Full Length Article

JHS(E)

Journal of Hand Surgery
(European Volume)
2022, Vol. 47(7) 750–754

DOI: 10.1177/17531934221081803
journals.sagepub.com/home/jhs

A comparative analysis of multidimensional computerized adaptive testing for the DASH and QuickDASH scores in Dupuytren's disease

Conrad Harrison[1], Andrew D. Clelland[2], Tim R. C. Davis[3], Brigitte E. Scammell[3,4], Weiya Zhang[4], Peter Russell[5], Sue Fullilove[6], Indranil Chakrabarti[7], Dominique Davidson[8] and Jeremy Rodrigues[9]

Abstract

The QuickDASH is a short-form version of the DASH questionnaire, the most widely used patient-reported outcome measure in hand surgery. Multidimensional computerized adaptive testing (MCAT) can produce shorter and more precise testing than static short forms, like QuickDASH. We used DASH responses from 507 patients with Dupuytren's disease to develop a MCAT. The algorithm was evaluated in a Monte Carlo simulation, where the standard error of measurement (SEm) of scores obtained from the 11-item QuickDASH was compared with scores obtained from an MCAT that could administer up to 11 items from the full 30-item DASH. The MCAT asked a mean of 8.51 items (SD 2.93) and 265/1000 simulated respondents needed to complete ≤five items. Median SEms were better for DASH MCAT: 0.299 (hand function) and 0.256 (sensory symptoms) versus 0.320 and 0.290, respectively, for QuickDASH. Our study showed that the DASH MCAT can produce more precise DASH measurement than the QuickDASH, from fewer items.

Keywords

Patient-reported outcome measures, PROM, multidimensional, computerized adaptive testing, CAT, Dupuytren's disease

Date received: 19th November 2021; revised: 2nd February 2022; accepted: 3rd February 2022

Introduction

Patient-reported outcome measures (PROMs) are important in hand surgery as they capture the patients' perspective. Comprising 30 items, the Disabilities of the Arm, Shoulder and Hand (DASH) score has been reported as the most used PROM in hand surgery research (Lloyd-Hughes et al., 2019). An abbreviated (short form) version, known as the QuickDASH, was derived by Beaton et al. (2005). This was produced by concept-retention method, an item reduction approach, which resulted in an 11-item outcome measure that performed comparably on a psychometric basis with the full DASH score when measuring patient symptoms and disability.

In patients with Dupuytren's disease, the DASH and QuickDASH measure two distinct domains, which might be interpreted as motor function and

[1]Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK
[2]Department of Surgery, University of Cambridge, Cambridge, UK
[3]Nottingham University Hospitals NHS Trust, Nottingham, UK
[4]Academic Unit of Injury, Inflammation and Recovery Sciences School of Medicine, University of Nottingham, Nottingham, UK
[5]Derby Teaching Hospitals NHS Foundation Trust, Derby, UK
[6]University Hospitals Plymouth NHS Trust, Plymouth, UK
[7]Rotherham NHS Foundation Trust, Rotherham, UK
[8]NHS Lothian, Livingston, UK
[9]Clinical Trials Unit, Warwick Medical School, Warwick, UK

Corresponding Author:
Conrad Harrison, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK.
Email: Conrad.harrison@medsci.ox.ac.uk
Twitter: @conrad_harrison; @AndrewDClelland; @mrjnrodrigues

sensory symptoms (i.e. pain and paraesthesia) (Rodrigues et al., 2016; Stirling et al., 2021). The application of item response theory might improve the validity of DASH scores, while not changing the questionnaire items or response options themselves. Unlike classical psychometric test theory, item response theory uses probabilistic modelling (statistical equations) to assess the measurement properties of individual items, allowing scores from the questionnaire to be calibrated on a true continuous (rather than ordinal) scale. Furthermore, item response theory can quantify the reliability of the score for each individual, with a personalized confidence interval, based on a reliability statistic called standard error of measurement (SEm), which measures the spread of possible 'true' scores for a given test score, similarly to standard deviation. Finally, item response theory items function independently; even when participants answer different combinations of items from within the PROM, their scores can still be validly compared. This is the basis of computerized adaptive testing (CAT).

By adapting the testing to the individual patients, CAT uses algorithms to make PROMs shorter and more personalized. The algorithms administer an item, predict a person's score after each item, then select the next best item for a person, based on their predicted score. This continues until a stopping rule is met, for example, a level of reliability has been reached.

Multidimensional CAT (MCAT) is a more advanced, and potentially more efficient form of CAT that uses multidimensional item response theory. In this theory, each questionnaire item can measure more than one health trait at a time, allowing data entry simultaneously on two different continuous scales. For example, in MCAT, items could measure both constructs (pain and function) on two different scales at the same time. This is more accurate and more valid than blending information about someone's pain and function into one numerical value. This is particularly relevant to Dupuytren's disease, which tends to impact motor function more than sensory symptoms.

The aim of this study was to develop an MCAT for the DASH in Dupuytren's disease and use a simulation experiment to compare this to QuickDASH in terms of measurement reliability (SEm) and the number of questions asked.

Methods

Data collection

Full-length DASH responses were obtained from 760 patients with primary or recurrent Dupuytren's disease, collected across five hand centres in the United Kingdom (UK) as part of an exercise independently approved as service evaluation at each participating site (Rodrigues et al., 2016). The multidimensional item response theory model and MCAT algorithm were then calibrated from responses to these full-length questionnaires. Responses from different individuals were collected at different time points: prior to surgery, and at 3 weeks, 6 weeks, 1 year and 5 years postoperatively.

Multidimensional item response theory modelling and MCAT simulation

Parameters were calculated for a multidimensional item response theory model (specifically, a graded response model), which allowed certain items to measure both motor function and sensory symptoms at the same time.

The multidimensional item response theory model was then used to create a MCAT algorithm in the R statistical computing environment from the real-world dataset. This was then tested in a simulated trial, using a simulated dataset of 1000 responses to the full-length DASH. This simulated dataset was based on bootstraps of the original sample data, to ensure it was realistic for patients with Dupuytren's disease.

During the simulation, the MCAT algorithm was able to pick any item from the full-length DASH, in any order. The MCAT continued to ask questions until it had either asked 11 items (i.e. the same number as the usual QuickDASH), or until it could measure both hand motor function and sensory symptoms with an SEm <0.3, which approximately equates to a marginal reliability of >0.90 (Walter, 2009). The DASH items asked during each simulated MCAT assessment were determined by the algorithm. These were chosen on a person-by-person basis and were not necessarily the same questions as those included in the QuickDASH, even when 11 items were posed. The MCAT algorithm was deliberately constrained to ask up to 11 items; this meant that if the measurement precision threshold (SEm <0.3) was not reached after 11 items, the MCAT algorithm would stop and a fair comparison could still be drawn between the 11-item QuickDASH and an 11-item MCAT algorithm. Otherwise, the MCAT algorithm may have achieved greater precision than the QuickDASH, but from more items. In that case, it would not be clear which approach achieved a preferable balance of measurement precision and response burden.

The SEms for motor function and sensory symptoms were recorded, along with the number of items

administered to each simulated respondent. The MCAT SEms were then compared with those that would have been obtained by using the 11 QuickDASH items for each of our simulated respondents, as if they had completed the usual QuickDASH static short form. Additional information for the development of the MCAT, including factor analysis and model fit, is included in the online Supplementary information (appendices S1 and S2).

Results

Data collection

Following listwise exclusion, 507 of the original 760 respondents were included in the analysis. Of the 253 excluded participants, 141 were excluded for failing to respond to item 21, which refers to sexual function in the original DASH questionnaire, and 112 were excluded for missing responses to other items. Demographics and treatment details are as summarized in Table 1.

Multidimensional item response theory modelling and MCAT simulation

The MCAT asked a mean of 8.51 items (SD 2.93) and as few as three items in some cases, 265/1000 simulated respondents needed to complete five items or fewer (Table 2). This compares to the QuickDASH short form where 11 items were administered in all cases.

The median SEms of QuickDASH were 0.320 for hand function and 0.290 for sensory symptoms, whereas the median SEms of DASH MCAT were 0.299 for hand function and 0.256 for sensory symptoms, indicating better precision than the QuickDASH.

Overall, we found that MCAT works most efficiently in patients with moderate–poor motor function and moderate–severe sensory symptoms, as illustrated in Figure 1. In our sample, the median logit score for hand function was 0.039 (IQR –0.954 to 0.859) and for sensory symptoms was 0.035 (IQR –0.833 to 0.547). These scores are lower than where MCAT is expected to have peak performance (Figure 1), meaning that MCAT is likely to perform even more efficiently in groups with more severe symptoms than were described by our cohort.

Discussion

In this study, we used multidimensional item response theory to develop a MCAT system to deploy the DASH in patients with Dupuytren's disease. We achieved a way of collecting PROM information that was more structurally valid and more reliable (as reflected by lower SEm) than QuickDASH, even though fewer questions were often posed. By intelligently selecting the most relevant items for an individual, this system personalizes the assessment to the respondent. The precision achieved in our simulation (SEm <0.3) is comparable with that achieved by the Patient-Reported Outcomes Information System (PROMIS) measures, which are widely considered to have excellent psychometric properties (Hung et al., 2013), and equivalent to an over 90% reliability. If feasible for deployment in clinical practice, an MCAT system of modelling, like the one demonstrated in this study, could potentially

Table 1. Demographics and clinical details, where available, for included participants ($n = 507$).

Demographic variable	Value
Gender (number)	
Male	431
Female	76
Age (years)	
Median	67
IQR	61–73
Procedure (number)	
Dermofasciectomy	88
Fasciectomy	315
Needle aponeurotomy	104

IQR: interquartile range.

Table 2. Numbers of items posed by the multidimensioanl computerized adaptive testing algorithm before the stopping rule was reached for simulated individuals.

Number of items posed in MCAT	Number of simulated individuals posed that number of items
3	31
4	120
5	133
6	85
7	38
8	28
9	32
10	20
11	513

MCAT: multidimensional computerized adaptive testing.

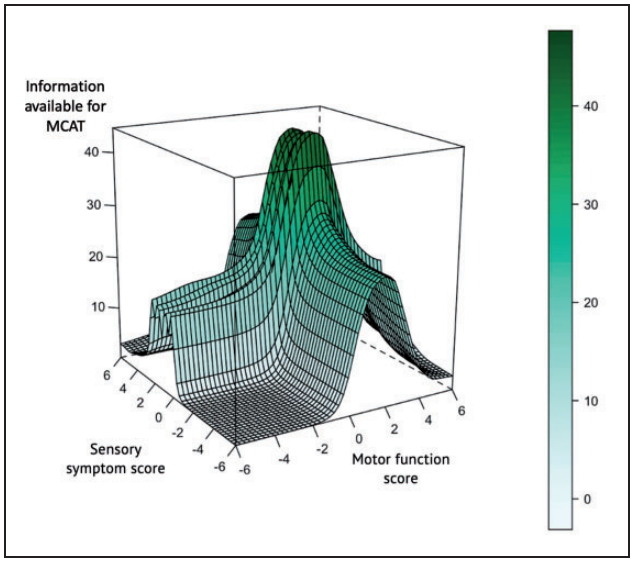

Figure 1. Multidimensional computerized adaptive testing (MCAT) efficiency at different test scores. The *X* and *Z* axes represent a person's motor function score and sensory symptom score, respectively, while the *Y* axis represents the amount of information available to the MCAT for efficient decision making. In this figure, scores are measured on a continuous logit scale, with a higher score indicating a poorer clinical state. The information peak occurs at a medium–high score in both motor function and sensory symptoms, meaning that the MCAT works most efficiently in patients with moderate–poor motor function and moderate–severe sensory symptoms.
MCAT: multidimensional computerized adaptive testing.

harness the multidimensional nature of DASH to get more valid PROM measurements for patients with Dupuytren's disease.

The multidimensionality of DASH has been criticized in the past, as blending sensory and motor scores into a single index of disability may be oversimplistic and introduce measurement error unnecessarily (Rodrigues et al., 2016). By using the MCAT system, the multidimensionality of DASH can be leveraged to administer the questionnaire more efficiently, while capturing two discrete scores. Given its greatly reduced burden for patients because of fewer questions that need answering, MCAT may also permit more frequent PROM sampling in research or clinical practice, thus providing deeper insights into day-to-day variations of symptom severity and functional impairment.

Differential analyses of the results from multidimensional PROMs like DASH and QuickDASH are particularly relevant to studying symptom severity and treatment effectiveness in a condition like Dupuytren's disease. Dupuytren's disease typically has a more severe impact on motor function than sensory symptoms (such as pain and paraesthesia) preoperatively. When Dupuytren's disease is treated surgically, we might expect to see an improvement in hand motor function, but it is also possible that the patients can experience postoperative pain. In this scenario, if we ignore the multidimensional nature of DASH and instead simply blend pain and function scores together as is often the case, it is possible for the improvement in function and deterioration in pain to cancel each other out somewhat. A person's scores may suggest no overall change has happened, when in fact two important changes have occurred – an improvement in function but a deterioration in pain. MCAT avoids this pitfall, providing more granular and actionable measurements.

In our study, we found that items 22, 23, 27 and 30 of the DASH capture information from both domains (Tables S1, S3, and S3 in the supplementary material). As such, we were able to sense check the data-driven suggestions that came from the analysis with clinical reasoning, for example, we deduct that items 22, 23 and 30 relate to social activities, work activities and confidence, respectively. Therefore, they are less explicit than the task-based or sensory symptom items and may reflect both sensory symptoms and impaired motor function. It is also possible that the word 'weakness' in item 27 was interpreted in different ways by different respondents. Qualitative interviews with affected patients could investigate these hypotheses in future.

Our study population had a relatively low degree of functional impairment (median logit score 0.039) and mild sensory symptom severity (median logit score 0.0352). As such, these patients will fall into the low-test information zone (Figure 1). Therefore, the algorithm may be functioning at a low level of efficiency in this cohort. In contrast, patients with conditions leading to 'worse' DASH scores, such as carpal tunnel syndrome, who may present with more severe sensory symptoms and higher degree of functional impairment, would have scores that are distributed at a higher level on both scales (i.e. closer to the zenith of the information curve, Figure 1). This means that our MCAT algorithm is likely to work even more efficiently for patients with conditions like carpal tunnel syndrome, although this remains to be proven.

While it might seem appealing to have allowed the MCAT algorithm to run with an SEm-based stopping rule only, so that more than 11 items could be presented, in this study we were specifically interested in the comparison with QuickDASH, and the broader question as to whether an item response theory-based questionnaire design is effective. We therefore elected to constrain the MCAT algorithm

to a maximum of 11 items, as discussed in the methods.

Despite the several advantages to using an MCAT, there are challenges to its implementation in clinical practice. For one, the MCAT is limited by the need for a computer/mobile device to provide a score, whereas QuickDASH may be completed with pen and paper. However, in the post-COVID era, remote monitoring through electronic PROMs may be particularly appealing. The parameters we have presented in our supplementary material may easily be applied as a mobile application to enable patient-friendly, remote reporting of symptoms.

There are limitations in this study. It is assumed that the order of items does not affect item response (in our simulation, all item responses were predetermined), however, studies from other fields suggest the impact of this is minimal (Li et al., 2012). Also, our study population is limited to patients with Dupuytren's disease who were treated in the UK, and the generalizability of our findings to other conditions and demographics remains unconfirmed.

In conclusion, by applying multidimensional item response theory to DASH responses from patients with Dupuytren's disease, an MCAT can be developed that can be more accurate and less burdensome than QuickDASH. In clinical practice, this could take the form of a smartphone application that administers frequent, short and personalized versions of DASH to patients so that hand surgeons or physiotherapists might monitor them remotely. In research, MCAT might improve DASH or QuickDASH completion rates (by lowering the response burden of the questionnaire) and provide higher quality measurement in clinical trials (by accounting for the multidimensional nature of DASH and QuickDASH). Future work may test the acceptability of these tools, and their generalizability across other conditions and patient groups. Similar work could also apply contemporary psychometric techniques to other existing PROMs, and the scoring of tools originally developed without these methods might be updated to reanalyse existing datasets, for example from previous trials.

Declaration of conflicting interests The authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Conrad Harrison is funded by National Institute for Health Research (NIHR) Doctoral Research Fellowship NIHR300684. Jeremy N. Rodrigues is funded by a NIHR Postdoctoral Fellowship PDF-2017-10-075. The views expressed in this publication are those of the author(s) and not necessarily those of the NIHR, NHS or the UK Department of Health and Social Care.

Ethical approval Ethical approval was not sought for the present study because secondary use of anonymized data primarily collected for service evaluation purposes does not require further ethical approval for secondary research use. Permission to work with the DASH was obtained. This study was completed in accordance with the Helsinki Declaration as revised in 2013.

ORCID iD Conrad Harrison https://orcid.org/0000-0002-1428-5751

Supplemental material Supplemental material for this article is available online.

References

Beaton DE, Wright JG, Katz JN Upper Extremity Collaborative Group. Development of the QuickDASH: comparison of three item-reduction approaches. J Bone Joint Surg Am. 2005, 87: 1038–46.

Hung M, Baumhauer JF, Latt LD et al. Validation of PROMIS® physical function computerized adaptive tests for orthopaedic foot and ankle outcome research. Clin Orthop Relat Res. 2013, 471: 3466–74.

Li F, Cohen A, Shen L. Investigating the effect of item position in computer-based tests: effect of item position in computer-based tests. J Educ Meas. 2012, 49: 362–79.

Lloyd-Hughes H, Geoghegan L, Rodrigues J et al. Systematic review of the use of patient reported outcome measures in studies of electively-managed hand conditions. J Hand Surg Asian Pac Vol. 2019, 24: 329–41.

Rodrigues J, Zhang W, Scammell B et al. Validity of the disabilities of the arm, shoulder and hand patient-reported outcome measure (DASH) and the QuickDASH when used in Dupuytren's disease. J Hand Surg Eur. 2016, 41: 589–99.

Stirling PHC, McEachan JE, Rodrigues JN, Harrison CJ. QuickDASH questionnaire items behave as 2 distinct subscales rather than one scale in Dupuytren's disease. J Hand Ther. 2021, S0894-1130: 00181–2.

Walter OB. Adaptive tests for measuring anxiety and depression. In: Walter OB (Ed). *Elements of adaptive testing*. New York, Springer, 2009: 123–36.