measures provide further support for the importance of sex-disaggregated analyses. Rasch analysis of the disability subscale by Cook et al (21) similarly identified D7 'difficulty carrying a heavy object' as misfitting alongside D4 'putting on shirt that buttons at the front' and D8 'removing something form your back pocket'.

Reasonable overall fit was achieved when deleting two items and could be one solution (25), however this can result in loss of clinically important information and affect targeting. The disability subscale in its current format does not provide true interval level measurement and further studies using Rasch model analysis need to be conducted to explore whether our findings can be replicated before adjustments are made to the SPADI disability subscale.

Our analysis of the structural validity of the SPADI subscales using the Rasch model concerns only one aspect of construct validity. It does not address other psychometric properties such as content validity, known-groups validity, test-retest reliability or responsiveness. The SPADI has been extensively studied using classical test theory approaches and there is strong support with regards to its test-retest reliability (coefficients ranging from 0.66-0.95), its ability to discriminate between known groups and its responsiveness to change (pain subscale effect size=2.1, disability subscale effect size=1.8) (9). However the content validity of the SPADI has not been investigated. The SPADI was developed in the 1990s and items selected by a panel of clinical experts consisting of 3 rheumatologist and one physical therapist. It does not meet current criteria for PROM development (42, 43) which recommend the use of in-depth interviews with patients to elicit key concepts as well as cognitive interviews to assess patient understanding of the new PROM. However, these standards were not in place when many PROMS used today were developed.

The SPADI has been translated for use in several countries including Turkish, Persian, Dutch, Danish, Arabic and German, however none of the studies have examined it for differential item

functioning by comparing responses across countries or used cognitive interviews to assess its content validity.

A number of participants were identified as 'extreme' when examining individual person-by-item fit statistics. They included participants who endorsed the highest score on one or both subscales (score of 10). Further examination of their clinical characteristics did not identify any consistent pattern with regards to gender, age or work status other than they all had low pain self-efficacy scores. It is also possible that patients did not read or fully understand the instructions and as the SPADI questionnaires were self-completed by patients there was no opportunity to clarify patient responses.

Clinical implications:

The SPADI is a widely used outcome measure in clinical practice and research, however the scores derived from it must be interpreted with caution. Firstly it should be treated as two separate subscales – a five-item pain subscale and 8 item disability subscale. The pain subscale is based on an 11-point ordinal scale applied to each of the five questions which can be summed into a total score. However, two items show response bias by age and sex. This means that, for example, a score of 5 for pain when pushing with the affected arm cannot be interpreted in the same way in men and women. It may be more meaningful to look at the change in score before and after an intervention.

The disability subscale showed significant misfit to the Rasch model which makes the interpretation of the summed score more problematic. Firstly its 8 items have to be rescored using a shorter 6-point ordinal scale. Secondly several items show response bias by age and gender. Only when deleting 2 items could overall fit to the Rasch model be achieved. This means that a summed score from the disability subscale in its current format cannot be treated as interval-level measurement. Deletion of some items does resolve this misfit, however this would result in a scale covering fewer items and which gives less clinical information upon which to plan treatment. Similarly, patients may consider the scale less relevant to their problems if it does not contain certain items. It also makes it difficult to compare results reported using the original scale with those from any revised versions. Clinicians should therefore exercise caution when interpreting score changes on the disability subscale and attempt to compare their scores to age and sex stratified data.

Strengths and limitations:

Our analysis is based on a large and relatively homogenous sample size of patients with musculoskeletal shoulder pain prospectively recruited to a multi-centre observational study from across a large geographical region in the UK. The power of analysis of fit was excellent, however large sample sizes (>500) can lead to inflated significance on the total item-trait chi-square statistics (44, 45). RUMM2030 has a function for adjusting sample sizes and when using only half the sample (n=507) the total item Chi-square statistics was not significant (Chi-square= 39.4, df=54, p=0.93) however this does not alter the individual item fit statistics. Although the Rasch model is sample

independent, this only applies if the data fit the Rasch model and Rasch studies by different authors do not always agree. Therefore a single study using Rasch model analysis is not sufficient justification to recommend modifications to a well-established PROM such as the SPADI and further studies are needed including patients with a wider range of shoulder conditions.

CONCLUSIONS

Rasch Model analysis of the SPADI has identified some strengths and limitations not previously observed using CTT methods. The SPADI should be treated as two separate subscales for pain (5 items) and disability (8 items). The pain subscale fits the Rasch model expectations well with the exception of some response bias by age and gender for some items. To accommodate this differential item functioning separate scores have to be calculated by gender and age groups. The disability subscale showed significant misfit to the Rasch model. Whilst some sources of misfit could be addressed by using fewer response categories (rescoring), generating separate item location estimates for men and women and by age to accommodate DIF or even deleting misfitting items, further Rasch analysis using samples with a wide range of shoulder condition is needed before modifications are made to the SPADI disability subscale. At present the disability subscale does not meet the criteria for true interval-level measurement required for use as a primary endpoint in clinical trials.

Table and Figure legends

Table 1: Summary of analysis stages for full SPADI and pain and disability subscales

Table 2: Item fit statistics for Pain subscale (n=1013) (based on analysis stage 3)

Table 3: Item fit statistics for disability subscale (n=1014) (based on analysis stage 8)

Figure 1: person-item threshold distribution of 5-item Pain subscale

Figure 2: Item characteristic curves for P1 and P5 plotted by person factors age and gender

Figure 3: Threshold map in location order for Disability subscale of SPADI (based on rescoring 00112233445)

Figure 4: person-item threshold distribution of reduced 6-item Disability subscale

Figure 5: Item characteristic curves for items D1 'washing hair' and D4 'putting on a shirt' by gender and D5 'putting on pants' by age

References

1. Linsell L, Dawson J, Zondervan K, Rose P, Randall T, Fitzpatrick R, et al. Prevalence and incidence of adults consulting for shoulder conditions in UK primary care; patterns of diagnosis and referral. Rheumatology. 2006;45(2):215-21.
2. Luime JJ, Koes BW, Hendriksen IJM, Burdorf A, Verhagen AP, Miedema HS, et al. Prevalence and incidence of shoulder pain in the general population; a systematic review. Scand J Rheumatol. 2004;33(2):73-81.
3. Page MJ, Huang H, Verhagen AP, Gagnier JJ, Buchbinder R. Outcome reporting in randomized trials for shoulder disorders: Literature review to inform the development of a core outcome set. Arthritis Care Res (Hoboken). 2017.
4. Buchbinder R, Page MJ, Huang H, Verhagen AP, Beaton D, Kopkow C, et al. A Preliminary Core Domain Set for Clinical Trials of Shoulder Disorders: A Report from the OMERACT 2016 Shoulder Core Outcome Set Special Interest Group. The Journal of rheumatology. 2017.
5. Roach KE, Budiman-Mak E, Songsiridej N, Lertratanakul Y. Development of a shoulder pain and disability index. Arthritis care and research : the official journal of the Arthritis Health Professions Association. 1991;4(4):143-9.
6. MacDermid JC, Solomon P, Prkachin K. The Shoulder Pain and Disability Index demonstrates factor, construct and longitudinal validity. BMC musculoskeletal disorders. 2006;7.
7. Angst F, Schwyzer HK, Aeschlimann A, Simmen BR, Goldhahn J. Measures of Adult Shoulder Function Disabilities of the Arm, Shoulder, and Hand Questionnaire (DASH) and Its Short Version (QuickDASH), Shoulder Pain and Disability Index (SPADI), American Shoulder and Elbow Surgeons (ASES) Society Standardized Shoulder Assessment Form, Constant (Murley) Score (CS), Simple Shoulder Test (SST), Oxford Shoulder Score (OSS), Shoulder Disability Questionnaire (SDQ), and Western Ontario Shoulder Instability Index (WOSI). Arthrit Care Res. 2011;63:S174-S88.
8. Dawson J, Harris KK, Doll H, Fitzpatrick R, Carr A. A comparison of the Oxford shoulder score and shoulder pain and disability index: factor structure in the context of a large randomized controlled trial. Patient-Relat Outcom. 2016;7.
9. Roy JS, MacDermid JC, Woodhouse LJ. Measuring Shoulder Function: A Systematic Review of Four Questionnaires. Arthrit Rheum-Arthr. 2009;61(5):623-32.
10. Thoomes-de Graaf M, Scholten-Peeters GGM, Schellingerhout JM, Bourne AM, Buchbinder R, Koehorst M, et al. Evaluation of measurement properties of self-administered PROMs aimed at patients with non-specific shoulder pain and "activity limitations": a systematic review. Qual Life Res. 2016;25(9):2141-60.
11. St-Pierre C, Desmeules F, Dionne CE, Fremont P, MacDermid JC, Roy JS. Psychometric properties of self-reported questionnaires for the evaluation of symptoms and functional limitations in individuals with rotator cuff disorders: a systematic review. Disabil Rehabil. 2016;38(2):103-22.
12. Hill CL, Lester S, Taylor AW, Shanahan ME, Gill TK. Factor structure and validity of the shoulder pain and disability index in a population-based study of people with shoulder symptoms. BMC musculoskeletal disorders. 2011;12:8.
13. Chester R, Jerosch-Herold C, Lewis J, Shepstone L. The SPADI and QuickDASH Are Similarly Responsive in Patients Undergoing Physical Therapy for Shoulder Pain. The Journal of orthopaedic and sports physical therapy. 2017;47(8):538-47.
14. Cano SJ, Hobart JC. The problem with health measurement. Patient preference and adherence. 2011;5:279-90.

15. Hobart JC, Cano SJ, Zajicek JP, Thompson AJ. Rating scales as outcome measures for clinical trials in neurology: problems, solutions, and recommendations. The Lancet Neurology. 2007;6(12):1094-105.
16. Andrich D. Rasch Models for Measurement. London: Sage; 1988. 94 p.
17. Hagquist C, Bruce M, Gustavsson JP. Using the Rasch model in nursing research: an introduction and illustrative example. International journal of nursing studies. 2009;46(3):380-93.
18. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. Journal of clinical epidemiology. 2010;63(7):737-45.
19. Wright BD, Linacre JM. Observations Are Always Ordinal - Measurements, However, Must Be Interval. Archives of physical medicine and rehabilitation. 1989;70(12):857-60.
20. Tennant A, McKenna SP, Hagell P. Application of Rasch analysis in the development and application of quality of life instruments. Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research. 2004;7 Suppl 1:S22-6.
21. Cook KF, Gartsman GM, Roddey TS, Olson SL. The measurement level and trait-specific reliability of 4 scales of shoulder functioning: an empiric investigation. Archives of physical medicine and rehabilitation. 2001;82(11):1558-65.
22. Chester R, Jerosch-Herold C, Lewis J, Shepstone L. Psychological factors are associated with the outcome of physiotherapy for people with shoulder pain: a multicentre longitudinal cohort study. Br J Sports Med. 2016.
23. Chester R, Shepstone L, Lewis JS, Jerosch-Herold C. Predicting response to physiotherapy treatment for musculoskeletal shoulder pain: protocol for a longitudinal cohort study. BMC musculoskeletal disorders. 2013;14:192.
24. Andrich D. Rating Formulation for Ordered Response Categories. Psychometrika. 1978;43(4):561-73.
25. Lundgren Nilsson A, Tennant A. Past and present issues in Rasch analysis: the functional independence measure (FIM) revisited. Journal of rehabilitation medicine. 2011;43(10):884-91.
26. Pallant JF, Tennant A. An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS). The British journal of clinical psychology / the British Psychological Society. 2007;46(Pt 1):1-18.
27. Smith EV, Jr. Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. Journal of applied measurement. 2002;3(2):205-31.
28. Hobart J, Cano S. Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods. Health technology assessment. 2009;13(12):iii, ix-x, 1-177.
29. Andrich D, Humphry SM, Marais I. Quantifying Local, Response Dependence Between Two Polytomous Items Using the Rasch Model. Appl Psych Meas. 2012;36(4):309-24.
30. Andrich D, Hagquist C. Real and Artificial Differential Item Functioning in Polytomous Items. Educ Psychol Meas. 2015;75(2):185-207.
31. Bland JM, Altman DG. Multiple Significance Tests - the Bonferroni Method .10. Brit Med J. 1995;310(6973):170-.
32. Fischer WJ. Reliability Statistics. Rasch Measurement Transactions 1992;6(3):238.
33. Tveita EK, Sandvik L, Ekeberg OM, Juel NG, Bautz-Holter E. Factor structure of the Shoulder Pain and Disability Index in patients with adhesive capsulitis. BMC musculoskeletal disorders. 2008;9.

34. Shea TL, Tennant A, Pallant JF. Rasch model analysis of the Depression, Anxiety and Stress Scales (DASS). BMC psychiatry. 2009;9:21.
35. Racine M, Tousignant-Laflamme Y, Kloda LA, Dion D, Dupuis G, Choiniere M. A systematic literature review of 10 years of research on sex/gender and experimental pain perception - Part 1: Are there really differences between women and men? Pain. 2012;153(3):602-18.
36. Racine M, Tousignant-Laflamme Y, Kloda LA, Dion D, Dupuis G, Choiniere M. A systematic literature review of 10 years of research on sex/gender and pain perception - Part 2: Do biopsychosocial factors alter pain sensitivity differently in women and men? Pain. 2012;153(3):619-35.
37. Michener LA, Snyder AR, Leggin BG. Responsiveness of the numeric pain rating scale in patients with shoulder pain and the effect of surgical status. Journal of sport rehabilitation. 2011;20(1):115-28.
38. Norman D, Streiner G. Health Measurement Scales: a practical guide to their development and use. 3rd edition ed. Oxford: Oxford University Press; 2003.
39. Raman J, MacDermid JC, Walton D, Athwal GS. Rasch analysis indicates that the Simple Shoulder Test is robust, but minor item modifications and attention to gender differences should be considered. Journal of hand therapy : official journal of the American Society of Hand Therapists. 2017.
40. Packham T, MacDermid JC. Measurement properties of the Patient-Rated Wrist and Hand Evaluation: Rasch analysis of responses from a traumatic hand injury population. Journal of Hand Therapy. 2013;26(3):216-24.
41. Johnson JL, Greaves L, Repta R. Better science with sex and gender: Facilitating the use of a sex and gender-based analysis in health research. Int J Equity Health. 2009;8.
42. Patrick DL, Burke LB, Gwaltney CJ, Leidy NK, Martin ML, Molsen E, et al. Content Validity-Establishing and Reporting the Evidence in Newly Developed Patient-Reported Outcomes (PRO) Instruments for Medical Product Evaluation: ISPOR PRO Good Research Practices Task Force Report: Part 1-Eliciting Concepts for a New PRO Instrument. Value in Health. 2011;14(8):967-77.
43. Patrick DL, Burke LB, Gwaltney CJ, Leidy NK, Martin ML, Molsen E, et al. Content Validity-Establishing and Reporting the Evidence in Newly Developed Patient-Reported Outcomes (PRO) Instruments for Medical Product Evaluation: ISPOR PRO Good Research Practices Task Force Report: Part 2-Assessing Respondent Understanding. Value in Health. 2011;14(8):978-88.
44. Hagell P, Westergren A. Sample Size and Statistical Conclusions from Tests of Fit to the Rasch Model According to the Rasch Unidimensional Measurement Model (Rumm) Program in Health Outcome Measurement. Journal of applied measurement. 2016;17(4):416-31.
45. Smith AB, Rush R, Fallowfield LJ, Velikova G, Sharpe M. Rasch fit statistics and sample size considerations for polytomous data. BMC Med Res Methodol. 2008;8.

Figure 1: Item characteristic curves for P1 and P5 plotted by person factors age and gender

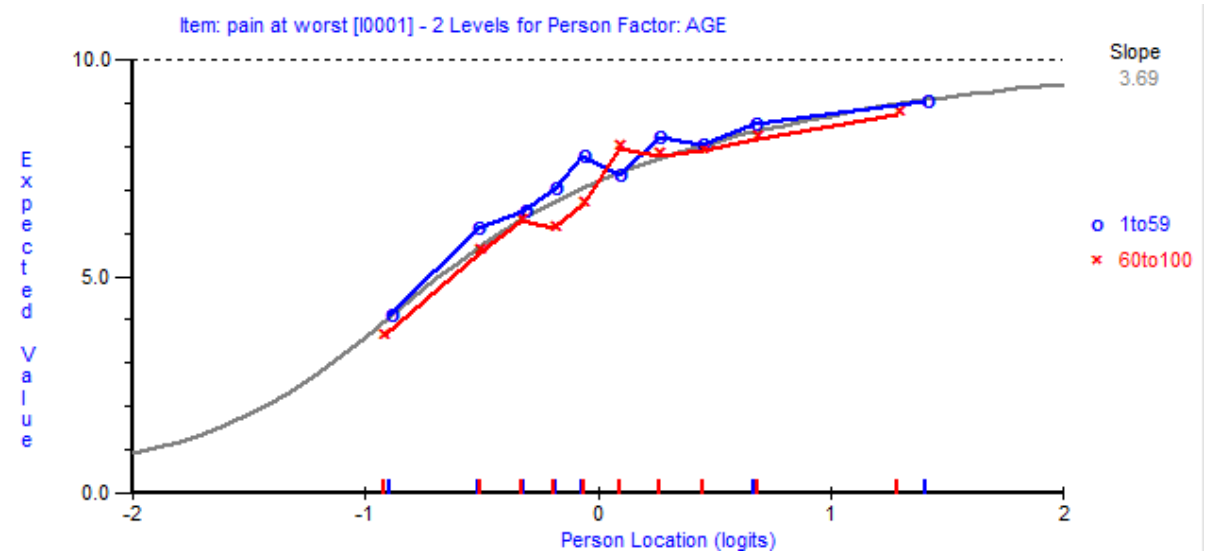
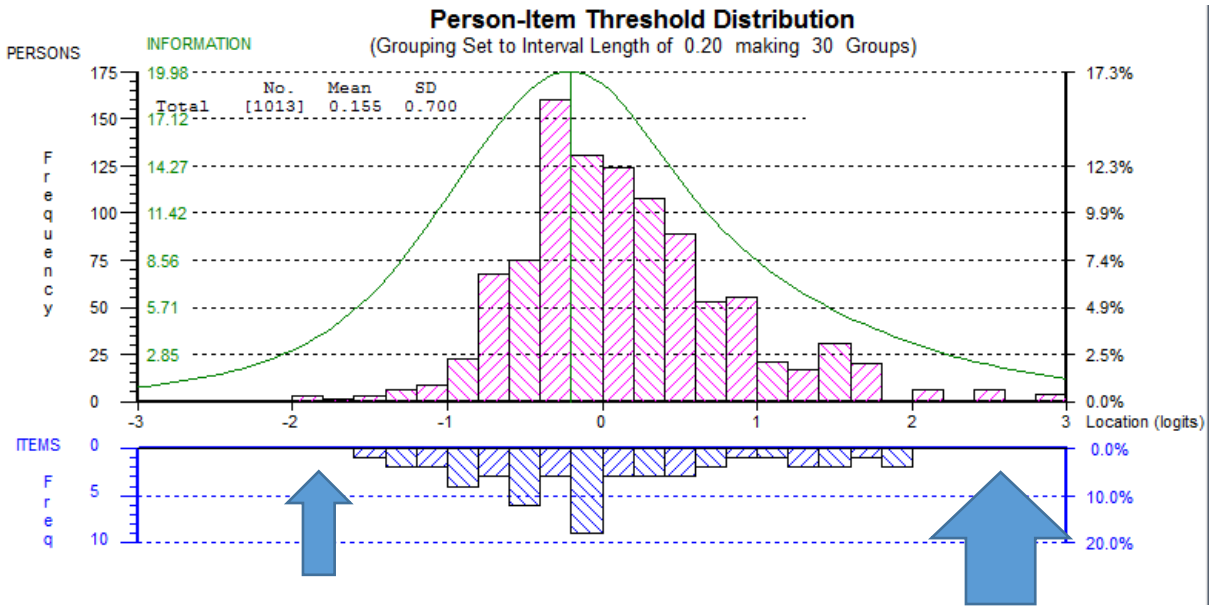


Figure 2: person-item threshold distribution of 5-item Pain subscale



Legend: Blue arrows indicate area where no available items on subscale to fit with persons. The highest point of the information curve (green line) indicates the area where the scale functions at its best

Figure 3: Threshold map in location order for Disability subscale of SPADI (based on rescoring 00112233445)

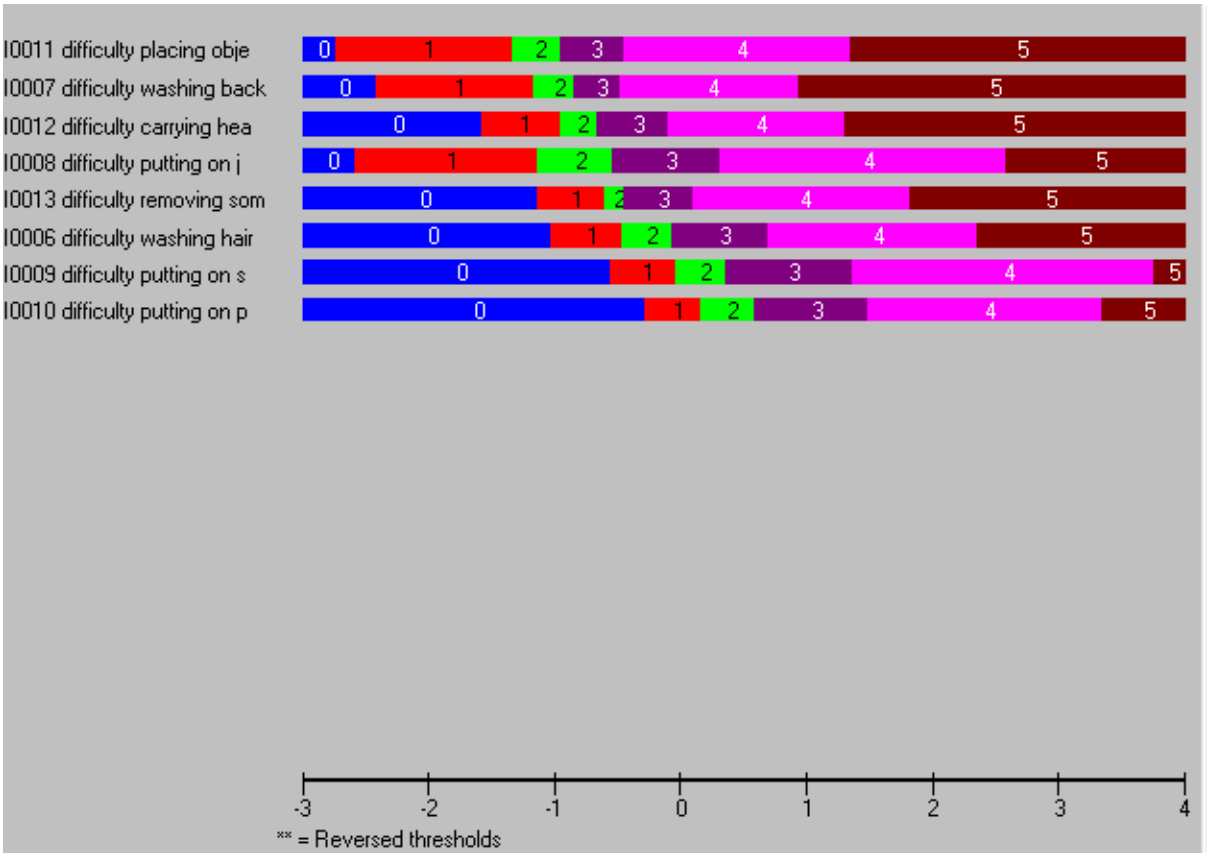
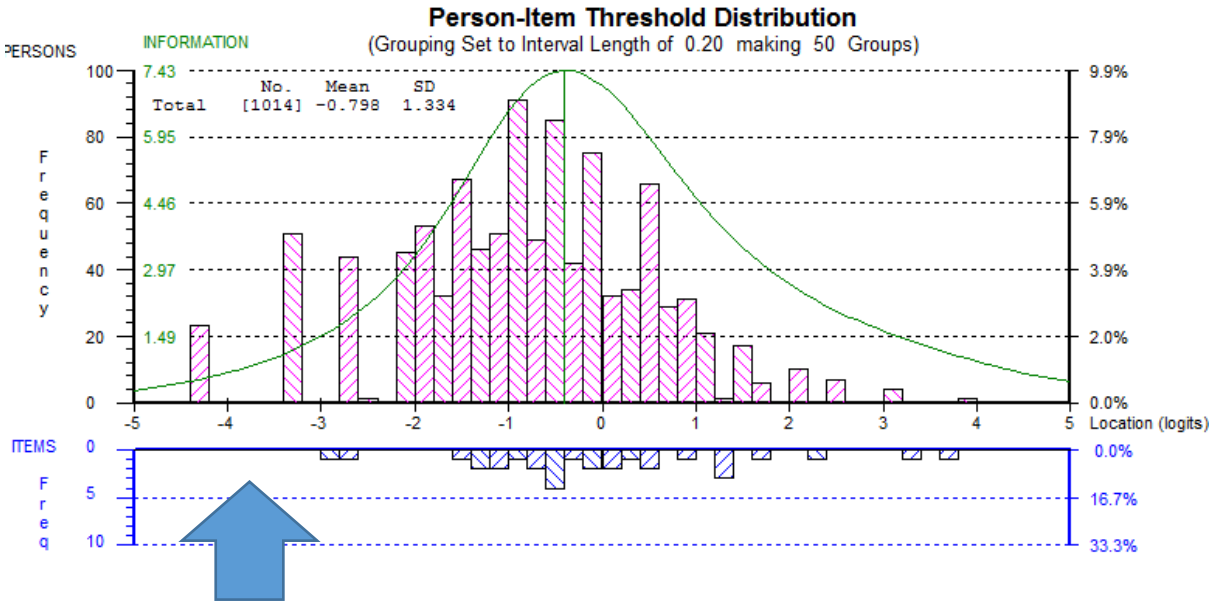
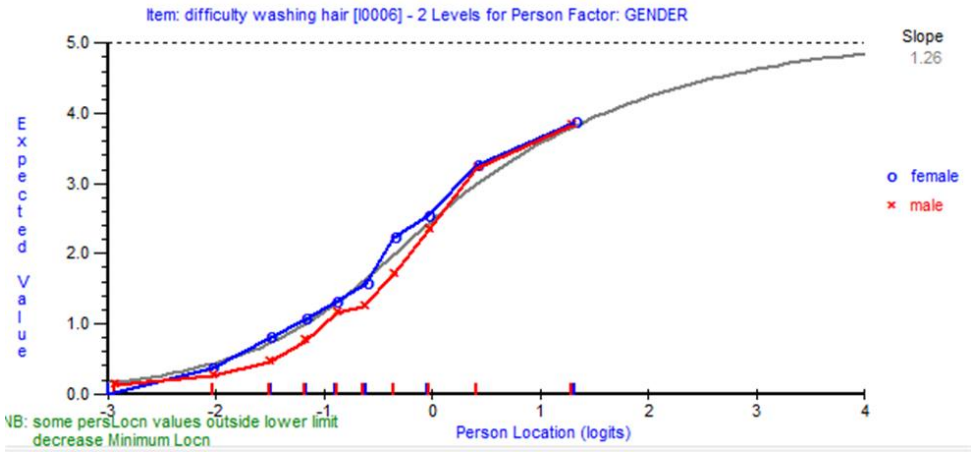


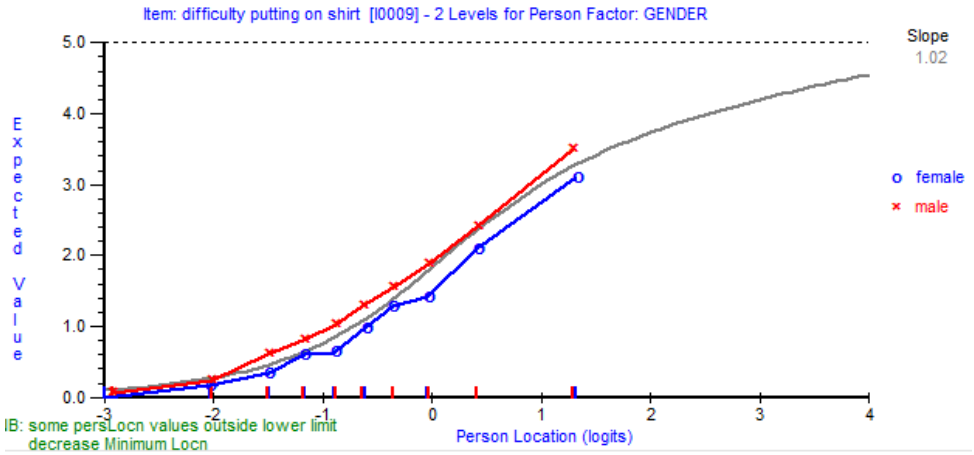
Figure 4: person-item threshold distribution of reduced 6-item Disability subscale



Legend: Blue arrow indicate area where no available items on subscale to fit with persons. The highest point of the information curve (green line) indicates the area where the scale functions at its best

Figure 5: Item characteristic curves for items D1 'washing hair' and D4 'putting on a shirt' by gender and D5 'putting on pants' by age





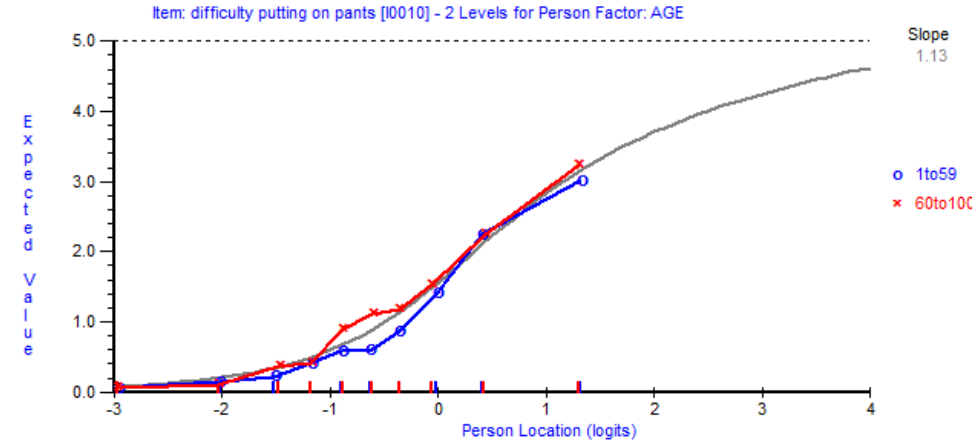


Table 1: Summary of analysis stages for full SPADI and pain and disability subscales

Stages of analysis	n=	Mean item fit residual mean (SD)	Mean person fit residual (SD)	Item-trait total chi-square		PSI	Test of unidimensionality[1] (95%CI)
				χ^2 (df)	*P*		
Ideal values		**mean=0, SD=1**	**mean=0, SD=1**		**>0.05**	**>0.85**	**<5%**
1. Initial analysis of full SPADI (11 items)	1030	0.15 (3.3)	-0.37 (1.4)	301.7 (117)	<0.001	n/a	10.1%
2. Pain subscale only	1030	0.42 (1.9)	-0.45 (1.2)	59.6 (45)	0.07	0.84	5.2%
3. Pain subscale (delete misfitting persons n=17)	1013	0.19 (1.5)	-0.44 (1.1)	57.2 (45)	0.1	0.83	4.55%
4. Disability subscale only	1030	-0.01 (3.3)	-0.37 (1.2)	209.5 (72)	<0.001	0.89	4.47%
5. Disability subscale only (delete misfitting persons n=16)	1014	0.007 (3.3)	-0.37 (1.2)	209.5 (72)	<0.001	0.89	4.47%
6. Disability subscale only rescore all (00112233445)	1014	-0.44 (2.9)	-0.37 (1.1)	164.2 (72)	<0.001	0.88	2.96%
7. Delete D7 (carry heavy object)	1014	-0.52 (1.7)	-0.38 (1.1)	99.2 (63)	0.0025	0.87	2.96%
8. Delete D3 (putting on undershirt or jumper)	1014	-0.54 (1.26)	-0.38 (1.0)	77 (54)	0.022	0.84	2.96%

[1] percentage of equating t-tests which are significant at p<0.05, a percentage below 5% or where the lower bound of the 95% CI straddles 5% indicates unidimensionality

Table 2: Item fit statistics for Pain subscale (n=1013) (based on analysis stage 3)

Item	description	Location	SE	FitResid	ChiSq	Prob*
P1	at its worst	-0.495	0.021	0.63	790.94	0.152
P2	lying on affected side	-0.033	0.018	2.00	800.48	0.935
P3	reaching for object on a high shelf	-0.134	0.018	-2.22	802.86	0.022
P4	touching the back of your neck	0.418	0.017	0.22	804.45	0.506
P5	pushing with involved arm	0.244	0.017	0.34	801.27	0.176

*probability adjusted by number of comparisons $p<0.01$,
SE= standard error, ChiSq= Chi-square statistic

Table 3: Item fit statistics for disability subscale (n=1014) (based on analysis stage 8)

Item description	Location	SE	FitResid	ChiSq	Prob*
D1 washing your hair	0.194	0.033	-2.24	19.09	0.024
D2 washing your back	-0.927	0.032	-1.27	15.37	0.081
D4 putting on a shirt that buttons at front	0.868	0.036	0.90	10.28	0.329
D5 putting on trousers	0.976	0.037	-1.39	12.04	0.211
D6 placing an object on high shelf	-0.955	0.033	0.25	8.78	0.457
D8 removing something from your back pocket	-0.156	0.031	0.51	11.42	0.248

fit residuals > ±2.5 highlighted; probability adjusted by number of comparisons p<0.005,
SE= standard error, ChiSq= Chi-square statistic