Rasch model analysis gives new insights into the structural validity of the Quick-DASH in patients with musculoskeletal shoulder pain

Christina Jerosch-Herold (DipCOT, PhD)[1] (corresponding author)

Rachel Chester (GradDipPhys, PhD)[1,2]

Lee Shepstone (PhD)[3]

[1]School of Health Sciences, Faculty of Medicine and Health Sciences, University of East Anglia, Norwich, Norfolk, United Kingdom.

[2]Physiotherapy Department, Norfolk and Norwich University Hospital, Norwich, Norfolk, United Kingdom.

[3]Norwich Medical School, Faculty of Medicine and Health Sciences, University of East Anglia, Norwich, Norfolk, United kingdom.

Corresponding author:

Dr Christina Jerosch-Herold, School of Health Sciences, Faculty of Medicine and Health Sciences, University of East Anglia, Norwich, Norfolk, NR4 7TJ United Kingdom. Email: c.jerosch-herold@uea.ac.uk

Word count: 3200 (excluding title, abstract, tables and figure legends)

Financial disclosure and conflict of Interest

We affirm that we have no financial affiliation (including research funding) or involvement with any commercial organization that has a direct financial interest in any matter included in this manuscript, except as disclosed in an attachment and cited in the manuscript. Any other conflict of interest (i.e., personal associations or involvement as a director, officer, or expert witness) is also disclosed in an attachment.

Grant support:

The lead (CJH) and second (RC) authors 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.

Institutional Review Board: National Research Ethics Service, East of England - Norfolk, UK, July 2011 (reference 11/EE/0212).

Acknowledgements: The authors would like to thank the Principal Investigators and physiotherapists who contributed and collected data for the study and to the patients who generously gave their time and participated in the study. We also thank the anonymous reviewers who offered very constructive and helpful feedback on an earlier version of this article.

Abstract

Study Design: Cross-sectional secondary analysis of a prospective cohort study

Background: The Quick-DASH is a widely used outcome measure which has been extensively evaluated using classical test theory (CTT). Rasch model analysis can identify strengths and weaknesses of rating scales which goes beyond CTT approaches. It uses a mathematical model to test the fit between the observed data and expected responses and converts ordinal-level scores into interval-level measurement.

Objective: To test the structural validity of the Quick-DASH using Rasch analysis

Methods: A prospective cohort study of 1030 patients with shoulder pain provided baseline data. Rasch analysis was conducted to i) assess how the Quick-DASH fits the Rasch model, ii) identify sources of misfit and iii) explore potential solutions to these.

Results: There was evidence of multidimensionality and significant misfit to the Rasch model (χ^2= 331.04, $p<0.001$). Two items had disordered threshold responses with strong flooring effects. Response bias was detected in most items for age and gender. Rescoring resulted in ordered thresholds, however the 11-item scale still did not meet the expectations of the Rasch model.

Conclusion: Rasch model analysis on the Quick-DASH has identified a number of problems which cannot be easily detected using traditional analyses. Whilst revisions to the Quick-DASH resulted in better fit, a ‘shoulder-specific’ version is not advocated at present. Caution needs to be exercised when interpreting results of the Quick-DASH outcome measure as it does not meet the criteria for interval level measurement and shows significant response bias by age and gender.

Key words: outcome measures, item-response theory, construct validity, Rasch Model

Intro/Background:

The Disabilities of the Arm, Shoulder and Hand Questionnaire (DASH) was developed as a 30-item patient-rated outcome measure (PROM) of symptoms and disability for upper extremity musculoskeletal disorders[20]. A modified, shorter 11-item version called the Quick-DASH was later generated to address possible item redundancy and improve speed and ease of administration[4].

The test-retest reliability, construct validity, internal reliability and responsiveness of both the full and shortened DASH have been extensively evaluated[2,3,4,10,21,29,30] with most studies applying classic test theory (CTT) methods. However CTT methods have their limitations[5,6,18] ,in particular, they cannot tell us whether ordinal scales like the Quick-DASH fulfil the linearity assumption met by continuous (interval level) measurement[31]. An alternative approach to CTT is to use Rasch model analysis which is recommended to test the structural validity of both existing and new PROMs especially where these are used in clinical trials of treatment effectiveness[5,19,32]. A full explanation of the Rasch model is beyond the scope of this paper and its application to patient-reported outcome measures has been described in more detail elsewhere[27]. Rasch model analysis uses item response theory (IRT) to quantify the interaction between a person's ability and a scale's individual item level of difficulty[31]. It examines the extent to which observed scores fit with the expected scores under the Rasch model, a fundamental assumption of which is that items follow an ordered hierarchy on a unidimensional scale. Furthermore it tests the assumption that ordinal-level scores approximate interval-level measurements by converting the raw (ordinal) scores into equal units on a 'logit' (log odds) scale on a linear scale[31].

Rasch analysis has been applied to the full 30-item DASH in several patient populations, including multiple sclerosis[5], Dupuytren' s disease[13] and mixed upper extremity musculoskeletal disorders [15,22]. All four studies identified some misfit of full DASH data to the Rasch model. Only one study has applied Rasch methods to the shorter Quick-DASH[14]. They used both CTT (exploratory factor analysis) as well as Rasch analysis to examine the structural validity of the Italian Quick-DASH in a sample of 283 patients with upper limb dysfunction affecting hand, elbow or shoulder. Of these 173 (61%) had disorders of the shoulder complex including surgically treated patients. They concluded that the Quick-DASH is not unidimensional and proposed a 10-item revised scale. However, recommending changes to the well-established and already shortened version of the Quick-DASH has considerable practical implications and should not be undertaken lightly. Therefore further studies using Rasch model analysis with larger samples and in which a range of fit solutions can be explored are warranted.

The objective of this study was to evaluate the structural validity of the Quick-DASH in a large cohort of patients treated conservatively for shoulder pain using Rasch analysis. Specific aims were to conduct an analysis to examine i) unidimensionality (a criterion for summing individual item responses into a single score), ii) targeting of items on the Quick-DASH with patient ability, iii) response thresholds including floor and ceiling effects; iii) independence between items and iv) response bias (do responses differ between persons based on other characteristics such as age or gender). Iterative analyses were used to test if modifications to the scale improve overall fit to the model and make recommendations for a revised scale to be used in future clinical practice and research.

Methods

Study population

Data were drawn from a large prospective cohort study designed to identify which factors assessed at baseline were associated with outcome following physiotherapy treatment for shoulder pain. Patients were included if they had shoulder or arm pain aggravated by shoulder movement. Fractures, traumatic dislocations, surgical treatment to the shoulder in the previous five years, and radiculopathy were excluded. The study was approved by the National Research Ethics Service, East of England, UK and all patients gave fully informed written consent. The Quick-DASH was one of the outcome measures collected at baseline and follow-up. The full study protocol and results have been reported elsewhere[7,8]. For this secondary analysis of the Quick-DASH we used the initial baseline scores obtained before patients underwent physiotherapy treatment as this is the point at which symptoms and disability are most likely to be present.

Data analyses:

The analysis approach followed recommendations from Lundgren, Nilsson and Tennant[23] for Rasch analysis of polytymous scales. Table 1 summarises each analysis stage. All analyses were performed using RUMM2030 for Windows (Andrich D, Lyne A, Sheridan B, Luo G., RUMM Laboratory, Perth 2003) software.

Firstly a likelihood ratio test (LRT) was performed to assess whether it was appropriate to use the partial credit model or not. Then an initial analysis was used to assess overall fit to the Rasch model by examining the fit between observed scores and expected scores using the total item-trait chi-square statistic (χ^2). A significant p-value, after Bonferroni corrections were applied, indicates misfit between observed and expected scores in the Rasch model.

Next, individual item fit was assessed for all 11 items of the Quick-DASH by examining the fit residuals. Values outside the range of ±2.5 and a statistically significant χ^2 indicate that an item does not fit the Rasch model. Sources of misfit for each item were then explored by examining i) response thresholds and ii) residual correlations to identify local dependence between items. Local independence assumes that responses to an item are independent of the responses to other items in the scale after controlling for the underlying trait. Inter-item residual correlations were examined to identify any correlations greater than the average ±0.2. High correlations can also be indicative of multidimensionality[17].

Unidimensionality was assessed using a two-stage process: firstly, a principal components analysis (PCA) of the residuals is used to identify clusters of items with positive and negative loadings >0.3 on the first component. Secondly, an independent t-test is applied on the two subsets. If the proportion of statistically significant tests does not exceed 5% with the lower binomial confidence interval overlapping 5% this is indicative that the items form a unidimensional scale[28].

Finally, response bias by age and gender for any items was assessed by identifying statistically significant differential item functioning.

Targeting was visually inspected using the person-item threshold map. A well targeted scale is one which covers a range of abilities and where the distribution of items mirrors the distribution of persons.

Reliability was examined using the person-separation index (PSI). A PSI of 0.7 or above is deemed acceptable and indicates that the measure can discriminate between at least 2 separate groups, whilst a PSI of 0.8 indicates discrimination between three or more groups[12].

As the available sample size was large, individual person fit residuals greater than ±2.5 were used to identify extreme persons and these removed from the analysis. Class intervals were inspected in each iteration to ensure that cases were equally distributed across intervals. A test of unidimensionality was applied at each iteration to check this was within the 5% level.

Results

Data were available on 1030 patients who completed the Quick-DASH at baseline. The mean age was 57 years (SD=15) and 44% were male. Average duration of shoulder symptoms was 14 months (SD=28).

The likelihood ratio test (LRT) was highly significant (χ^2=641.76, DF=29, p<0.001) indicating that the partial credit model is appropriate. This is a less restrictive model which provides greater specificity and is appropriate to use for polytomous rating scales (for more than two response options)[25].

The initial analysis (analysis 1) showed a highly significant item-trait total χ^2 statistic (Table 1). Thirty-three persons with residuals greater than ±2.5 were identified and removed from the analysis (analysis 2) leaving a sample of 997 available for analysis. Upon removal of the extreme persons several items still showed significant misfit (analysis 3). Sources of misfit were explored sequentially, starting with response categories. Figure 1 shows disordered thresholds on two items: '*using a knife to cut food'* (Q5) and '*tingling'* (Q10). Both these items also had a strong ceiling effect with over 60% of patients endorsing 'no difficulty' or 'none', respectively (see Figure 2). Cutting food with a knife requires primarily hand dexterity with comparatively little shoulder movement and strength than other items, which may also explain why this task posed no difficulty for 70% of respondents. Of the 1030 participants, data from the physiotherapy assessment indicated that only 110 reported paraesthesia in the affected upper quadrant and those with shoulder pain secondary to radiculopathy were excluded in this study. It is likely that the item '*tingling*' was therefore not relevant to most patients. Rescoring of these two items to 00112 resulted in ordered thresholds (see Figure 3). To assess local dependence mean inter-item residual correlations were explored. Dependence occurs when either items duplicate each other or they both share the same underlying trait. Using a parsimonious threshold[24] of any correlations greater than the average of all correlations +0.2, we identified dependence between items '*open a tight jar*' '*heavy household chores*' and *'carrying shopping or a briefcase*'. Dependence was also found between '*interference with social activities*' and '*limiting work or usual activities*'. Therefore, questions 1, 2 and 3 were combined into a '*household testlet*' and questions 7 and 8 were grouped into a '*participation testlet'*. No further response dependence was found.

As local dependence can also contribute to multidimensionality this was tested using principal component analysis (PCA) of the residuals (see table 2). For the first component three items had high positive loadings and four had high negative loadings (>0.3) indicating that there is not a single underlying construct and that the scale is not unidimensional. This was further confirmed by an equating t-test between two sets of positively and negatively loaded items, which was significant for 8.93% exceeding the 5% recommended threshold. Following the creation of testlets (analysis 4) to deal with local dependence and re-applying the equating t-test procedure the lower bound of the confidence interval fell to the 5% level (analysis 4). However the total item-trait Chi-Square statistic remained highly significant. Individual item fit statistics revealed Questions 5 and 11 as misfitting with residuals greater than ±2.5 and a significant chi-square test in Q5 and Q9 (see table 3).

Response bias was explored by examining the item characteristic curves and probability of significant differential item functioning (DIF) by two person factors: gender (male or female) and age (two groups: up to 59 years or 60 and above). DIF occurs when the responses are affected by a factor other than the underlying trait. For example pain may be perceived and therefore rated differently by men and women. We assessed differential item functioning both before and after creating subtests (based on analysis stages 3 and 4, respectively). Item split was done sequentially selecting the item with highest F-statistic on DIF analysis first in order to distinguish real from artificial DIF[1]. In the initial analysis for DIF (analysis 3), questions 1, 2 and 7 showed significant uniform DIF by age and gender, questions 4, 6 and 10 by age only and questions 3, 8 and 11 by gender only (Table 4). Whilst a sequential process of splitting items by gender and age was explored in several iterations this did not resolve local dependence or cleared any remaining DIF. After creating 2 testlets (analysis 4) DIF remained in both subtests by age and gender. Significant uniform DIF by age was also present for questions 4, 6 and 10. Splitting both testlets by gender cleared any remaining DIF by gender. However significant uniform DIF by age remained for the 'household' testlet in men and women, the 'participation' testlet in women, Q4, Q6 and Q10. Given the significant misfit of Q5 and 11 a third option in which these two questions were deleted was also explored (see analysis 5b). Significant uniform DIF by gender was observed in testlets 1 and 2 and these items split (analysis 6). Individual item fit was good with residuals for all items within ±2.5 but the overall item-trait chi-square statistic remained significant (p=0.007).

Targeting was visually inspected using the person-item threshold map (Figure 4). Person and item thresholds are skewed to left which indicates higher ability and easier items, respectively. Overall the targeting between items and persons is good and with a wide spread across ± 4 logits, however there are some gaps on the Quick-DASH at the higher ability/easier item end.

Reliability remained high with a PSI > 0.8 throughout each iteration indicating that it can discriminate between at least three subgroups.

Discussion

Applying Rasch model analysis has provided novel insights into the structural validity of the Quick-DASH in a large cohort of patients referred to physiotherapy with musculoskeletal shoulder pain. The 11-item Quick-DASH shows significant misfit to the Rasch model. In particular, the assumptions of local independence and unidimensionality were not met. Two items show strong flooring effect and patients with shoulder pain do not distinguish between available response categories correctly. Finally it also appears that there is significant response bias by age and gender.

Our findings concur with Franchignoni et al's study which used both CTT and Rasch methods[14] . Although a more heterogeneous and smaller sample than our study they also identified the Quick-DASH as multidimensional. In their sample of 283 patients with a range of upper limb conditions the item 'tingling' did not fit and was subsequently removed to generate a 10-item scale. They also found that the item '*using a knife to cut food*' showed disordered thresholds which they resolved by collapsing adjacent response categories and rescoring. All other items had ordered thresholds and the 5-point ordinal scale was working well across the other items.

Our finding that the scale is not unidimensional also concurs with other studies using conventional PCA analysis[11,16] . Fayad et al[11] used exploratory factor analysis of the French Quick-DASH in 153 patients with shoulder disorders including humeral fractures. Two factors explained 59% of variance, however the authors do not comment further on the implications of this or whether summing all items into a single score is appropriate or not. Gabel et al[16] studied patients with a wide range of upper extremity disorders. They suggested removing the items '*tingling*' and '*pain affecting sleep*' and proposed a modified 9-item scale. Subsequent PCA of the 9-item scale showed unidimensionality.

Both Franchignoni et al[14] and Gabel et al[16] proposed removing items from the Quick-DASH. Removal of Q5 and Q11 was explored in our analyses and did improve the overall fit to the Rasch model, however, we would argue that for patients with shoulder pain, the sleep item is particularly relevant as sleep disturbance is associated with many shoulder problems with a reported prevalance of over 80% [9] . Retaining the content validity of an already existing and shortened version is an important consideration. In our study over 60% of respondents affirmed the responses of 'no difficulty' or 'none' for the items on '*using a knife to cut food*' and '*tingling*', however discarding items can reduce the coverage of a construct for the intended population[32]. For the Quick-DASH this is a wide range of MSK disorders of the upper limb[20] including carpal or cubital tunnel syndrome where tingling may be a clinically important symptom. Proposing revised versions also has considerable implications for the use of the scale in future practice and research and should be based on repeated studies in several populations before vaible alternative forms are recommended[26].

Retaining all 11 items but re-ordering thresholds by collapsing adjacent responses for two items also has practical consequences. At present clinicians administering the Quick-DASH can use a simple algorithm to convert the total raw score (ranging from 11 to 55 points) into a percentage from 0-100%. Applying a different scoring method to two items would take longer and limit comparability of results against existing published data. Furthermore whilst rescoring did result in ordered thresholds several further challenges to achieving fit to the Rasch model remained.

Multidimensionality can also be a manifestation of local dependence[17] which conventional PCA cannot detect. Using testlets for '*household activities*' (questions 1, 2 and 3) and '*participation*' (questions 7 and 8) removed the effect of local dependence and a subsequent t-test for unidimensionality brought this within an acceptable threshold. This means that all 11 items can be summed into a single score. However there was still evidence of individual items misfitting as well as significant response bias by age or gender for 9 items. For example for the '*household testlet'* parameter estimates for women were located at the more able end (-0.452) whereas men were located closer to zero (+0.094) which suggests that women find these items easier. However contrary to other studies we did not see response bias by age for the items on pain (questions 9). There is evidence that pain thresholds differ by age group[33] and significant DIF by age has been observed previously in pain related items found in other upper extremity PROMs such as the Patient-Rated Elbow Evaluation Questionnaire[34].

Targeting of item difficulty to person ability was also good, although skewed towards the higher ability and with gaps in the Quick-DASH at the more able spectrum.

The person-separation index (PSI) ranged from 0.867 to 0.814 at different analyses and indicates that the reliability of the Quick-DASH of discriminating statistically between at least three subgroups remained high.

Strengths and limitations:

Our study was based on a large sample of patients with MSK shoulder pain which excluded surgical cases and those with radiculopathy. A systematic approach was taken to the analysis and exploration of sources of misfit and potential solutions. However, there are also some limitations: only response bias by age and gender was examined and other factors such as response bias by countries for cross-cultural comparison of translated versions of the Quick-DASH needs to be explored in future. Our analysis was cross-sectional and did not include longitudinal analysis to examine responsiveness. Finally we recognise that the application of Rasch analysis to make 'improvements' to the psychometric properties of an existing PROM and ensure interval-level measurement can result in multiple versions and compromises comparability across trial results. Moreover whilst Rasch parameter estimates should be sample independent we cannot rule out that the extent of misfit may be magnified by this large and homogenous sample of patients with shoulder pain only.

Conclusions

In patients with MSK shoulder pain the original 11-item Quick-DASH showed significant misfit with the Rasch model. Further studies using Rasch analysis and CTT methods are needed to assess the consistency of our findings in patients with shoulder pain before the ordinal scores on the Quick-DASH can be considered as linear interval-level measures required for mathematical calculations such as effect sizes.

Key points:

Findings: In patients with musculoskeletal shoulder pain the original Quick-DASH does not fit with the Rasch Model. It is not a unidimensional scale, shows response bias by age and gender and for two items shows strong flooring effects and disordered response thresholds. The Quick-DASH can discriminate between three or more groups of patients with shoulder pain and shows good targeting to person ability.

Implications: Clinicians and researchers need to be aware, when using the 11-item Quick-DASH as an outcome measure for patients with musculoskeletal shoulder pain, that the original 11-item Quick-DASH does not meet interval-level measurement criteria.

Cautions: Further studies on a wider range of MSK upper extremity disorders using Rasch model analysis are needed to assess the consistency of findings before any modifications can be recommended.

References

1. Andrich D, Hagquist C. Real and Artificial Differential Item Functioning in Polytomous Items. *Educ Psychol Meas* 2015;75(2):185-207.

2. Beaton D, Davis A, Hudak P, McConnell S. The DASH (Disabilities of the Arm, Shoulder and Hand) outcome measure: what do we know about it now? *Br J Hand Ther* 2001;6(4):109-118.

3. Beaton D, Katz J, Fossel A, Wright J, Tarasuk V, Bombardier C. Measuring the whole or parts? validity, reliability, and responsiveness of the DASH outcome measure in different regions of the upper extremity. *J of Hand Ther* 2001; 14:128-146.

4. Beaton D, Wright J, Katz JN, Upper Extremity Collaborative. Development of the QuickDASH: comparison of three item-reduction approaches. *J Bone Joint Surg Am* 2005; 87(5):1038-1046.

5. Cano SJ, Barrett LE, Zajicek JP, Hobart JC. Beyond the reach of traditional analyses: using Rasch to evaluate the DASH in people with multiple sclerosis. *Mult Scler* 2011; 17(2): 214-222.

6. Cano SJ, Hobart JC . The problem with health measurement. *Patient Prefer Adherence* 2011; 5: 279-290.

7. 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* Published Online First [21 July 2016] doi:10.1136/bjsports-2016-096084.

8. Chester R, Shepstone L, Lewis J, Jerosch-Herold C. Predicting response to physiotherapy treatment for musculoskeletal shoulder pain: protocol for a longitudinal cohort study. *BMC Musculoskelet Disord* 2013; 14: 192.

9. Cho CH, Jung SW, Park JY, Song KS , Yu KI. Is shoulder pain for three months or longer correlated with depression, anxiety, and sleep disturbance? *J Shoulder Elbow Surg* 2013; 22(2): 222-228.

10. Dowrick AS, Gabbe BJ, WilliamsonOD, Cameron PA. Outcome instruments for the assessment of the upper extremity following trauma: a review. *Injury* 2005;36(4): 468-476.

11. Fayad F, Lefevre-Colau MM, Gautheron V, et al. Reliability, validity and responsiveness of the French version of the questionnaire Quick Disability of the Arm, Shoulder and Hand in shoulder disorders. *Man Ther* 2009;14(2): 206-212.

12. Fischer WJ. Reliability Statistics. *Rasch Measurement Transactions* 1992; 6(3): 238.

13. Forget NJ, Jerosch-Herold C, Shepstone L, Higgins J. Psychometric evaluation of the Disabilities of the Arm, Shoulder and Hand (DASH) with Dupuytren's contracture: validity evidence using Rasch modeling. *BMC Musculoskelet Disord* 2014;15: 361.

14. Franchignoni F, Ferriero G, Giordano A, Sartorio F, Vercelli S, Brigatti E. Psychometric properties of QuickDASH - a classical test theory and Rasch analysis study. *Man Ther* 2011;16(2): 177-182.

15. Franchignoni F, Giordano A, Sartorio F, Vercelli S, Pascariello B, Ferriero G. Suggestions for refinement of the Disabilities of the Arm, Shoulder and Hand Outcome Measure (DASH): a factor analysis and Rasch validation study. *Arch Phys Med Rehabil* 2010;91(9): 1370-1377.

16. Gabel CP, Yelland M, Melloh M, Burkett B . A modified QuickDASH-9 provides a valid outcome instrument for upper limb function. *BMC Musculoskelet Disord* 2009; 10:161.

17. Hagquist C, Bruce M, Gustavsson JP. Using the Rasch model in nursing research: an introduction and illustrative example. *Int J Nurs Stud* 2009; 46(3):380-393.

18. Hobart J, Cano S. Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods. *Health Technol Assess* 13(12): iii, ix-x, 1-177.

19. Hobart JC, Cano SJ, Zajicek JP, Thompson AJ. Rating scales as outcome measures for clinical trials in neurology: problems, solutions, and recommendations. *Lancet Neurol* 2007;6(12): 1094-1105.

20. Hudak P, Amadio P, Bombardier C, T. U. E. C. G. (UECG). Development of an upper extremity outcome measure: The DASH (Disabilities of the Arm, Shoulder, and Hand). *Am J Ind Med* 1996; 29: 602-608.

21. Kennedy CA, Beaton DE, Smith P et al. Measurement properties of the QuickDASH (Disabilities of the Arm, Shoulder and Hand) outcome measure and cross-cultural adaptations of the QuickDASH: a systematic review. *Qual Life Res* 2013;22(9): 2509-2547.

22. Lehman L A, Woodbury M, Velozo CA. Examination of the factor structure of the Disabilities of the Arm, Shoulder, and Hand questionnaire. *Am J Occup Ther* 2011;65(2): 169-178.

23. Lundgren Nilsson A., Tennant A. Past and present issues in Rasch analysis: the functional independence measure (FIM) revisited. *J Rehabil Med* 2011;43(10): 884-891.

24. Marais I, Andrich D. Effects of varying magnitude and patterns of local dependence in the unidimensional Rasch model. *J Appl Meas* 2008;9(2): 1-20.

25. Masters G. A Rasch model for partial credit scoring. *Psychometrika* 1982;47: 149-174.

26. Packham T, MacDermid JC. Measurement properties of the Patient-Rated Wrist and Hand Evaluation: Rasch analysis of responses from a traumatic hand injury population. *J Hand Ther* 2013;26(3): 216-224.

27. Pallant JF, Tennant A. An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS). *Br J Clin Psychol* 2007;46(1): 1-18.

28. Smith EV. Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *J Appl Meas* 2002;3(2): 205-231.

29. Solway S, Beaton DE, McConnell S, Bombardier C. *The DASH and QuickDASH Outcome Measure User's Manual*. Toronto, Ontario, Institute of Work & Health, 2002

30. SooHoo NF, McDonald AP, Seiler JG, McGillivary GR. Evaluation of the construct validity of the DASH questionnaire by correlation to the SF-36. *J Hand Surg Am* 2002;27(3): 537-41.

31. Tennant A, Conaghan PG. The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis and Rheum* 2007; 57(8): 1358-62.

32. Tennant A, McKenna SP, Hagell P. Application of Rasch analysis in the development and application of quality of life instruments. *Value Health* 2004; 7:S22-6.

33. Vincent JI, MacDermid JC, King GJ, Grewal R. Validity and sensitivity to change of patient-reported pain and disability measures for elbow pathologies. *J Orthop Sports Phys Ther* 2013; 43(4): 263-74.

34. Vincent JI, MacDermid JC, King GJ, Grewal R. Rasch analysis of the Patient Rated Elbow Evaluation questionnaire. *Health Qual Life Outcomes* 2015; 13:84.

Tables

Table 1: Summary of Rasch analyses of the Quick-DASH

Stages of analysis		**Item fit residual**	**Person fit residual**	**Item-trait total chi-square**		**PSI**	**Test of unidimensionality[1] (95%CI)**
	n=	**mean ±SD**	**mean ±SD**	**χ^2 (df)**	***P***		
1. Initial analysis	1030	0.081 ±3.03	-0.266 ±1.12	331.0 (99)	<0.001	0.859	
2. delete extreme persons	997	0.036 ±3.071	-0.223 ±1.02	328.3 (99)	<0.001	0.861	
3. Rescore Q5 and Q10 to 00112	997	0.032 ±2.797	-0.26 ±0.982	247.55 (99)	<0.001	0.867	8.93% (7.6 to 10.3)
4. create 2 testlets (subtest 1 and 2)	997	0.125 ±1.914	-0.276 ± 0.912	201.48 (72)	<0.001	0.841	6.31% (5.0 to 7.7)
5a. split subtests 1 & 2 by gender	997	-0.062 ±1.86	-0.285 ±0.910	200 (90)	<0.001	0.845	
5b. analysis 4+ delete Q5 & 11	997	0.199 ± 1.49	-0.296 ± 0.953	129.99 (54)	<0.001	0.814	4.79%
6. analysis 5b+split subtests 1 & 2 by gender	997	-0.087 ± 1.49	-0.31 ±0.952	105.2 (72)	0.007	0.820	
Ideal values		**mean=0, SD<1.4**	**mean=0, SD<1.4**		**>0.05**	**>0.85**	**<5%**

[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: Principal Component Analysis of residuals of 11-item scale (loadings >0.3 highlighted in bold) showing loadings for principal components (PC) with Eigenvalues >1 (analysis 3)

Item number and descriptor	PC1	PC2	PC3	PC4
Eigenvalue (cumulative percent)	2.1 (19.1%)	1.34 (31.3%)	1.28 (43%)	1.15 (53.5%)
1: open a jar	**0.590**	0.039	-0.194	-0.108
2: do heavy household chores	**0.629**	-0.105	-0.076	-0.083
3: carry shopping or a briefcase	**0.516**	-0.257	**-0.365**	-0.005
4: wash your back	0.105	**0.652**	**0.495**	**-0.306**
5: use knife to cut food	0.245	-0.274	0.068	**-0.353**
6: recreational activity (golf, hammering, tennis)	0.017	-0.262	**0.608**	**0.688**
7: interference with social activities	**-0.606**	**-0.398**	0.025	-0.102
8: limited work or usual activities	**-0.501**	**-0.471**	0.053	**-0.363**
9: arm, shoulder or hand pain	**-0.461**	0.197	-0.171	**-0.314**
10. tingling in arm, shoulder or hand	-0.168	-0.011	**-0.588**	**0.371**
11. difficulty sleeping because of arm pain	**-0.401**	**0.523**	**-0.339**	0.252

Table 3: Item fit statistics in location order for 11-item Quick-DASH after creating two testlets (analysis 4)

item	location	SE	Fit residual	Chi-sq	Prob	F-stat	Prob
subtest 1 'household'	-0.29	0.019	-0.727	5.131	0.8227	0.616	0.7843
subtest 2 'participation'	-0.172	0.025	0.501	17.521	0.0412	2.138	0.0242
Q4	-0.952	0.033	0.84	13.18	0.1547	1.625	0.1035
Q5	2.213	0.092	**-2.863**	49.989	**<0.0001**	9.468	**<0.0001**
Q6	-1.141	0.033	1.758	20.757	0.0138	2.357	0.0124
Q9	-1.131	0.047	-1.902	54.635	**<0.0001**	8.153	**<0.0001**
Q10	1.74	0.078	0.348	25.79	0.0022	3.274	**0.0006**
Q11	-0.266	0.037	**3.043**	14.476	0.1064	1.557	0.1235
Ideal values			**< ± 2.5**		**>0.05***		**>0.05***

Significant p-value (adjusted for 8 items to p<0.00125) and fit residuals ±2.5 in bold

Table 4: Differential Item functioning for 11-item Quick-DASH based on initial analysis excluding extremes (highlighted if p-value <0.0015 Bonferroni adjusted)

		Uniform DIF by AGE			**Uniform DIF by Gender**		
Item	**description**	**F**	**DF**	**prob**	**F**	**DF**	**prob**
Q1	open a jar	33.5661	1	**<0.0001**	85.4736	1	**<0.0001**
Q2	do heavy household chores	30.7267	1	**<0.0001**	19.8655	1	**<0.0001**
Q3	carry shopping or a briefcase	0.9992	1	0.3178	84.8174	1	**<0.0001**
Q4	wash your back	13.01	1	**0.0003**	1.1029	1	0.2939
Q5	use knife to cut food	4.1657	1	0.0415	1.1176	1	0.2907
Q6	recreational activity (golf, hammering, tennis)	19.9508	1	**<0.0001**	7.0288	1	0.0081
Q7	interference with social activities	21.2661	1	**<0.0001**	45.7023	1	**<0.0001**
Q8	limited work or usual activities	4.1182	1	0.0427	15.7035	1	**<0.0001**
Q9	arm, shoulder or hand pain	0.0243	1	0.8760	7.6343	1	0.0059
Q10	tingling in arm, shoulder or hand	26.0624	1	**<0.0001**	0.0008	1	0.9781
Q11	difficulty sleeping because of arm pain	5.7041	1	0.0171	10.1594	1	0.0015

Figure 1: Category probability curves showing examples of ordered and disordered thresholds

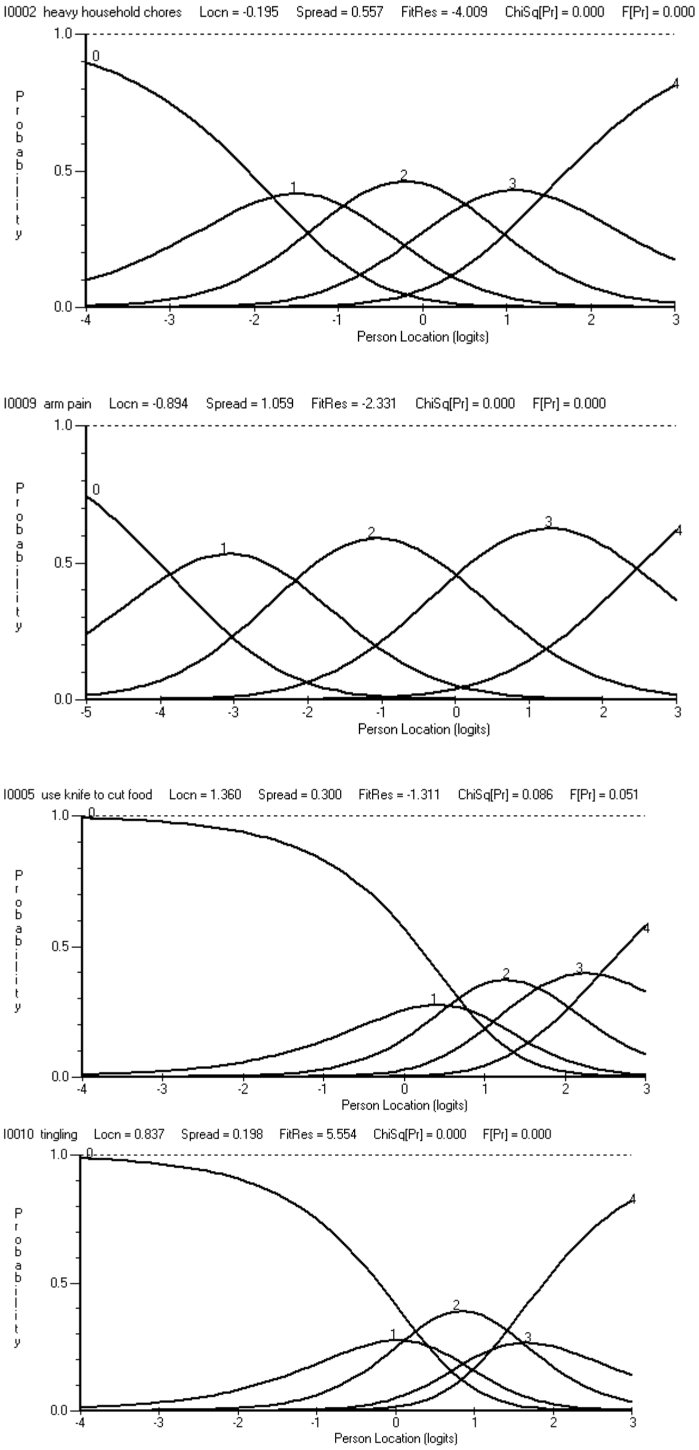


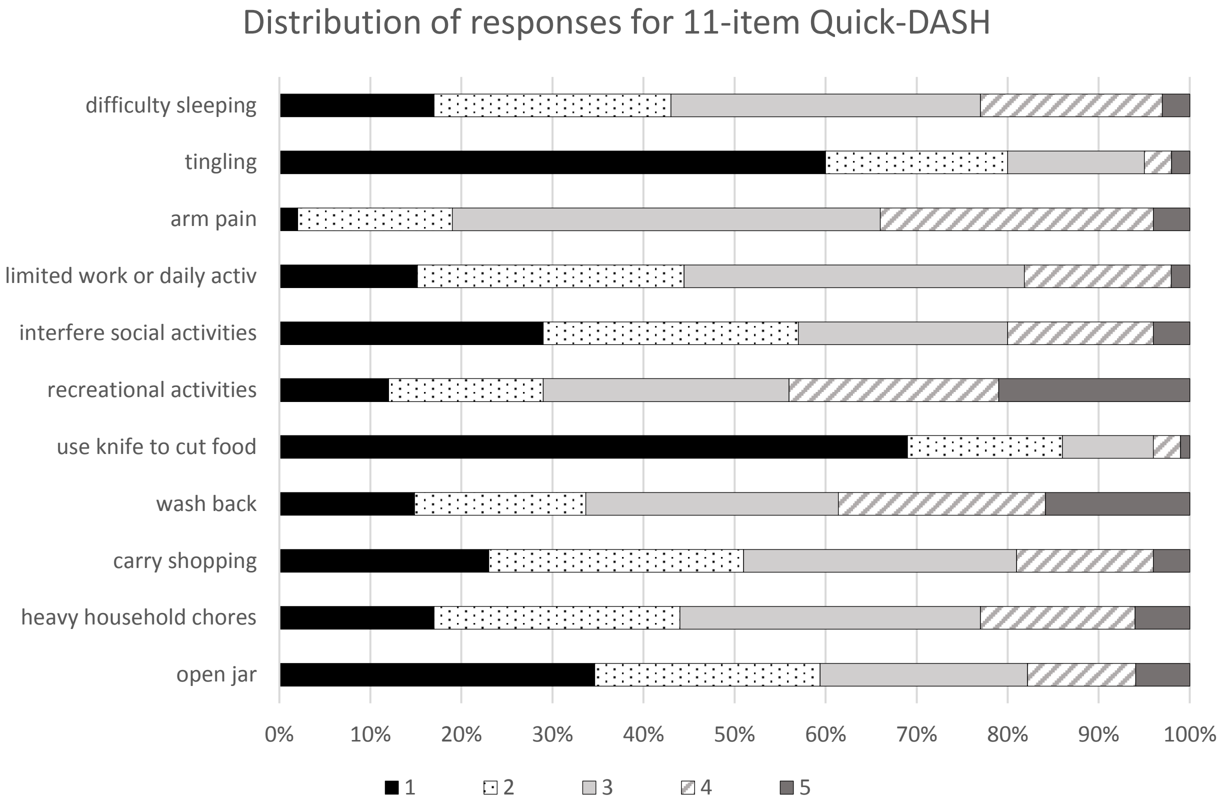
Fig 1 a: heavy household chores – ordered thresholds (each response category has a clear ‘peak’)

Figure 1b: arm, shoulder or hand pain – ordered thresholds

Figure 1 c: use a knife to cut food – disordered thresholds (first two response categories overlap)

Figure 1 d: Tingling – disordered thresholds (first two and last two response categories overlap)

Figure 2: Proportion (%) of responses in each category for 11 items (n=1030)



Score of 1 to 5 represents responses from ‘none’ or ‘no difficulty’ to ‘unable’ or ‘extreme’

Figure 3: Threshold map for all 11 items in location order with Q5 and Q10 rescored as 00112

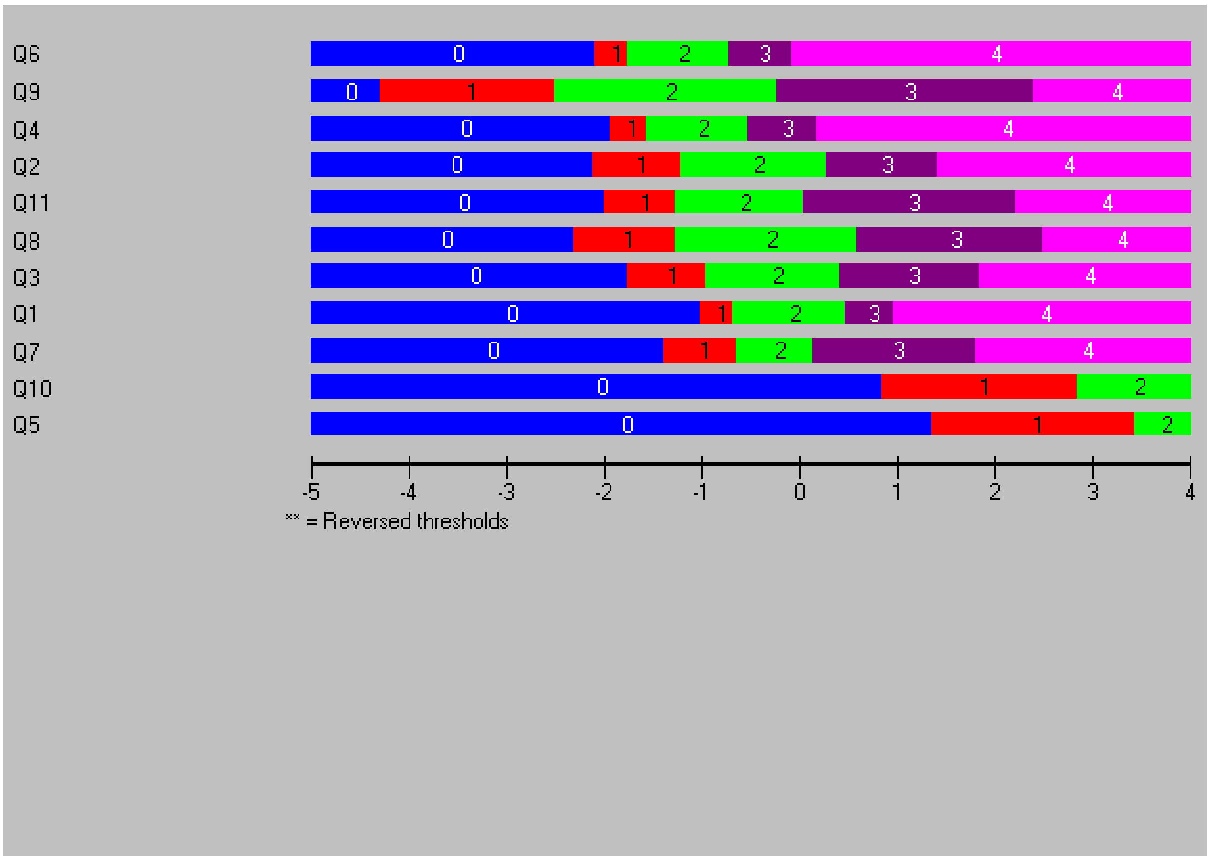


Figure 4: Person-Item threshold map of Quick-DASH with 2 testlets (analysis 3)

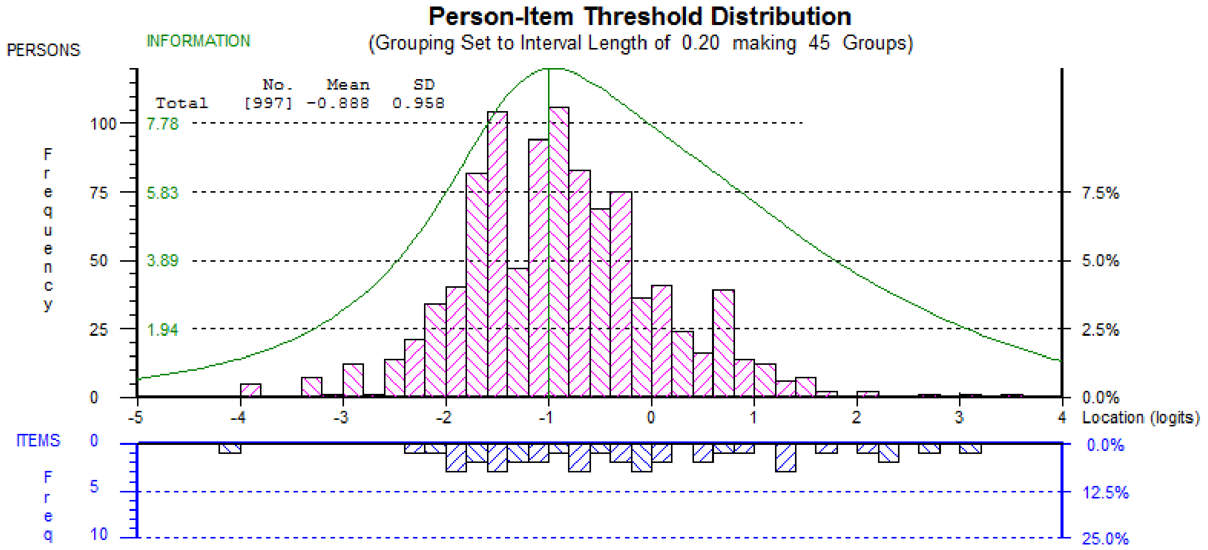


Figure legend: Quick-DASH represented on bottom of histogram and persons on the top. There is a very good overlap between person ability and item difficulty with a spread of difficulty within -4 and +3 logits. The frequency of persons is skewed to the left which represent more able persons and easier items.