Optimizing Oxford Shoulder Scores with computerized adaptive testing reduces redundancy while maintaining precision

an NHS England National Joint Registry analysis

From University Hospitals of Leicester NHS Trust, Leicester, UK

Cite this article:
Bone Joint Res 2024;13(8):
392–400.

DOI: 10.1302/2046-3758.
138.BJR-2023-0412.R1

Correspondence should be sent to Ahmed Barakat
ahmedharoonbarakat@gmail.com

A. Barakat,[1] J. Evans,[2,3] C. Gibbons,[4] H. P. Singh[3,5]

[1]University Hospitals of Leicester NHS Trust, Leicester, UK
[2]University of Exeter, Exeter, UK
[3]Exeter Hip Unit, Royal Devon University Healthcare NHS Foundation Trust, Exeter, UK
[4]Division of Internal Medicine, Department of Symptom Research, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA
[5]University Hospitals of Leicester NHS Trust, Leicester, UK

Aims

The Oxford Shoulder Score (OSS) is a 12-item measure commonly used for the assessment of shoulder surgeries. This study explores whether computerized adaptive testing (CAT) provides a shortened, individually tailored questionnaire while maintaining test accuracy.

Methods

A total of 16,238 preoperative OSS were available in the National Joint Registry (NJR) for England, Wales, Northern Ireland, the Isle of Man, and the States of Guernsey dataset (April 2012 to April 2022). Prior to CAT, the foundational item response theory (IRT) assumptions of unidimensionality, monotonicity, and local independence were established. CAT compared sequential item selection with stopping criteria set at standard error (SE) < 0.32 and SE < 0.45 (equivalent to reliability coefficients of 0.90 and 0.80) to full-length patient-reported outcome measure (PROM) precision.

Results

Confirmatory factor analysis (CFA) for unidimensionality exhibited satisfactory fit with root mean square standardized residual (RSMSR) of 0.06 (cut-off ≤ 0.08) but not with comparative fit index (CFI) of 0.85 or Tucker-Lewis index (TLI) of 0.82 (cut-off > 0.90). Monotonicity, measured by H value, yielded 0.482, signifying good monotonic trends. Local independence was generally met, with Yen's Q3 statistic > 0.2 for most items. The median item count for completing the CAT simulation with a SE of 0.32 was 3 (IQR 3 to 12), while for a SE of 0.45 it was 2 (IQR 2 to 6). This constituted only 25% and 16%, respectively, when compared to the 12-item full-length questionnaire.

Conclusion

Calibrating IRT for the OSS has resulted in the development of an efficient and shortened CAT while maintaining accuracy and reliability. Through the reduction of redundant items and implementation of a standardized measurement scale, our study highlights a promising approach to alleviate time burden and potentially enhance compliance with these widely used outcome measures.

Article focus

- Patient-reported outcome measures (PROMs) provide valuable insights for the healthcare community, but often have redundancy issues.

- The Oxford Shoulder Score (OSS), despite its validity, has low patient engagement.
- This study explored the use of item response theory (IRT) and computerized adaptive testing (CAT) to reduce patient burden while maintaining accuracy in assessing shoulder conditions.

Key messages

- After assessing over 16,000 OSS, CAT testing has demonstrated > 80% reduction in the 12-item PROM to a median of two items with 80% precision and a median of three items with 90% precision compared to the full-length questionnaire.
- It was shown in this study that the application of modern psychometric analysis to the world's largest repository of shoulder arthroplasty PROMs has led to 80% reduction in the number of items required to estimate the patient-specific impact of shoulder disease and without compromising precision.

Strengths and limitations

- Embracing innovative methods such as CAT for assessing the OSS presents a chance to enhance patient-centred outcome evaluation and streamline data collection.
- This promises better patient compliance when completing these questionnaires, and reduced the burden on the National Joint Registry (NJR) for England, Wales, Northern Ireland, the Isle of Man, and the States of Guernsey administrators when compiling the data and scores (saving a total of 49,450 minutes (34.27 days) per year).
- The broader practical implications, including resource allocation and cost-effectiveness, require further exploration to assess the feasibility of implementing these innovative methods in routine clinical practice.

Introduction

Patient-reported outcome measures (PROMs) have become invaluable tools in the healthcare sector, providing important insights into the impact of a treatment from the patient's own perspective. These standardized assessments play a crucial role in understanding the effectiveness of medical interventions and improving patient care. One substantial challenge associated with the use of PROMs is the potential for redundancy, wherein patients are required to answer multiple similar questions or undergo repetitive time-consuming assessments. There is a mounting demand for more representative PROMs that avoid data redundancy and burden on both the patients and healthcare system through more effective data collection. These aspects have most recently been highlighted as part of the Cumberlege report and recommendation to NHS England public consultation on the current PROM system.[1,2]

The Oxford Shoulder Score (OSS) is a validated PROM used to assess shoulder pain and function.[3] It has 12 items, with each item offering five possible response options. The OSS ranges from 0 (representing the least favourable outcome) to 4 (representing the most favourable outcome), with a cumulative score ranging from 0 to 48. The National Joint Registry (NJR) in England, Wales, Northern Ireland, the Isle of Man, and the States of Guernsey has attempted to collect OSS data for individuals undergoing shoulder arthroplasties, encompassing preoperative assessments (time zero) as well as postoperative evaluations at six months, three years, and five years since 2012. The completion rates for the preoperative OSS were found to be as low as 45% and only 75% for the six-month scores, highlighting the limited patient engagement with these PROMs.[4]

Traditionally, PROMs assessment of non-observable latent traits such as pain and function are based on assumptions of classical test theory (CTT), where their True Score (TS) is the questionnaire's observed score plus a fixed measurement error applying to all patients indiscriminately. Additionally, in CTT, the entire questionnaire must be fulfilled to be scored. As each question is equally weighted, a potential problem is not being able to discern subtle changes in latent traits, as in the case of equal overall scores on different assessments despite changes in different domains or individual items balancing themselves out.

Alternatively, a more comprehensive approach is item response theory (IRT) methodology, where latent traits are assessed based on individual questions/item responses even if the questionnaire responses were truncated and not complete. This mathematical approach factors the items score as the main measure, rather than the test score as with CTT. This aims to provide more accurate and richer descriptions of latent trait performance, as well as helping to identify the most measure-reflective items. Moreover, in IRT, the measurement error is modelled at the individual level, meaning that for each pattern of responses one can quantify the potential error in the resulting score, thus leading to more accurate assessment of TSs. It has been reported that use of IRT can be more dependable than the conventional CTT approach for this reason.[5,6]

Applying IRT approach paves the way for use of computerized adaptive testing (CAT), where items in a PROM are delivered sequentially based on previous item responses, and the assessment stops whenever a predetermined desired measurement precision is reached. This dynamic approach ensures that patients are asked only relevant and informative questions, thus reducing the overall number of items needed to derive accurate and reliable scores. This has already been successfully implemented to other PROMs delivering fewer items while maintaining precision of a full-length questionnaire.[7]

In this context, the present study explored the implementation of CAT for the OSS. Exploring this model-based approach on OSS promises to reduce the number of questions, translating to reduced burden on patients and NJR database input, while avoiding an impact on PROM validity.

Methods

The anonymized preoperative NJR UK PROM dataset, spanning from its inception in April 2012 to April 2022, was used for this study. Initially, the dataset was partitioned into two distinct subsets: a 'calibration set' and a 'testing set' in a ratio of 80:20, respectively.[8]

Statistical analysis

The foundational assumptions of IRT — namely, unidimensionality, monotonicity, and local independence — were evaluated for each of the 12 OSS items.[9] Confirmatory

factor analysis (CFA) served to verify the unidimensionality of each item (R package 'LAVAAN' version 0.6 to 7; R Foundation for Statistical Computing, Austria), ensuring that it captured a single principal latent trait that fundamentally drives the observed responses. Monotonicity was assessed through Mokken scaling utilizing Loevinger's H coefficient ≥ 0.3 per item as a cut-off (R package 'Mokken' version 2.8.4; R Foundation for Statistical Computing), confirming that as the level of the latent trait increases, the likelihood of a higher corresponding response also rises.

Local independence was assessed through residual covariance. A high level of residual covariance may indicate that the items are too similar and therefore redundant, one item response affects the other, or that they together measure a second unintended latent trait. This was undertaken through an examination of the CFA residual correlation matrix with the Yen's Q3 statistic cut-off set to a correlation between two items of < 0.2 demonstrating locally independent items (i.e. response to one item will not affect response to other remaining items).

Having established the IRT model assumptions, a graded response model (GRM) was applied to the categorical item responses (R package 'mirt' version 3.3.2; R Foundation for Statistical Computing). The GRM entails two characteristics for each item: item discrimination, representing the item's ability to discriminate between people with similar latent trait levels; and item difficulty thresholds, which relate to the severity of each response. Model fit was evaluated through indices such as the root mean square error of approximation (RMSEA) ≤ 0.07 with 90% confidence intervals, and statistical significance ($p \leq 0.05$), root mean square standardized residual (RSMSR) ≤ 0.08, comparative fit index (CFI) and Tucker-Lewis index (TLI) > 0.9 to indicate good model fit.[10]

The reliability of this IRT model was quantified as marginal reliability and juxtaposed with the classical Cronbach's α estimate for the OSS using CTT, where values exceeding 0.8 indicate excellent reliability.

Subsequently, the constructed GRM model facilitated the development of a CAT simulation. This simulation enabled the comparison of precision, denoted by the standard error (SE) of the latent trait estimate, between shortened versions of the test (where specific items were selected based on their difficulty and discrimination attributes identified via GRM) and the full-length PROM. The simulation was automatically concluded upon achieving a predetermined SE level, conventionally set at SE < 0.32 and SE < 0.45 (equivalent to reliability coefficients of 0.90 and 0.80, respectively). This process aimed to unveil the most informative items required to attain the desired precision levels of 0.90 and 0.80, expressed as a percentage of their use in the simulation.

Variables derived from the simulation were the correlation (intraclass correlation coefficient (ICC)) between the latent trait estimation of the full-length questionnaire and the CAT, and the mean and SD, median and IQR of items required to derive estimates of the latent trait at the two levels of precision. The items selected by the CAT were reported by their percentage of use within the simulation. Differences in the item use between full-length and CAT administration are presented as percentage differences. Time-saving between full-length and CAT administration was calculated against the estimate that each item takes between ten and 75 seconds per item to complete.[11]

All data analysis was executed in RStudio (Rstudio PBC, USA). The CAT simulation was performed using Firestar for R (version 1.3.2; R Foundation for Statistical Computing). Chi-squared test was used to report statistical significance, with $p < 0.05$ considered statistically significant.

Results

Dataset analysis

A total of 16,238 preoperative OSS were available in the period from April 2012 to April 2022. Among these, 70% (n = 11,366) were female and 30% (n = 4,872) were male patients. The mean preoperative OSS was 16.7 (SD 8.6). For the purpose of calibration, any incomplete scores were excluded from the analysis, resulting in a final pool of 15,375 available for the final dataset. Consequently, the dataset has been divided into 12,300 instances for the calibration subset and 3,075 instances designated for CAT testing as per the predetermined 80:20 ratio.

IRT assumptions

CFA showed a good fit with RSMSR of 0.06 (cut-off ≤ 0.08) but not with CFI of 0.85 (cut-off > 0.90), TLI of 0.82 (cut-off > 0.90), or RMSEA of 0.11 (95% CI = 0.11 to 0.12; $p < 0.05$) (cut-off ≤ 0.07).

The analysis yielded a H value of 0.48 (range 0.30 to 0.54, SE 0.004), confirming monotonic items and that as the trait being measured (pain or function in this case) increases, the probability of a higher response category being selected also increases.

Local independence between items was confirmed for nine items with correlations < 0.2. There was, however, a local dependence of 0.33 between items 1 and 8, 0.25 between items 8 and 12, and 0.26 between items 1 and 12, indicating a moderate level of mutual influence between these items. The IRT modelling of the 12 OSS items is depicted in Figure 1.

GRM calibration

Item and GRM fit were assessed using the M2 statistic ($p < 0.05$). M2 was 798.15 ($p < 0.05$), SRMSR = 0.07, RMSEA was 0.05 (95% CI 0.04 to 0.05, $p < 0.05$)), CFI = 0.96, and TLI = 0.93, indicating a good model fit. The reliability of the IRT model was assessed using marginal reliability, which evaluates overall precision by averaging the mean SEs at different ability levels. This metric can be compared to Cronbach's α in CTT, with scores > 0.8 signifying excellent reliability. In this study, marginal reliability was calculated as 0.91.

CAT simulation

At the designated stopping points, specifically SE = 0.32 (i.e. precision of 90%), the mean number of items needed for CAT was 4.15 (SD 1.83). Similarly, when aiming for a SE of 0.45 (i.e. precision of 80%), the mean number of items required for CAT was 2.33 (SD 0.63, mean SE = 0.36).

The median number of items needed to complete the CAT simulation with a SE of 0.32 was 3 (IQR 3 to 12). Meanwhile, for a SE of 0.45, the median number of items required was 2 (IQR 2 to 6) (Figure 2). This median number of items constituted only 25% and 16.6% of the original 12-item full-length questionnaire at a SE of 0.32 and 0.45, respectively.

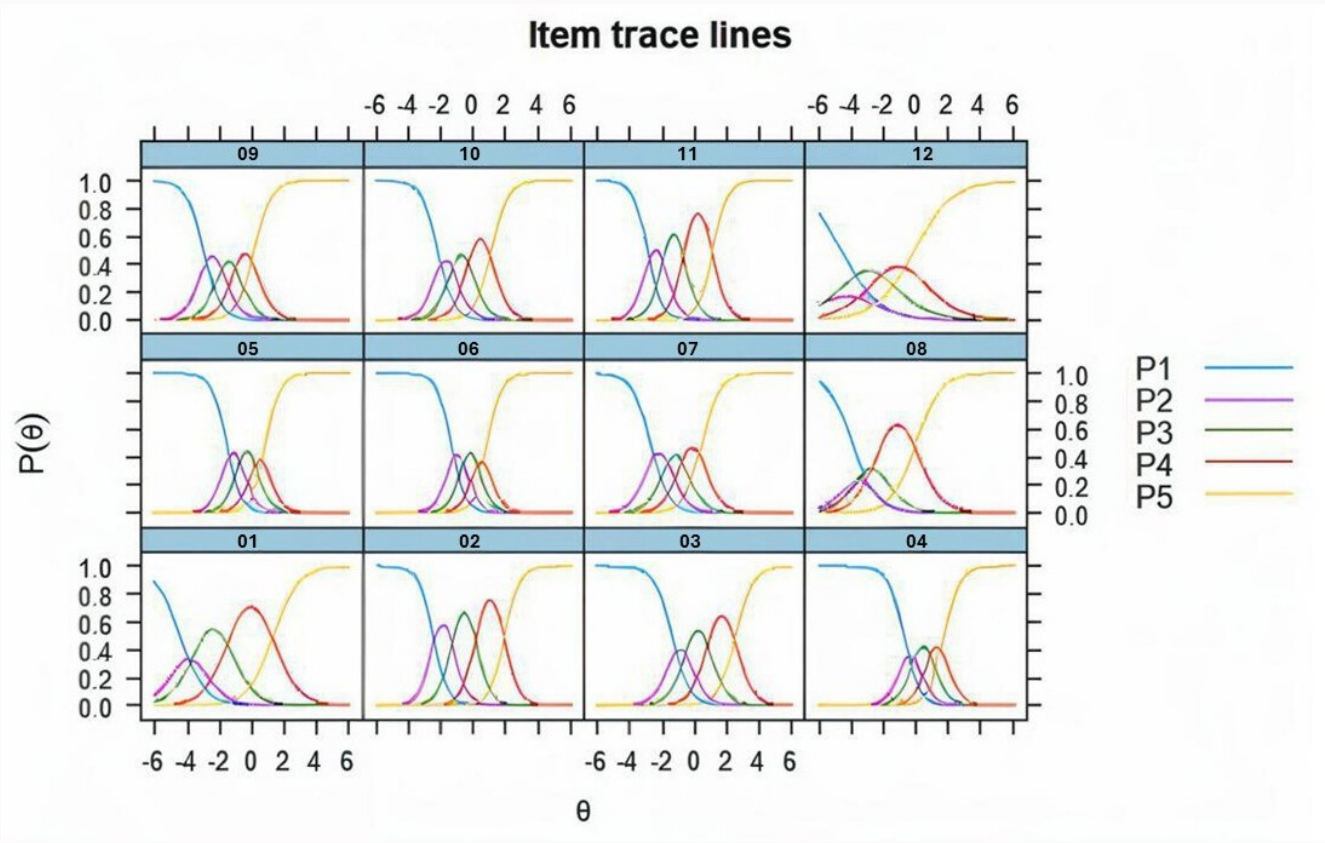

Fig. 1
Item response theory (IRT) trace line graphs for the 12 items of the Oxford Shoulder Score (OSS). The x-axis of the graph represents the underlying latent trait (θ) (pain and function in this case). The y-axis of the graph represents the probability of a correct response to the item given a specific level of the latent trait (θ).

At SEs of 0.32 and 0.45, the ICC between the CAT and the full-length OSS latent trait estimates was r = 0.94 and r = 0.88, respectively (Figure 3).

This translates to time taken for the CAT assessment, ranging from < one minute (40 seconds) to five minutes (306 seconds) compared to a range from two minutes (120 seconds) to 15 minutes (900 seconds) for the full-length pen and paper questionnaire, assuming that it takes ten to 75 seconds per item.[11]

The items underwent quantification based on their use frequency, expressed as percentages (Figure 4).

During the adaptive testing process, item 6 ("During the past 4 weeks, could you carry a tray containing a plate of food across a room?") consistently served as the starting item in 100% (n = 3,075) of the simulations. After compiling all the items, it was found that item 6 constituted 25% (n = 769) of all the items used in the simulations. Following this, items 5 ("During the past 4 weeks, could you do the household shopping on your own?") and 4 ("During the past 4 weeks, have you been able to use a knife and fork - at the same time?") were chosen in 23% (n = 707) and 15% (n = 461) of the simulations, respectively. These observations suggest that these items exhibit higher discriminative power concerning the latent ability of interest. Conversely, items 8 ("During the past 4 weeks, how would you describe the pain you usually had from your shoulder?") and 12 ("During the past 4 weeks, have you been troubled by pain from your shoulder in bed at night?") demonstrated notably lower use rates, appearing in only 1% (n = 31) of the simulations. The observed results provide guidance concerning the varying effectiveness of items and their contributions to the precision of the adaptive test (Figure 5).

Discussion

Patient and provider interviews have been critical of the length of the questionnaires delivered by the national PROMs programme.[12] Furthermore, time taken to complete these PROMS has been recognized as a key factor affecting patients' responses and compliance.[13] Applying IRT modelling and subsequent use of CAT for OSS delivery will simulate individually tailored OSS assessment, with the goal of reducing the number of questions a patient must complete without compromising measurement accuracy. This electronic capture of data aims to reduce patient and administrative burden with potentially increasing completion compliance. In this study, we have demonstrated that applying CAT to the OSS substantially reduces the full-length questionnaire to just 25%, all while maintaining a high precision level of 90%. Notably, during the adaptive testing process in our study, certain items were rarely used in the simulation, such as items 8 and 12 appearing in only 1% (n = 31) of simulations. This trend may be attributed to the fact that these items either possess extremely low or excessively high difficulty levels. CAT simulations mainly utilize items of moderate difficulty as the primary items in the sequence, as these items are best targeted to the majority of respondents. The most frequently

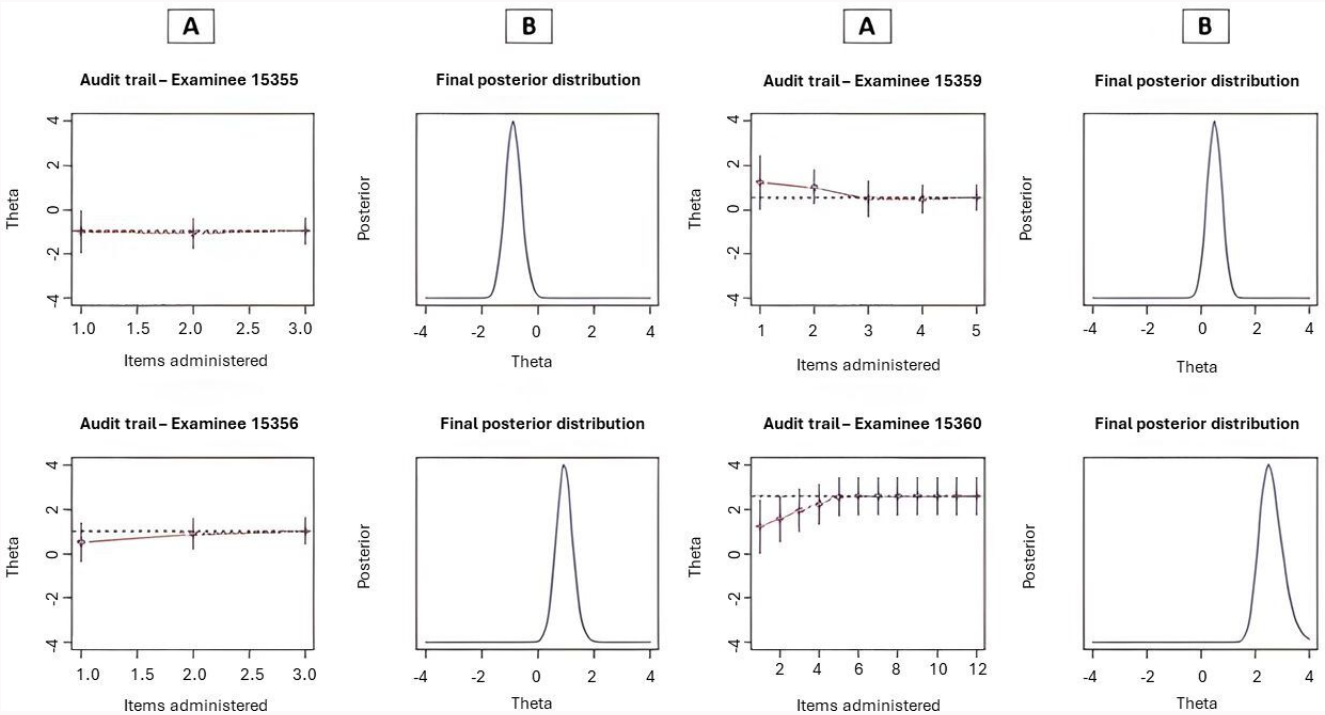

Fig. 2

Computerized adaptive testing (CAT) graphical representations for four Oxford Shoulder Scores (OSS). a) Items were adaptively selected until the patient's latent trait (θ) was reached with a predetermined level of accuracy (standard error). b) The final posterior distribution represents the probability distribution of the examinee's true ability level after completing the CAT session.

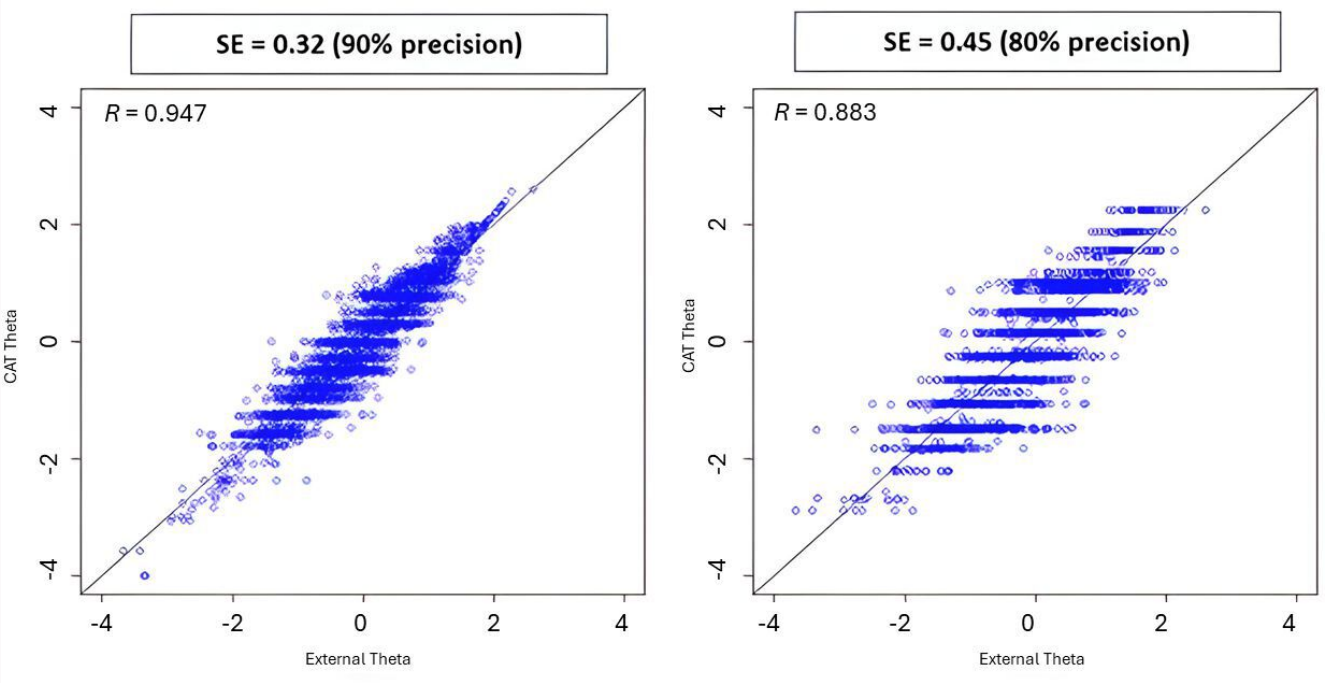

Fig. 3

Scatter plots depicting the intraclass correlation coefficient (ICC) between the computerized adaptive testing (CAT) and full-length questionnaire estimates of the latent trait (r = 0.94 in the case of standard error (SE) = 0.32, r = 0.88 in the case of SE = 0.45).

used items in this study were items 6, 5, and 4 appearing in 25% (n = 767), 23% (n = 707), and 15% (n = 461), respectively. It is worth mentioning that those items most frequently used were related to function rather than pain. This trend is consistent with other similar qualitative PROM assessments such as in the Oxford Hip Score (OHS) and Oxford Knee Score (OKS).[14] It has been postulated that patients prefer questions

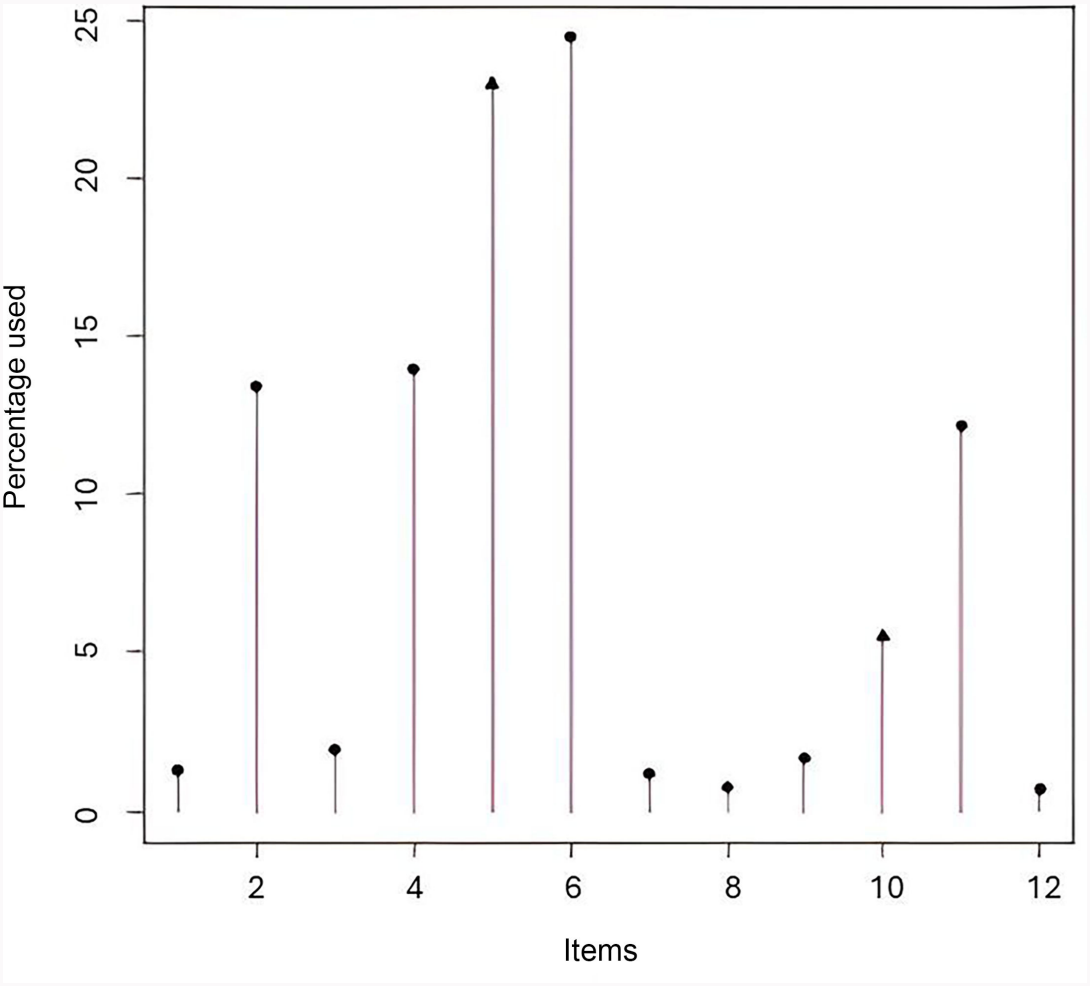

Fig. 4
Line graph representing use percentage for each item during the computerized adaptive testing as a proportion of the total number of items used. Note that item 6 was the starting item for 100% of the simulations, and it constituted a 25% proportion of all items used collectively.

about their function because they find them clearer and easier to answer compared to questions about their pain.

Applying IRT to legacy PROMs is not new. The Patient-Reported Outcomes Measurement Information System (PROMIS) network is an example of how IRT modelling and CAT administration have been used, with a plethora of published literature demonstrating greater reliability and validity, as well as decreased cost and collection resources when compared to conventional PROMs.[15,16] This methodology has been applied on two other Oxford Innovation scores, the OHS and the OKS, demonstrating large reductions in the number of items required to estimate the latent trait, without compromising efficiency.[17] In the context of shoulder PROMs, CAT was similarly applied to 2,763 American Shoulder and Elbow Shoulder Assessments (ASES), demonstrating a 40% reduction compared to the full-length ASES with negligible effect on score integrity (ICC = 0.99).[18] CAT implementation to the OSS was recently assessed by Harrison et al[8] using machine-learning algorithms on a sample size of 561 scores. Although the sample scores in that study were heterogenous (preoperative and postoperative scores) and not solely arthroplasty patients, the authors nevertheless demonstrated high accuracy with CAT modelling. Our study has implemented the entire NJR preoperative dataset for arthroplasty OSS spanning a ten-year period and encompassing more than 15,000 preoperative scores, substantially enhancing reported accuracy of the implemented CAT simulation for this specific shoulder arthroplasty population. With the established satisfactory IRT model fit to the OSS, it paves the way to allow placing individuals and items on the same scale of measurement. This means that regardless of the specific test or set of items being used, IRT provides a way to compare individuals' abilities or traits on a standardized scale. This is one of the strengths of IRT, and allows for the comparison of individuals' scores across different tests or measurement instruments that are designed to measure the same underlying trait. This 'cross-walking' between scores has already been demonstrated between different PROMs to produce equivalent scores.[19,20]

From the patient and administrative perspectives, CAT efficiently homes in on the specific level of function or discomfort a patient experiences. This targeted questioning not only ensures a comprehensive assessment but also minimizes the number of questions needed, saving time for both patients and healthcare providers. For example, a patient might start by answering a general question about their overall shoulder function. If they report minimal limitations, the system might bypass questions about more severe functional impairments and concentrate on specific daily

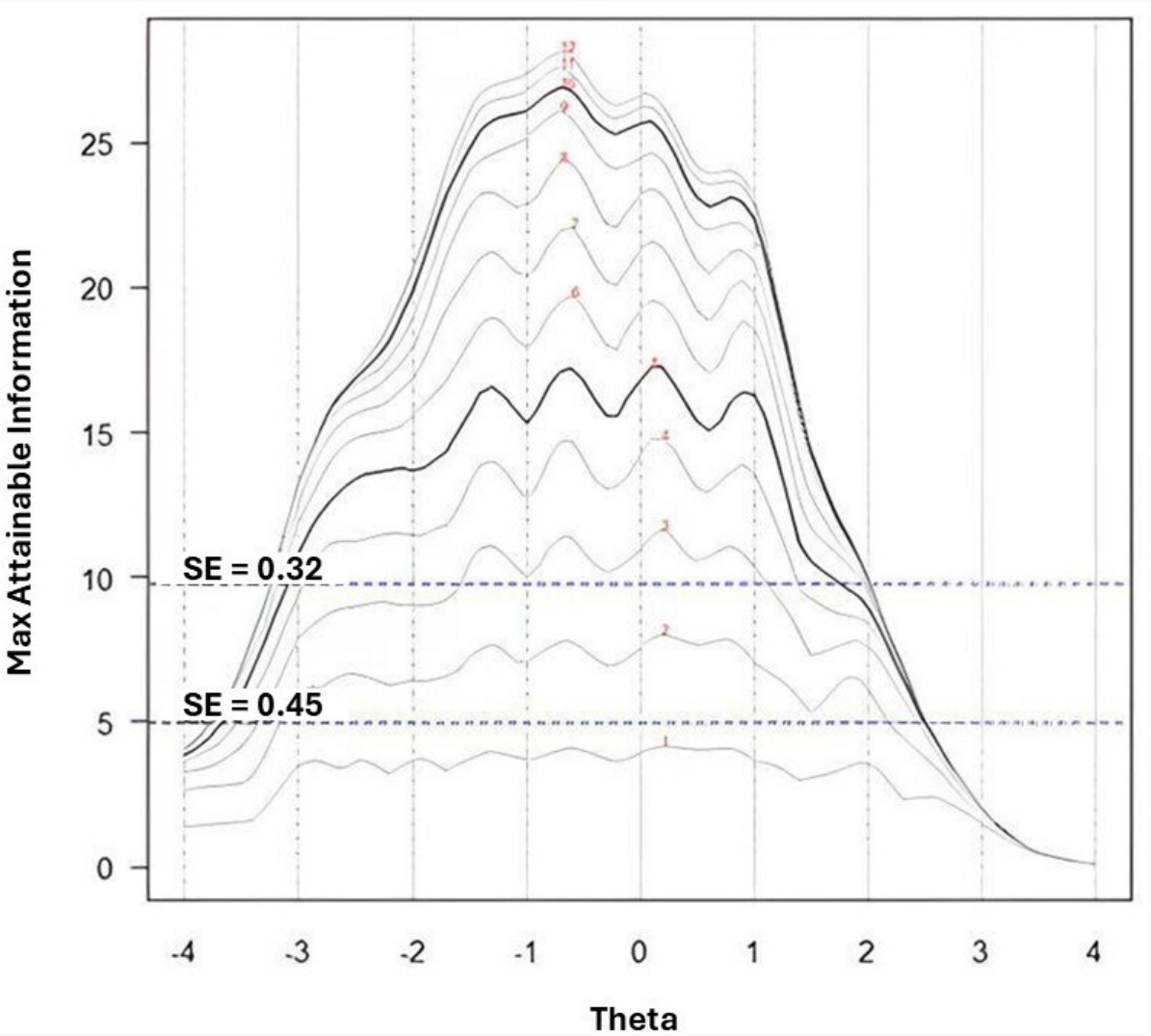

Fig. 5
Graph displaying the maximum attainable information over the trait continuum, for the 12-item Oxford Shoulder Score (OSS). The outermost curve represents the item pool information function, while the inner curves represent the maximum information that would result from the administration of a given number of items at the two predetermined standard errors of 0.32 and 0.45.

activities impacted by their mild limitations, thus tailoring the assessment. This not only expedites the process but also enhances patient engagement, making them more likely to complete the assessment accurately.

Our results suggest that CAT assessments can last anywhere from under one minute (40 seconds) to five minutes (306 seconds), in contrast to conventional pen-and-paper questionnaires, which typically take between two minutes (120 seconds) and 15 minutes (900 seconds) to complete based on an assumed time of ten to 75 seconds per item. This can also be extrapolated to substantial time-saving for the NJR administrators, cutting down analysis time which will now be computerized. Considering the 8,600 shoulder arthroplasties registered in the NJR for the year 2020 to 2021, this could amount to a total of 49,450 minutes (34.27 days) saved for the total 8,600 scores.[21]

We recognize that this modelling study has limitations inherent to its retrospective data analysis. Within the IRT modelling for the OSS, despite the single-dimension structure fitting satisfactorily on RSMSR, the RMSEA did not meet the cut-off criterion of ≤ 0.07, while CFI and TLI fell short of the > 0.9 cut-off. We ascribe this to the composite nature of the OSS measuring two traits (pain and function) rather than a single unidimensional score. This instrument has been widely adopted in clinical practice with previous CFA conducted by the original designers, supporting its appropriateness for such composite usage while recognizing both single- and two-factor models.[22] Future work could utilize a multidimensional IRT methodology, which may optimize the difficulty and discrimination estimation. Additionally, local independence was demonstrated for all but three items. This highlights the potential for redundancy due to the interdependence between these items being too similar to each other. As the OSS has become a standardized assessment, at this stage we would not advocate removal of items before further validation studies have been performed. We also recognize that, as an unmandated score, an element of selection bias may exist in the analyzed sample, however this remains one of the world's largest repositories of arthroplasty-related shoulder scores. From a resource perspective, the implementation of

CAT, where the questions must be delivered electronically, is often viewed as challenging. However, in the post COVID-19 era, the yielding of data through digital means continues to expand and evolve rapidly, and should therefore not be viewed as an insurmountable barrier.

In conclusion, by embracing innovative approaches such as CAT for the assessment of OSS, there is an opportunity to attain more effective, patient-centred outcome evaluation and optimize costly data collection resources. The application of modern psychometric analysis to the world's largest repository of shoulder arthroplasty PROMs has isolated a potential 80% reduction in the number of items required to estimate the patient-specific impact of shoulder disease, without compromising precision. Further implementation studies will be needed to validate the delivery of OSS using this computerized technique over its current pen-and-paper form.

References

1. **Haskell H**. Cumberlege review exposes stubborn and dangerous flaws in healthcare. *BMJ*. 2020;370:m3099.
2. **No authors listed**. National Patient Reported Outcome Measures (PROMs) Programme Consultation. NHS England. 2016. https://www.engage.england.nhs.uk/consultation/proms-programme (date last accessed 12 July 2024).
3. **Dawson J**, **Rogers K**, **Fitzpatrick R**, **Carr A**. The Oxford shoulder score revisited. *Arch Orthop Trauma Surg*. 2009;129(1):119–123.
4. **Achakri H**, **Ben-Shlomo Y**, **Blom A**, **et al**. The National Joint Registry 20th Annual Report 2023. London: National Joint Registry. 2023. https://reports.njrcentre.org.uk/Portals/0/PDFdownloads/NJR%2020th%20Annual%20Report%202023.pdf (date last accessed 21 June 2024).
5. **Jabrayilov R**, **Emons WHM**, **Sijtsma K**. Comparison of classical test theory and item response theory in individual change assessment. *Appl Psychol Meas*. 2016;40(8):559–572.
6. **Singh HP**, **Haque A**, **Taub N**, **et al**. Floor and ceiling effects in the Oxford Shoulder Score: an analysis from the National Joint Registry. *Bone Joint J*. 2021;103-B(11):1717–1724.
7. **Kane LT**, **Namdari S**, **Plummer OR**, **Beredjiklian P**, **Vaccaro A**, **Abboud JA**. Use of computerized adaptive testing to develop more concise patient-reported outcome measures. *JB JS Open Access*. 2020;5(1):e0052.
8. **Harrison CJ**, **Plummer OR**, **Dawson J**, **Jenkinson C**, **Hunt A**, **Rodrigues JN**. Computerized adaptive testing for the Oxford Hip, Knee, Shoulder, and Elbow scores: accurate measurement from fewer, and more patient-focused, questions. *Bone Jt Open*. 2022;3(10):786–794.
9. **Reise SP**, **Revicki DA**. *Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment*. London: Routledge, 2014.
10. **Hu LT**, **Bentler PM**. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*. 1999;6(1):1–55.
11. **McMurray R**, **Heaton J**, **Sloper P**, **Nettleton S**. Measurement of patient perceptions of pain and disability in relation to total hip replacement: the place of the Oxford hip score in mixed methods. *Qual Health Care*. 1999;8(4):228–233.
12. **Kyte D**, **Cockwell P**, **Lencioni M**, **et al**. Reflections on the national patient-reported outcome measures (PROMs) programme: where do we go from here? *J R Soc Med*. 2016;109(12):441–445.
13. **Hutchings A**, **Neuburger J**, **Grosse Frie K**, **Black N**, **van der Meulen J**. Factors associated with non-response in routine use of patient reported outcome measures after elective surgery in England. *Health Qual Life Outcomes*. 2012;10:34.
14. **Wylde V**, **Learmonth ID**, **Cavendish VJ**. The Oxford hip score: the patient's perspective. *Health Qual Life Outcomes*. 2005;3:66.
15. **Haws BE**, **Khechen B**, **Bawa MS**, **et al**. The Patient-Reported Outcomes Measurement Information System in spine surgery: a systematic review. *J Neurosurg Spine*. 2019;30(3):405–413.
16. **Flynn KE**, **Dew MA**, **Lin L**, **et al**. Reliability and construct validity of PROMIS® measures for patients with heart failure who undergo heart transplant. *Qual Life Res*. 2015;24(11):2591–2599.
17. **Evans JP**, **Gibbons C**, **Toms AD**, **Valderas JM**. Use of computerised adaptive testing to reduce the number of items in patient-reported hip and knee outcome scores: an analysis of the NHS England National Patient-Reported Outcome Measures programme. *BMJ Open*. 2022;12(7):e059415.
18. **Plummer OR**, **Abboud JA**, **Bell J-E**, **et al**. A concise shoulder outcome measure: application of computerized adaptive testing to the American Shoulder and Elbow Surgeons Shoulder Assessment. *J Shoulder Elbow Surg*. 2019;28(7):1273–1280.
19. **Edelen MO**, **Rodriguez A**, **Herman P**, **Hays RD**. Crosswalking the Patient-Reported Outcomes Measurement Information System Physical Function, Pain Interference, and Pain Intensity scores to the Roland-Morris Disability Questionnaire and the Oswestry Disability Index. *Arch Phys Med Rehabil*. 2021;102(7):1317–1323.
20. **Harrison CJ**, **Plessen CY**, **Liegl G**, **et al**. Overcoming floor and ceiling effects in knee arthroplasty outcome measurement. *Bone Joint Res*. 2023;12(10):624–635.
21. **Ben-Shlomo Y**, **Blom A**, **Boulton C**, **et al**. The National Joint Registry 18th Annual Report 2021. London: National Joint Registry, 2021. https://reports.njrcentre.org.uk/Portals/12/PDFdownloads/NJR%2018th%20Annual%20Report%202021.pdf (date last accessed 21 June 2024).
22. **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 Outcome Meas*. 2016;7:195–203.

Author information

A. Barakat, MBBCH, MSc, MRCS, Specialty Registrar in Trauma & Orthopaedics,
University Hospitals of Leicester NHS Trust, Leicester, UK.

J. Evans, BMedSci, BMBS, MSc, MD, FRCS(Tr&Orth), RAF (Rtd), BMBS, MSc, MD, Clinical Senior Lecturer in Trauma & Orthopaedics, Honorary Consultant in Trauma & Orthopaedics, University of Exeter, Exeter, UK; Exeter Hip Unit, Royal Devon University Healthcare NHS Foundation Trust, Exeter, UK.

C. Gibbons, MBBS, PhD, Associate Professor, Division of Internal Medicine, Department of Symptom Research, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA.

H. P. Singh, MBBS, MSc, PhD, FRCS (Tr&Orth), Honorary Consultant in Trauma & Orthopaedics, Consultant in Trauma & Orthopaedics, Exeter Hip Unit, Royal Devon University Healthcare NHS Foundation Trust, Exeter, UK; University Hospitals of Leicester NHS Trust, Leicester, UK.

Author contributions

A. Barakat: Formal analysis, Investigation, Writing – original draft.
J. Evans: Conceptualization, Funding acquisition, Software, Supervision, Writing – review & editing.
C. Gibbons: Methodology, Project administration, Supervision, Writing – review & editing.
H. P. Singh: Conceptualization, Methodology, Project administration, Visualization, Writing – review & editing.

Funding statement

The authors disclose receipt of the following financial or material support for the research, authorship, and/or publication of this article: funding for the article processing charges from the National Institute for Health and Care Research (NIHR) Exeter Biomedical Research Centre (BRC), as reported by J. Evans.

ICMJE COI statement

J. Evans reports funding for the article processing charges from the National Institute for Health and Care Research (NIHR) Exeter Biomedical Research Centre (BRC).

Data sharing

The data that support the findings for this study are available to other researchers from the corresponding author upon reasonable request.

Acknowledgements

We would like to acknowledge and thank the National Joint Registry for England, Wales, Northern Ireland, the Isle of Man, and the States of Guernsey for approving provision of the anonymized data repository for the analysis.

Ethical review statement

The study was approved by the National Joint Registry for England, Wales, Northern Ireland, the Isle of Man, and the States of Guernsey for access, analysis, and publication of anonymized depersonalized data repository.

Open access funding

The authors report that they received open access funding for their manuscript from the National Institute for Health and Care Research (NIHR) Exeter Biomedical Research Centre (BRC).