BMJ Open

Computerised Adaptive Testing Dramatically Reduces the Number of Items in Patient Reported Hip and Knee Outcome Scores An analysis of the UK National PROMs programme

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2021-059415
Article Type:	Original research
Date Submitted by the Author:	28-Nov-2021
Complete List of Authors:	Evans, Jonathan; University of Exeter Medical School, Health Services and Policy Research Group ; Royal Devon and Exeter Hospital Gibbons , Christopher; The University of Texas MD Anderson Cancer Center, Center for INSPiRED Cancer Care (Integrated Systems for Patient-Reported Data) Toms, Andrew ; Royal Devon and Exeter NHS Foundation Trust Valderas, Jose; University of Exeter Medical School, Health Services and Policy research Group
Keywords:	Quality in health care < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, Adult orthopaedics < ORTHOPAEDIC & TRAUMA SURGERY, Hip < ORTHOPAEDIC & TRAUMA SURGERY, Knee < ORTHOPAEDIC & TRAUMA SURGERY

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

Computerised Adaptive Testing Dramatically Reduces the Number of Items in Patient Reported Hip and Knee Outcome Scores

An analysis of the UK National PROMs programme

Jonathan P Evans[1,2], Christopher Gibbons[3], Professor Andrew Toms[2], Professor Jose Valderas[1]

1. Health Services and Policy Research, Exeter Collaboration for Academic Primary Care (APEx), University of Exeter, Magdalen Campus, Smeall Building, Room JS02, Exeter, EX1 2LU, UK
2. Princess Elizabeth Orthopaedic Centre, Royal Devon and Exeter Hospital, Exeter, EX2 5DW, UK
3. Division of Internal Medicine, Department of Symptom Research, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

Corresponding Author

Jonathan P Evans

Address: Health Services and Policy Research, Exeter Collaboration for Academic Primary Care (APEx), University of Exeter, Magdalen Campus, Smeall Building, Room JS02, Exeter, EX1 2LU, UK

Email: j.p.evans2@exeter.ac.uk

Contributions

JP Evans MBChB, MSc, MD(res), FRCS (Tr and Orth) is an NIHR Academic Clinical Lecturer in health services research and is a senior fellow in trauma and orthopaedics

C Gibbons PhD is Deputy Chair and Associate Professor in symptom research and patient reported outcome measures

A Toms MB ChB FRCS (Ed) MSc Eng FRCS (Tr & Orth) is a Consultant Trauma and Orthopaedic Surgeon and Honorary Clinical Professor

JM Valderas MPH, PhD is Professor of health services and policy research

Computerised Adaptive Testing Dramatically Reduces the Number of Items in Patient Reported Hip and Knee Outcome Scores

An analysis of the NHS England National PROMs programme

Abstract

Objective

Over 160,000 participants per year complete the 12-item Oxford Hip and Knee Scores (OHS/OKS) as part of the NHS England Patient Reported Outcome Measures (PROMs) programme. We used a modern computational approach, known as computerized adaptive testing (CAT), to simulate individually-tailored OHS and OKS assessment, with the goal of reducing the number of questions a patient must complete without compromising measurement accuracy.

Methods

We calibrated the 2018/2019 PROMs data to an item response theory (IRT) model. We assessed IRT model assumptions alongside reliability. We used parameters from the IRT model with data from 2017/2018 to simulate CAT assessments. Two simulations were run until a prespecified standard error of measurement was met (SE = .32 and SE = .45). We compared the number of questions required to meet each cut-off and assessed the correlation between the full-length and CAT administration.

Results

We conducted IRT analysis using 40,432 OHS and 44,714 OKS observations. The OHS and OKS were both unidimensional (Root Mean Square Error of Approximation (RMSEA) .08 and .07 respectively) and marginal reliability .91 and .90. The CAT, with a precision limit of SE = .32 and SE = .45 required a median of 4 items (IQR 1) and 2 items (IQR 1) respectively for the OHS, and median of 4 items (IQR 2) and 2 items (IQR 0) for the OKS. This represents a potential 82% reduction in PROM length. In the context of 160,000 yearly assessments, these methodologies could result in the omission of some 1,280,000 redundant questions per year which equates to 40,000 hours of patient time.

Conclusion

The application of IRT to the OHS and OKS produces an efficient and substantially reduced CAT. We have demonstrated a path to reduce the burden and potentially increase the compliance for these ubiquitous outcome measures without compromising measurement accuracy.

Strengths and limitations of this study

- **Our study is the first application of computer adaptive testing on the worlds largest repository of patient reported outcome measures.**
- **Over 35,000 responses were used in each modelling and simulation group**
- **The Oxford Hip and Knee scores are very widely used at an international level**
- **This secondary database analysis requires validation on a prospectively collected cohort**
- **The available datasets are limited due to attrition that is attributed to the linking of PROMs to health records**

Introduction

The ability to assess a patient's perspective about their health is central to holistic clinical decision making, medical research, and health policy construction.[1] For hip and knee replacement surgery, patients often complete questionnaires called patient reported outcome measures (PROMs) before and after their operation. Since 2009, over 160,000 patients per year undergoing a hip or a knee replacement complete PROMs as part of the NHS England's PROMs Programme.[2]

The PROMs used as part of this programme include the Oxford Hip Score (OHS) and Oxford Knee Score (OKS) which are filled in using pen and paper. Outside of the UK, they are also collected routinely as part of arthroplasty registries in Australia, New Zealand, Canada and the Netherlands.[3] The completion rates across England for the 2018/2019 pre-operative OHS and OKS were 85.7% and 86.1% respectively,[4] however at the hospital trust level, the completion rate varies from 30% to 100%.[5] Attrition is evident when obtaining completed post-operative PROMs (70% completion), further reduction in the data is caused by the process of transcribing the scores to a digital platform and linking with health records which reduces the number of usable records to below 50%.[4] It has been recognised that PROM questionnaires collected using paper and pen for the England PROMs programme are resource-intensive, inefficient for providers and burdensome for patients.[6] The time required to complete orthopaedic PROMs is seen as a key barrier by patients, and the risk of non-

completion is highest in those from the most deprived quintile of socioeconomic status and those with poorer general health.[5,7]

Patient-reported outcome measures are composed of a series of questions (items) that ask patients about aspects of their health. These are scored with a structured format to give an estimate of a continuous construct known as a latent trait (i.e. a variable that is not directly observable).[8] Latent traits in orthopaedics typically include pain and physical function. The OHS and OKS were developed using a methodological process called Classical Test Theory (CTT), whereby fixed-length questionnaires were given an overall score, without weighting or standardisation, which estimates the latent trait. More recently, PROMs developed using advanced psychometric techniques have emerged. Influenced by state-of-the-science psychological tests, modelling approaches including Rasch analysis and Item Response Theory (IRT) focus on the individual item within the scale, in contrast to CTT methods, which focus on the total score of all the items together.[9] The ability to calibrate each item individually dramatically increases the versatility of the resulting PROM. Within the IRT paradigm, valid measurement can be obtained using any number of questions from the scale whereas under CTT each item must be administered for the score to be deemed valid. Another limitation of CTT is that it can only identify items that are not related to the construct being measured, it does not identify items which are redundant (e.g., too similar to others) and can incentivise the inclusion of redundant items.[10] This flexibility is leveraged by a computational technique known as computerised adaptive testing (CAT).[11] A CAT method iteratively select the most informative and relevant items for a particular individual, thereby individualising the assessment to the patient, typically resulting in reducing assessment length without sacrificing accuracy. Importantly, IRT analyses can be retrospectively applied to legacy PROMs that were initially designed using CTT. Rasch analysis of the OHS and OKS has previously been undertaken, with all studies demonstrating improvements in precision and group discrimination.[12–16] The development of OHS/OKS CAT could improve the efficiency of administration and reduce the administrative burden

of the PROMs programme while offering the opportunity to implement a CAT at an unprecedented scale.

The purpose of this study is to assess whether the application of IRT and CAT to the OHS and OKS confers a reduction in questionnaire burden whilst maintaining precision.

Methods

Data

The OHS was developed in 1996, and the OKS in 1998 [17,18] Each PROM contains 12 items that assess joint-specific symptoms over the last four weeks. Each item has five response options that grade the severity of symptoms and functional limitations. Developed following interviews with joint replacement patients they were found to be the best performing condition-specific instruments available in a standardised comparison of the measurement properties.[19]

All Individual item level pre-operative OHS and OKS scores were extracted from the 2018/2019 data release for hip and knee replacements for IRT model development. A second sample for simulation of the CAT was extracted from the 2017/2018 data release.[4] Raw PROMs data are released annually on the NHS digital platform following pre and postoperative linking, health record linking and validation and data cleaning.[4]

Development of the IRT model

We assessed the number of missing responses at the item level and presented them as a percentage difference. We assessed the IRT assumptions of unidimensionality, local independence and monotonicity.[20] To confirm that all items measure a single underlying construct, we assessed unidimensionality using confirmatory factor analysis (CFA). Model fit for CFA was assessed through root mean square error of approximation (RMSEA) with a borderline model fit set at ≤0.08 and good fit ≤0.06, and comparative fit index (CFI) and Tucker-Lewis index (TLI) with borderline model fit set to >0.90 and good fit >0.95 (R package `LAVAAN` version 0.5–23.1097).[21] We confirmed the dimensional structure of each scale using Mokken scaling and assessed scalability (monotonicity) of the items. This assesses whether the probability of scoring the item along its scale of symptom severity increases with a higher level of the underlying construct. A Loevinger's H value of ≥0.3 per item was deemed acceptable (R package `Mokken` version 2.8.4).[22]

Assessment of local independence of items was undertaken to ensure that all items only relate to the dominant construct being measured, and not to a further independent construct. This was assessed through a correlation between items residuals revealing significant covariance that may indicate that the items are too similar and therefore redundant. 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 above 0.2 demonstrating locally dependent items.[23]

Following confirmation of IRT model fit assumptions a Graded Response Model (GRM), which is appropriate when item responses can be categorised as ordered categorical[24], was fit to the item response data (R package `mirt` version 3.3.2). This model yields two-item parameters, the item difficulty (a) which is a representation of the level of information about the underlying construct each item provides, and the discrimination (b) thresholds which locate the response categories and their transitions along a contiguous scale. If the item-characteristic curves revealed disordered thresholds, where the response category does not accord with the latent trait score, reordering of adjacent response options was undertaken. Item and model fit was assessed using the RMSEA. Reliability in the IRT model was estimated as marginal reliability where the overall reliability of the test was based on the average conditional standard errors.[25] This overall index of precision can be compared to the classical internal consistency (Cronbach's alpha) reliability estimate for CTT, where scores >0.8 indicate excellent reliability.

Computer Adaptive Testing Simulation

The production of item thresholds and difficulty information from the IRT models allows the construction of a CAT. The administration of a CAT utilises algorithms, which match participants to the most informative items within a PROM and once an acceptable level of precision is reached, denoted by the reliability (Standard Error (SE)), of the latent trait estimate, no further items are required.[26] Within a CAT simulation, the estimate of the latent trait from the full-length PROM can be compared to the delivery of shortened versions where particular items are selected. This

simulation can provide information on the number of items needed to provide estimates of the latent trait at predetermined levels of precision. Through the simulation, the items that provide the highest level of information, and thereby the greatest utility in shortened versions, can be determined.

We performed a CAT simulation using Firestar for R (version 1.3.2).[27] Two separate simulations were conducted for OHS and OKS with the 2017/18 dataset with predetermined stopping criteria (precision) denoted as a SE of the latent trait estimate of <0.32 and <0.45. These SE values are equivalent to a reliability coefficient of 0.90 and 0.80 respectively. Variables derived from the simulation include the correlation (Intraclass Correlation Coefficient, (ICC)) between the latent trait estimation of the full-length questionnaire and the CAT, and the mean and standard deviation, median and interquartile range (IQR) 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 is presented as a percentage difference. Time-saving between full-length and CAT administration were calculated against the estimate that each item takes between 10 seconds and 75 seconds per item to complete.[28]

All data analysis was conducted in R (RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA).

Public and Patient Involvement Statement

Formal patient and public involvement was not undertaken for this analysis of public domain data. The national PROMs programme and the data held within have themselves been evaluated via public consultation. Response to this evaluation from a multiple stakeholders taskforce highlighted the need to improve efficiency of data collection.[6]

Data sharing

Difficulty and discrimination parameters of the IRT model for both the OHS and OKS are available in Appendix 1. All data are available from NHS digital and can be used in accordance with the open government licence for public sector information.

Results

Dataset characteristics

Of the 40,172 preoperative OHS scores and 44,264 OKS scores in the 2018/19 data, 1,704 were revision hip replacements and 1,162 revision knee replacements, which were excluded. Further exclusion of incomplete questionnaires resulted in 37,995 OHS and 42,558 OKS observations. Missing responses to items were found 4118 (0.90%) and 4803 (0.93%) times for the OHS and OKS. Although the first two items had few missing responses (0.15% OHS, 0.10% OKS), the remaining ten had substantially more (1.04% OHS, 1.1% OKS), indicating a possible patient preference for shorter measures.

For hip replacements, 59.4% were undertaken in females, 93.8% were ≥50 years, and 51.9% were ≥70 years old. 14.8% of respondents had assistance completing the questionnaires, median symptom duration was one to five years. For knee replacements, 56.5% were undertaken in females, 97.4% were ≥50 years and 51.8% were ≥70 years old. 14.6% had assistance completing their questionnaires, median symptom duration was one to five years. All demographic features of the PROMS dataset were equivalent to that of the full National Joint Registry.

IRT model assumptions

The criterion of unidimensionality was met at a borderline level for both OHS and OKS with an RMSEA of 0.08 (OHS) and 0.07 (OKS), CFI of 0.93 (OHS) and 0.94 (OKS), TFI of 0.91 (OHS) and 0.93 (OKS). Mokken scaling corroborated this finding of unidimensionality and produced overall scalability coefficients (H) of 0.49 (range 0.41 – 0.58) (OHS) and 0.46 (0.38 – 0.55) (OKS). Local independence of items was confirmed for both OHS and OKS with all item correlations below 0.02. There were no misfitting items within the GRM model.

Following production of the IRT item characteristic curves disordered thresholds (where the curve lies under the line created by an adjacent curve) were noted in items 5, 6, 9, 10 and 12 for the OHS

and items 4, 6 and 8 for the OKS. Items with disordered thresholds were rescored, giving them the same score as the adjacent item whose area it lay within (Fig 1a & 1b). Item level RMSEA was good for both scores with all items RMSEA <0.02 (Appendix 1). The marginal reliability of the model was 0.91 for OHS and 0.90 for OKS. Overall model fit was adequate for the OHS (RMSEA 0.09) and good for OKS (RMSEA 0.06)

CAT simulation

We conducted a CAT simulation using the derived IRT parameters and utilising the preoperative OHS and OKS item responses from the 2017/2018 data release as the testing set. For the OHS 36,516 participants scores were included, and for the OKS 45,122. Incomplete records (i.e. less than 12 item scores) were included as the IRT method accounts for missing data, using all available responses to gain the best estimate of the latent trait. At the standard error threshold of 0.32 (corresponding to a reliability of 0.9) the ICC between full-length and CAT latent trait estimates was r = 0.96 (OHS) and r = 0.96 (OKS) (fig 2). For the OHS CAT, the mean number of items required was 3.98 (SD 1.26) with a median of 4 (IQR 1)). For the OKS CAT, the mean number of items required was 4.22 (SD 1.32) with a median of 4 (IQR 2) (fig 3).

With a precision SE threshold of 0.45 (corresponding to a reliability of 0.8), the concordance between full-length and CAT simulations decreased marginally to r = 0.90 (OHS) and r = 0.91 (OKS) (fig 2). The OHS CAT required a mean of 2.27 (SD 0.45) items, median of 2 (IQR 1). The OKS CAT required 2.13 (SD 0.45) items, median of 2 (IQR 0) (fig 3).

The items used most frequently within the 0.35 SE CAT were items 8 (24.9%) and 11 (21.3%) for OHS and items 9 (23.5%), 11 (23.4%) and 12 (16.5%) for the OKS, all other items were used in less than 16% of simulations. At 0.45 SE, OHS items 3 (19.9%), 8 (43.7%) and 11 (32.9.0%) were used most frequently in the simulations, four items were not used in any simulations and all other items were used less than 3% of the time. For the OKS items, 9 (46.1%) and 11 (45.7.0%) were used most

frequently, items 12 (5.1%) was minimally utilised leaving two items that were not required in any simulations (including amongst them items 6, and 8 whose response options had been identified as disordered) and all others were used in less than 1.4% of simulations (fig 4)(Appendix 1).

The items utilised most frequently in estimating the level of the latent trait within the OHS were item 8 (*During the past 4 weeks… After a meal (sat at a table), how painful has it been for you to stand up from a chair because of your hip?*) and within the OKS were item 9 (*During the past 4 weeks… How much has pain from your knee interfered with your usual work (including housework)?*) and item 11 (*During the past 4 weeks… Could you do the household shopping on your own?*).

Out of a potential 438,192 items for the OHS scores, only 145,462 items were used by the CAT at 0.32 SE, and 82,980 at 0.45 SE. This represents a 100.3% and 136.3% difference. Taking the whole 2018/19 NJR dataset before exclusions of 95,977 total hip replacements, at 0.45 SE this represents a potential time saving of 2583 – 19374 hours for pre-operative scores. Out of a potential 541,464 items for the OKS scores, only 190,410 items were used by the CAT at 0.32 SE, and 96,922 at 0.45 SE, representing a 100.3% and 136.3% difference For the entire NJR dataset in 2018/19, at 0.45 SE this represents a potential saving of between 2832.2 – 21241.5 hours for collection of pre-operative scores.

Discussion

The use of PROMs in the outcome assessment of hip and knee replacements is widely accepted. The best PROMs for patients, researchers and clinicians are easy to understand, free from redundancy, and psychometrically robust. In this study, we have applied a modern psychometric approach to the world's largest repository of orthopaedic PROMs. The OHS and OKS conformed to IRT assumptions by demonstrating unidimensionality, monotonicity, and local independence. Computerised adaptive testing simulations demonstrated the possibility to dramatically reduce the length of these 12-item PROMs to as little as two items without compromising precision. In large-scale data collection, the potential time saving from the deployment of a CAT is equivalent to more than a million redundant questionnaire items per year and more than 4 years of collective patient time annually.

Although Rasch analysis, a type of IRT methodology, has previously been applied to both the OHS and OKS,[13,15] no research to date has explored the possibility of using these methods to reduce the burden of assessment using CAT. Computerised adaptive testing achieved the goal of minimising the burden of a PROM by only delivering the most relevant and informative items required to measure a patient's level of hip or knee pain and function.[29] The simulations performed within this study were able to reduce the number of items required by 67% for the OHS and 65% for the OKS at 0.32 SE (equivalent to 90% precision) and as much as 81% for both PROMs at 0.45 SE (80% precision). As a comparator, the reliability, (taken as a proxy marker of precision) of the 12-item OHS and OKS delivered, (using the classical test theory derived scoring system) has a test-retest Intraclass Correlation Coefficient (ICC) of 0.82 – 0.94.[30,31] Although this is excellent, delivery of the full test does not demonstrate superiority over a CAT administration. A minimal reliability threshold of 0.70 is commonly accepted for PROMs, such as those used in the NHS England PROMS programme. The standard error at this reliability level is 0.55 of a standard deviation, which is roughly equivalent to a reliability of .70.[32] Similarly, a SE of 0.45, (equivalent to .80 reliability), in this simulation, a median of only two items were required to estimate patients pain and function dramatically reducing the burden on a patient. Interestingly, although the overall completion rate of the PROMs was high,

within this sample, the non-completion of items substantially increased after the first two items, and then remained stable for the remaining 10 items. Minimising respondent fatigue by simply asking two items is likely to improve completion rates both at the start of data collection and longitudinally, thereby optimising the utility of this valuable data.[33] The value and reliability of PROMs is vastly improved by regular administration over time, the ability to conduct this with targeted highly condensed PROMs that retain their ability to precisely estimate the latent trait is only possible through IRT analysis and CAT administration.[26] Furthermore, the two items used most frequently in the CAT deployment for OHS (Items 8 and 11), have been judged by patients as having the most clarity and fewest limitations.[34] Of interest, within both questionnaires, the pre-operative items pertaining to function rather than pain were selected by the simulation as most valuable.

The use of IRT-derived PROMs is becoming increasingly prevalent in efforts to advance high-value care and improve shared decision-making.[35] The ability to score on a simple continuum (eg 0-100) and derive population norms (eg a score of 50), vastly improves patient comprehension of their score. Patients understanding of the relevance of their PROM score improves their compliance with future assessment and optimises the use of a PROM as a decision aid.[36] The use of this latent trait continuum that is independent of the PROM also allows comparison of the OHS and OKS scores with other scores assessing the same trait. Therefore, so-called "cross-walks" can be derived to compare the scores derived from the OHS and OKS with other hip scores such as the Hip disability and Knee injury and Osteoarthritis Outcome Scores (HOOS and KOOS) that have also undergone IRT analysis,[37] or contemporarily designed PROMs such as the PROMIS physical function and pain interference scores. This attribute can have a profound effect on the translatability of research findings.

The authors recognise limitations inherent to this study. We recognise that the dimensionality of both the OHS and OKS could be contested on the basis of the borderline results. It has previously been identified that both one-factor and two-factor models fit these scores.[38,39] As the most commonly applied scoring method utilises the total score for this very common PROM, it was

deemed appropriate to maintain a unidimensional model. The strength of this analysis is the very large sample size for the IRT model construction. The CAT simulation requires validation on patients with both qualitative and quantitative analysis of validity and acceptability. The significant limitation to the practical application of IRT and CAT is the availability of a computer and an appropriate interface. However, the utilisation of the PROMIS system in the USA highlights that these barriers can be overcome, furthermore, the increasing ubiquity of tablet and smartphone interfaces and the often-underappreciated technological literacy of this patient population suggest that this problem is far from insurmountable. Both during and in the post-pandemic era, remote medicine is becoming the norm; refined PROMs collection has a vital role to play in this process.

Conclusion

The collection of hip and knee outcome measures for the NHS England National PROMs programme has been criticised as remote from patient care. By applying modern psychometric analysis to the world's largest repository of hip and knee patients PROMs, we have demonstrated up to an 80% reduction in the number of items required to estimate the patient-specific impact of joint disease without compromising precision. Widespread adoption of this system has the potential to reduce participant burden and increase completion rates, thereby maximising the reliability and utility of longitudinal data.

Contribution

Authors JPE, CG and JMV devised the project. JPE managed the data and analysed the results with oversight from CG and JMV. AT contributed to the data management and interpretation. JPE wrote the manuscript. CG, AT and JMV edited the manuscript. All authors read and approved the final manuscript.

Ethics approval

Formal ethics approval was not required. All data used in accordance with the open government licence for public sector information. The National Archives. Open Government Licence for public sector information. 2020. http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/ (accessed Dec 21, 2020).

Transparency statement

The lead author (Jonathan Evans) affirms that the manuscript is an honest, accurate, and transparent account and that no important aspects of the study have been omitted.

Role of the funding source

Jonathan Evans is in receipt of an NIHR Academic Clinical Lecturer award. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. There was no involvement of the funder in study design, data collection, data analysis, manuscript preparation or publication decisions. All authors had complete access to the study data that support the publication.

Conflict of Interest

Jonathan Evans, Christopher Gibbons, Andrew Toms, and Jose Valderas declare that they have no conflict of interest

Figure Legends

Figure 1: Item response theory (IRT) item traces for the 12- items of the Oxford Hip Score (OHS)(a) and Oxford Knee Score (OKS)(b)

Figure 2: Scatter plot and correlation between the theta estimation (values of the latent trait) between the full 12-item administration and the Computerised Adaptive Test (CAT) for the Oxford His Score (OHS) (a & b) and Oxford Knee Score (OKS) (c & d) at 0.32 Standard Error (SE) and 0.45 SE.

Figure 3: Bar chart showing the number of items used per participant at 0.32 Standard Error (SE) and 0.45 SE for the OHS (a, b) and OKS (c, d) Computerised Adaptive Test (CAT).

Figure 4: Bar chart showing the proportional use of each item at 0.35 Standard Error (SE) and 0.45 SE for the OHS (a, b) and OKS (c, d) Computerised Adaptive Test (CAT).

References

1 Black N, Burke L, Forrest CB, *et al.* Patient-reported outcomes: pathways to better health, better services, and better societies. *Qual Life Res* 2016; **25**: 1103–12.

2 Health & Social Care Information Centre. National PROMs programme. .

3 Wilson I, Bohm E, Lübbeke A, *et al.* Orthopaedic registries with patient-reported outcome measures. *EFORT Open Rev* 2019; **4**: 357–67.

4 NHS Digital. Finalised PROMs data release. Patient Reported Outcome Measures (PROMs) in England for Hip and Knee Replacement Procedures (April 2018 to March 2019). 2020. https://digital.nhs.uk/data-and-information/publications/statistical/patient-reported-outcome-measures-proms/hip-and-knee-replacement-procedures-april-2019-to-march-2020 (accessed Dec 21, 2020).

5 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.

6 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**: 441–5.

7 Rowland C, Walsh L, Harrop R, Roy B, Skevington SM. What Do U.K. Orthopedic Surgery Patients Think About PROMs? Evaluating the Evaluation and Explaining Missing Data. *Qual Health Res* 2019; **29**: 2057–69.

8 Gorter R, Fox J-P, Twisk JWR. Why item response theory should be used for longitudinal questionnaire data analysis in medical research Data analysis, statistics and modelling. *BMC Med Res Methodol* 2015; **15**. DOI:10.1186/s12874-015-0050-x.

9 Cappelleri JC, Lundy JJ, Hays RD. Overview of classical test theory and item response theory

for the quantitative assessment of items in developing patient-reported outcomes measures. *Clin Ther* 2014; **36**: 648–62.

10 Sijtsma K. On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika* 2009; **74**: 107.

11 Cella D, Gershon R, Lai J-S, Choi S. The future of outcomes measurement: Item banking, tailored short-forms, and computerized adaptive assessment. *Qual Life Res* 2007; **16**: 133–41.

12 Ko Y, Lo NN, Yeo SJ, *et al.* Comparison of the responsiveness of the SF-36, the Oxford Knee Score, and the Knee Society Clinical Rating System in patients undergoing total knee replacement. *Qual Life Res* 2013; **22**: 2455–9.

13 Ko Y, Lo NN, Yeo SJ, *et al.* Rasch analysis of the Oxford Knee Score. *Osteoarthr Cartil* 2009; **17**: 1163–9.

14 Norquist JM, Fitzpatrick R, Dawson J, Jenkinson C. Comparing alternative Rasch-based methods vs raw scores in measuring change in health. *Med Care* 2004; **42**. DOI:10.1097/01.mlr.0000103530.13056.88.

15 Fitzpatrick R, Norquist JM, Jenkinson C, *et al.* A comparison of Rasch with Likert scoring to discriminate between patients' evaluations of total hip replacement surgery. *Qual Life Res* 2004; **13**: 331–8.

16 Fitzpatrick R, Norquist JM, Jenkinson C, *et al.* A comparison of Rasch with Likert scoring to discriminate between patients' evaluations of total hip replacement surgery. *Qual Life Res* 2004; **13**: 331–8.

17 Dawson J, Fitzpatrick R, Carr A, Murray D. Questionnaire on the perceptions of patients about total hip replacement. *J Bone Joint Surg Br* 1996; **78**: 185–90.

18 Dawson J, Fitzpatrick R, Murray D, Carr A. Questionnaire on the perceptions of patients about

total knee replacement. *J Bone Joint Surg Br* 1998; **80**: 63–9.

19 Harris K, Dawson J, Gibbons E, *et al.* Systematic review of measurement properties of patient-reported outcome measures used in patients undergoing hip and knee arthroplasty. *Patient Relat Outcome Meas* 2016; **Volume 7**: 101–8.

20 Cella D, Yount S, Rothrock N, *et al.* The Patient-Reported Outcomes Measurement Information System (PROMIS): Progress of an NIH roadmap cooperative group during its first two years. *Med Care* 2007; **45**: S3–11.

21 Rosseel Y. Lavaan: An R package for structural equation modeling and more. Version 0.5–12 (BETA). *J Stat Softw* 2012; **48**: 1–36.

22 Van der Ark LA. Mokken scale analysis in R. *J Stat Softw* 2007; **20**: 1–19.

23 Yen WM. Scaling Performance Assessments: Strategies for Managing Local Item Dependence. *J Educ Meas* 1993; **30**: 187–213.

24 Hays RD, Morales LS, Reise SP. Item Response Theory and Health Outcomes Measurement in the 21st Century NIH Public Access. .

25 Green BF, Bock RD, Humphreys LG, Linn RL, Reckase MD. Technical guidelines for assessing computerized adaptive tests. *J Educ Meas* 1984; **21**: 347–60.

26 Gibbons CJ. Turning the page on pen-and-paper questionnaires: combining ecological momentary assessment and computer adaptive testing to transform psychological assessment in the 21st Century. *Front Psychol* 2017; **7**: 1933.

27 Choi SW. Firestar: Computerized adaptive testing simulation program for polytomous item response theory models. *Appl Psychol Meas* 2009; **33**: 644.

28 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. *BMJ Qual Saf* 1999; **8**: 228–33.

29 Cook KF, O'Malley KJ, Roddey TS. Dynamic assessment of health outcomes: time to let the CAT out of the bag? *Health Serv Res* 2005; **40**: 1694–711.

30 Gagnier JJ, Huang H, Mullins M, *et al.* Measurement properties of patient-reported outcome measures used in patients undergoing total hip arthroplasty: A systematic review. *JBJS Rev* 2018; **6**. DOI:10.2106/JBJS.RVW.17.00038.

31 Gagnier JJ, Mullins M, Huang H, *et al.* A Systematic Review of Measurement Properties of Patient-Reported Outcome Measures Used in Patients Undergoing Total Knee Arthroplasty. J. Arthroplasty. 2017; **32**: 1688-1697.e7.

32 Reeve BB, Wyrwich KW, Wu AW, *et al.* ISOQOL recommends minimum standards for patient-reported outcome measures used in patient-centered outcome1. Reeve BB, Wyrwich KW, Wu AW, Velikova G, Terwee CB, Snyder CF, et al. ISOQOL recommends minimum standards for patient-reported outcome measures us. *Qual Life Res* 2013; **22**: 1889–905.

33 Krosnic J, Presser S. Question and Questionnaire Design. In "Handbook of Survey Research", 2nd edn. Elsevier, 2013.

34 Wylde V, Learmonth ID, Cavendish VJ. The Oxford hip score: the patient's perspective. *Health Qual Life Outcomes* 2005; **3**: 1–8.

35 Brodke DJ, Hung M, Bozic KJ. Item response theory and computerized adaptive testing for orthopaedic outcomes measures. *JAAOS-Journal Am Acad Orthop Surg* 2016; **24**: 750–4.

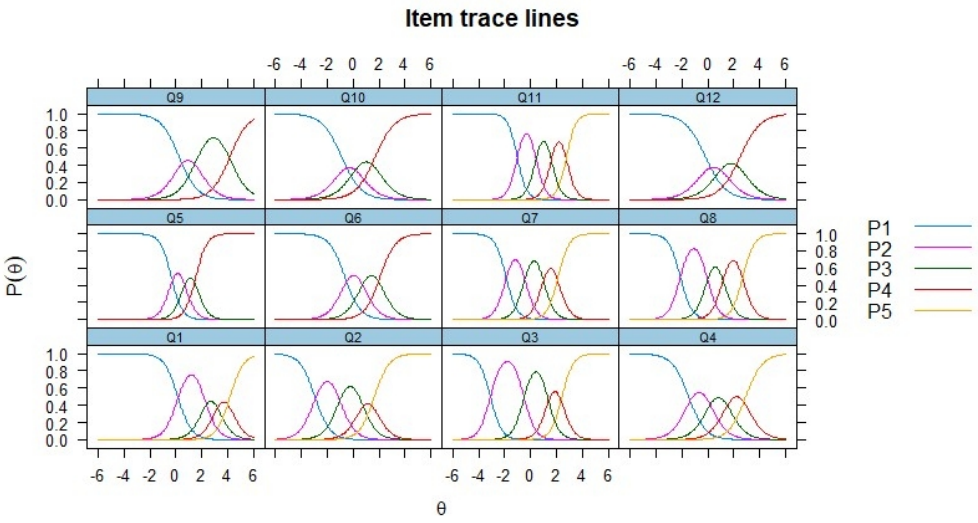
36 Porter I, Gonçalves-Bradley D, Ricci-Cabello I, *et al.* Framework and guidance for implementing patient-reported outcomes in clinical practice: evidence, challenges and opportunities. *J Comp Eff Res* 2016; **5**: 507–19.

37 Gandek B, Roos EM, Franklin PD, Ware JE. Item selection for 12-item short forms of the Knee

injury and Osteoarthritis Outcome Score (KOOS-12) and Hip disability and Osteoarthritis Outcome Score (HOOS-12). *Osteoarthr Cartil* 2019; **27**: 746–53.

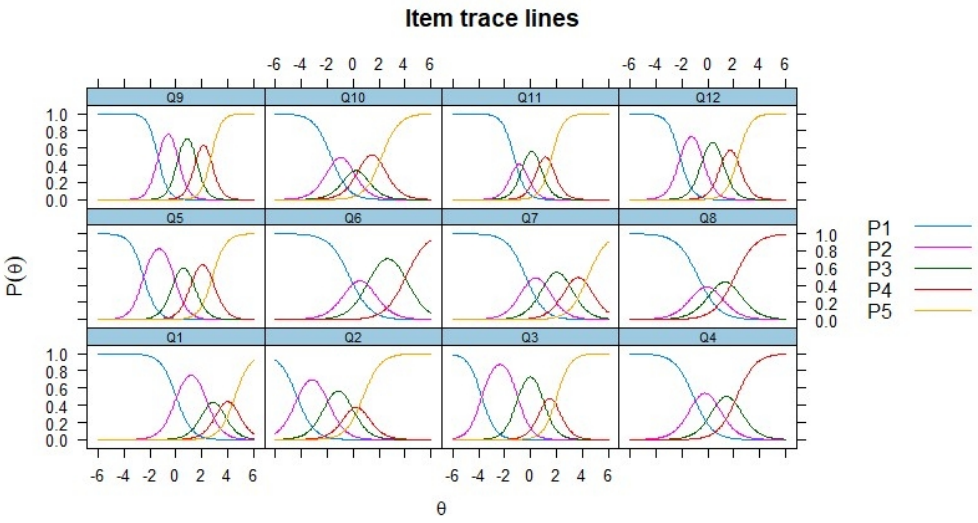
38 Harris KK, Price AJ, Beard DJ, Fitzpatrick R, Jenkinson C, Dawson J. Can pain and function be distinguished in the Oxford Hip Score in a meaningful way?: An exploratory and confirmatory factor analysis. *Bone Jt Res* 2014; **3**: 305–9.

39 Harris K, Dawson J, Doll H, *et al.* Can pain and function be distinguished in the Oxford Knee Score in a meaningful way? An exploratory and confirmatory factor analysis. *Qual Life Res* 2013; **22**: 2561–8.



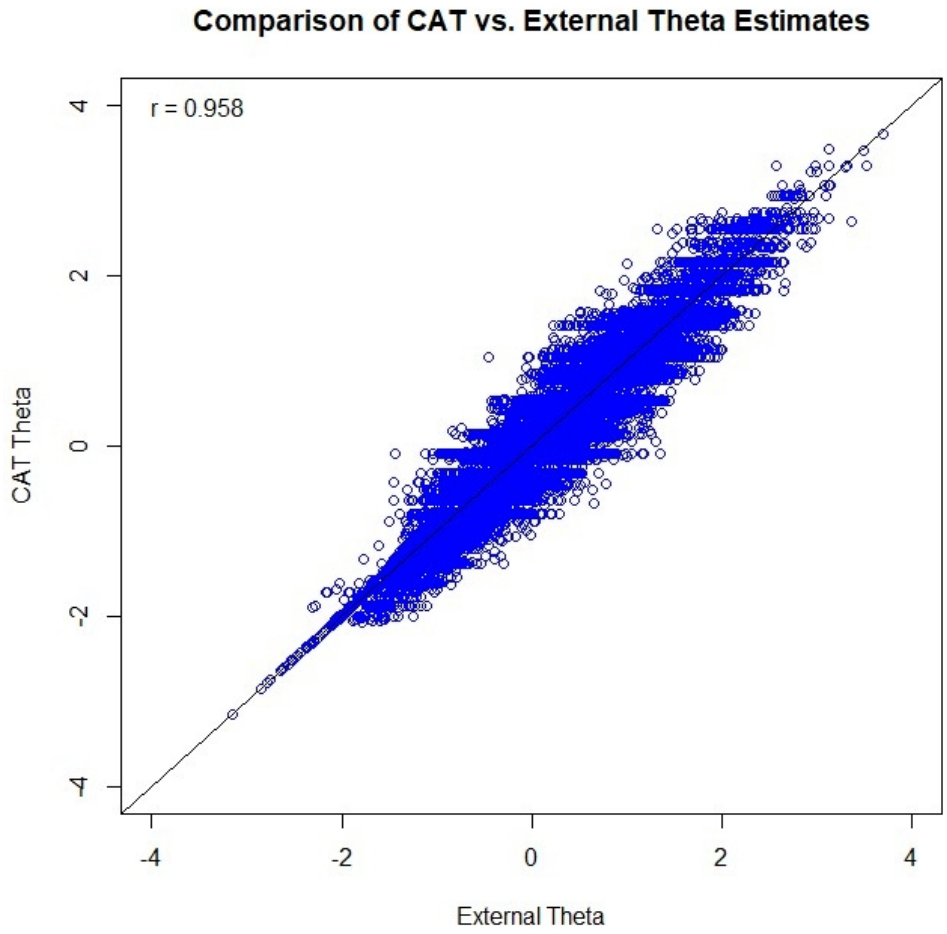
Item response theory (IRT) item traces for the 12- items of the Oxford Hip Score (OHS)(a) and Oxford Knee Score (OKS)(b)

212x114mm (96 x 96 DPI)



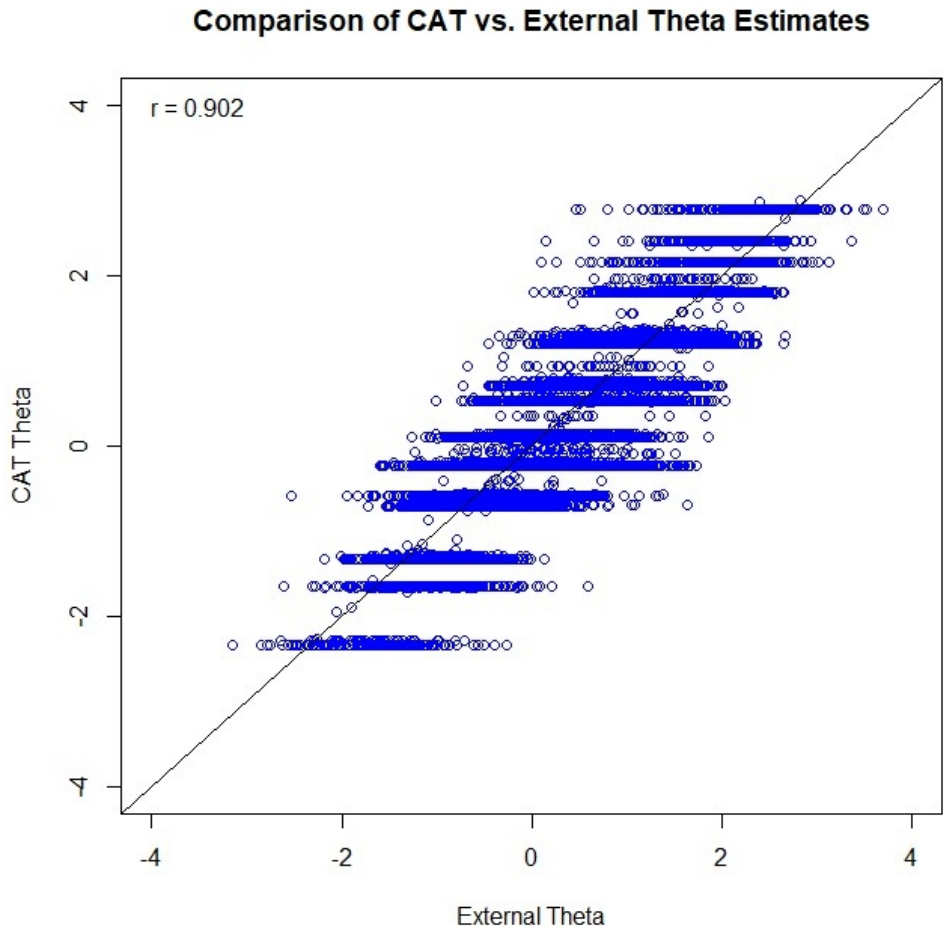
Item response theory (IRT) item traces for the 12- items of the Oxford Hip Score (OHS)(a) and Oxford Knee Score (OKS)(b)

212x114mm (96 x 96 DPI)



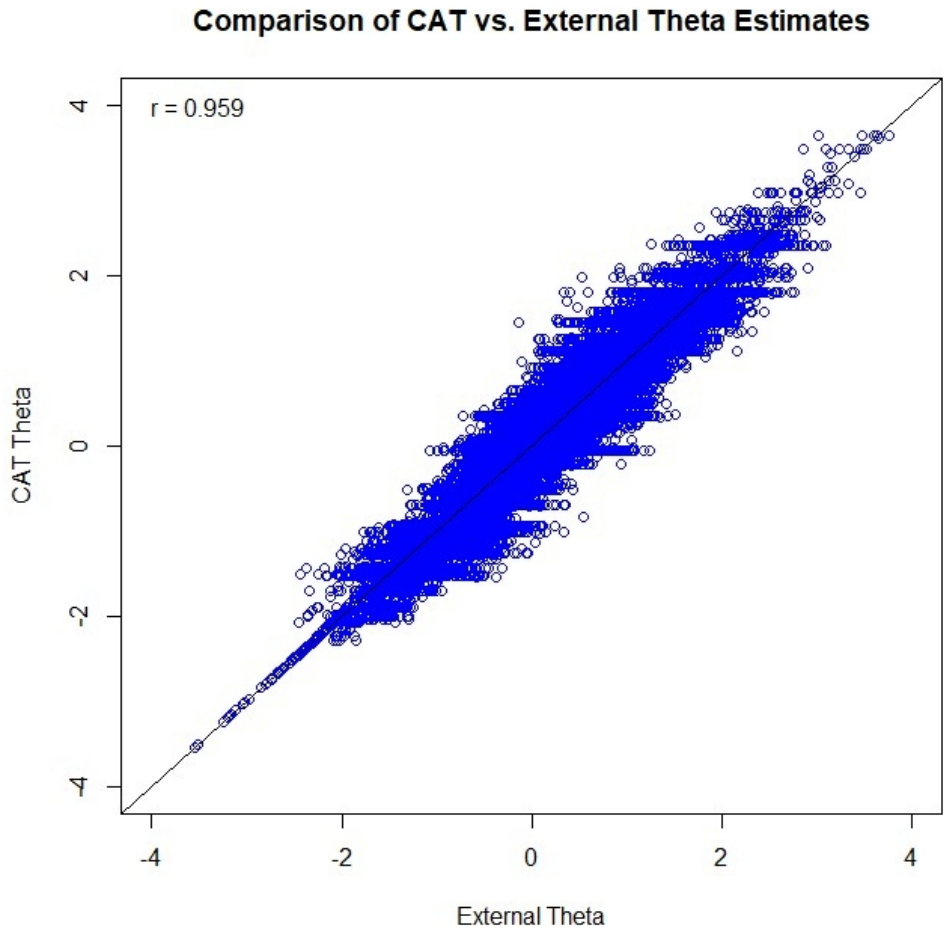
Scatter plot and correlation between the theta estimation (values of the latent trait) between the full 12-item administration and the Computerised Adaptive Test (CAT) for the Oxford His Score (OHS) (a & b) and Oxford Knee Score (OKS) (c & d) at 0.32 Standard Error (SE) and 0.45 SE

175x175mm (96 x 96 DPI)



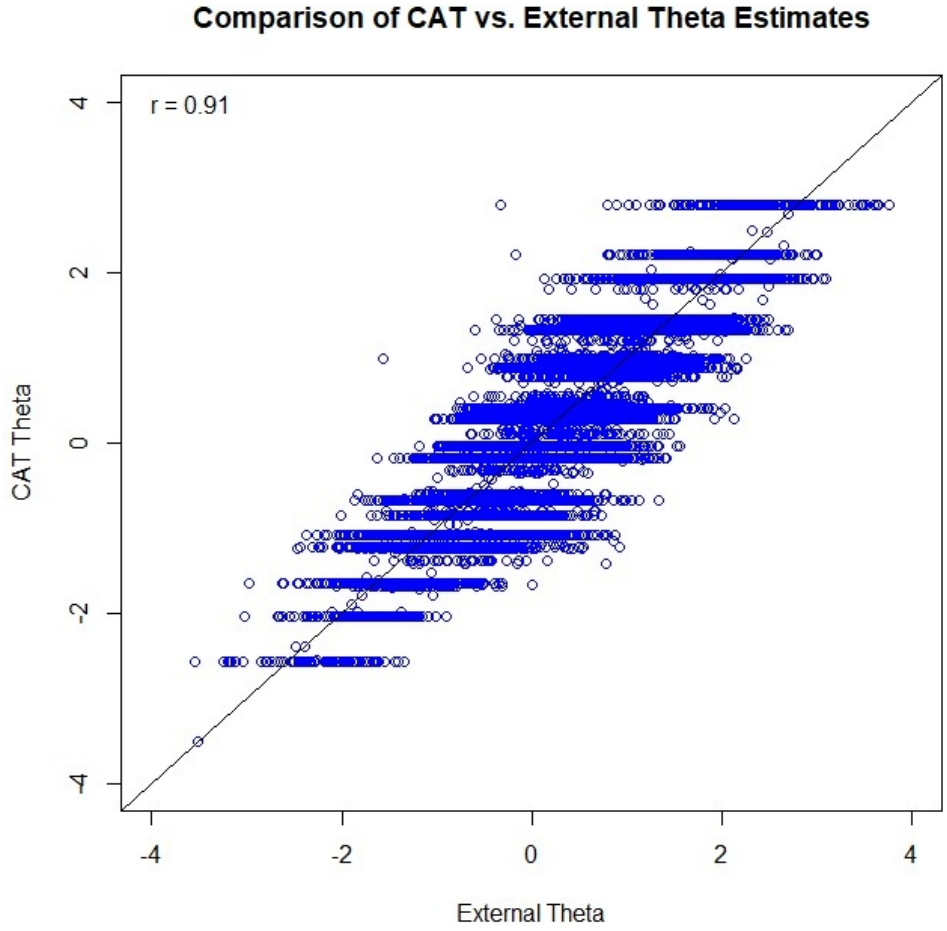
Scatter plot and correlation between the theta estimation (values of the latent trait) between the full 12-item administration and the Computerised Adaptive Test (CAT) for the Oxford His Score (OHS) (a & b) and Oxford Knee Score (OKS) (c & d) at 0.32 Standard Error (SE) and 0.45 SE

175x175mm (96 x 96 DPI)



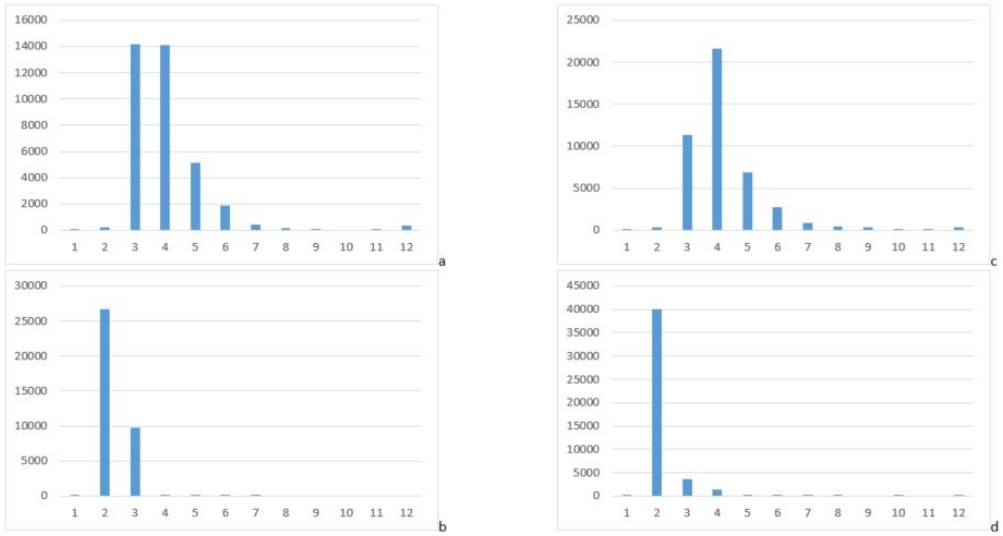
Scatter plot and correlation between the theta estimation (values of the latent trait) between the full 12-item administration and the Computerised Adaptive Test (CAT) for the Oxford His Score (OHS) (a & b) and Oxford Knee Score (OKS) (c & d) at 0.32 Standard Error (SE) and 0.45 SE

175x175mm (96 x 96 DPI)



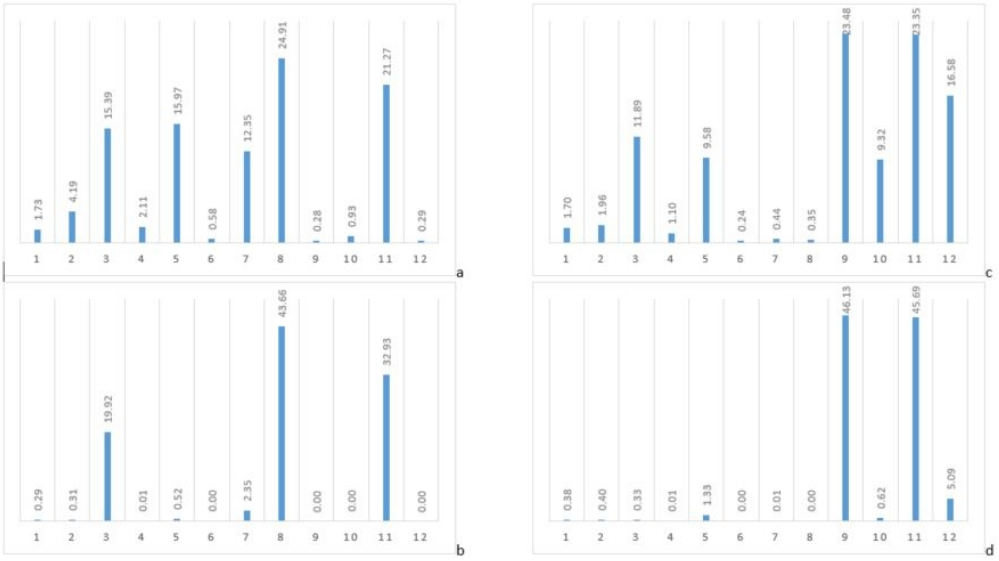
Scatter plot and correlation between the theta estimation (values of the latent trait) between the full 12-item administration and the Computerised Adaptive Test (CAT) for the Oxford His Score (OHS) (a & b) and Oxford Knee Score (OKS) (c & d) at 0.32 Standard Error (SE) and 0.45 SE

175x175mm (96 x 96 DPI)



Bar chart showing the number of items used per participant at 0.32 Standard Error (SE) and 0.45 SE for the OHS (a, b) and OKS (c, d) Computerised Adaptive Test (CAT).

242x129mm (96 x 96 DPI)



Bar chart showing the proportional use of each item at 0.35 Standard Error (SE) and 0.45 SE for the OHS (a, b) and OKS (c, d) Computerised Adaptive Test (CAT).

232x130mm (96 x 96 DPI)

Appendix 1

	Item	Difficulty a	Discrimination b1	Discrimination b2	Discrimination b3	Discrimination b4	Item RMSEA
1	During the past 4 weeks..... How would you describe the pain you usually had from your hip?	1.914691405	0.164325049	2.206224388	3.209345224	4.186925039	0.011
2	During the past 4 weeks..... Have you had any trouble with washing and drying yourself (all over) because of your hip?	1.740856269	-2.940926073	-1.052648162	0.609238754	1.637069089	0.003
3	During the past 4 weeks..... Have you had any trouble getting in and out of a car or using public transport because of your hip? (whichever you tend to use)	2.321961482	-3.093433431	-0.496468303	1.340835438	2.431036075	0.014
4	During the past 4 weeks..... Have you been able to put on a pair of socks, stockings or tights?	1.547750826	-1.475337619	0.125804202	1.50632625	2.924283288	0.009
5	During the past 4 weeks..... Could you do the household shopping on your own?	2.376360472	-0.348672227	0.659403283	1.549282667	NA	0.008
6	During the past 4 weeks..... For how long have you been able to walk before pain from your hip becomes severe? (with or without a stick)	1.668616909	-0.642568044	0.719789561	2.086871457	NA	0.015
7	During the past 4 weeks..... Have you been able to climb a flight of stairs?	2.360767142	-1.889430738	-0.433126661	0.981278344	2.144165736	0.007
8	During the past 4 weeks..... After a meal (sat at a table), how painful has it been for you to stand up from a chair because of your hip?	2.258019654	-2.12187152	-0.056100661	1.205103327	2.706143566	0.012
9	During the past 4 weeks..... Have you been limping when walking, because of your hip?	1.42746619	0.251525777	1.633812881	4.171461153	NA	0.007
10	During the past 4 weeks..... Have you had any sudden, severe pain - 'shooting', 'stabbing' or 'spasms' - from the affected hip?	1.324519295	-0.892084132	0.289697801	1.709666854	NA	0.011
11	During the past 4 weeks..... How much has pain from your hip interfered with your usual work (including housework)?	2.775690212	-1.028278183	0.415519561	1.607031821	2.776328526	0.006
12	During the past 4 weeks..... Have you been troubled by pain from your hip in bed at night?	1.260879482	-0.194492803	1.068727384	2.493862719	NA	0.011

Table 1 : Oxford Hip Score items with associated IRT derived difficulty and discrimination parameters.

Appendix 1

	Item	Difficulty a	Discrimination b1	Discrimination b2	Discrimination b3	Discrimination b4	Item RMSEA
1	During the past 4 weeks..... How would you describe the pain you usually have from your knee?	1.683615138	0.035557101	2.328208621	3.433951121	4.561327448	0.005
2	During the past 4 weeks..... Have you had any trouble with washing and drying yourself (all over) because of your knee?	1.492252738	-4.326258935	-2.018457992	-0.292460352	0.770869766	0.016
3	During the past 4 weeks..... Have you had any trouble getting in and out of a car or using public transport because of your knee? (whichever you would tend to use)	1.932656761	-3.747375182	-0.968344445	0.945305979	2.030550642	0.007
4	During the past 4 weeks..... For how long have you been able to walk before pain from	1.387915921	-1.095979526	0.648683	2.23788967	NA	0.010
5	During the past 4 weeks..... After a meal (sat at a table), how painful has it been for you to stand up from a chair because of your knee?	1.973493643	-2.473540467	-0.110575	1.302091221	2.835473629	0.003
6	During the past 4 weeks..... Have you been limping when walking, because of your knee?	1.263415959	-0.270638346	1.272615	4.061594039	NA	0.011
7	During the past 4 weeks..... Could you kneel down and get up again afterwards?	1.413075377	-0.361382764	1.150702	2.902993876	4.420398561	0.018
8	During the past 4 weeks..... Have you been troubled by pain from your knee in bed at night?	1.23865998	-0.757873886	0.547077	2.065465542	NA	0.005
9	During the past 4 weeks..... How much has pain from your knee interfered with your usual work (including housework)?	2.563072755	-1.375193192	0.177463	1.561747127	2.724240742	0.008
10	During the past 4 weeks..... Have you felt that your knee might suddenly 'give way' or let you down?	1.507070288	-1.693738699	-0.259655	0.672066751	2.202410973	0.008
11	During the past 4 weeks..... Could you do the household shopping on your own?	2.235209642	-1.264960065	-0.483151	0.662939289	1.641645648	0.019
12	2 During the past 4 weeks..... Could you walk down one flight of stairs?	2.135163585	-2.14417636	-0.398779	1.11031124	2.341481914	0.011

Table 2 : Oxford Knee Score items with associated IRT derived difficulty and discrimination parameters.

BMJ Open

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

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2021-059415.R1
Article Type:	Original research
Date Submitted by the Author:	01-Jun-2022
Complete List of Authors:	Evans, Jonathan; University of Exeter Medical School, Health Services and Policy Research Group ; Royal Devon and Exeter Hospital Gibbons , Christopher; The University of Texas MD Anderson Cancer Center, Center for INSPiRED Cancer Care (Integrated Systems for Patient-Reported Data) Toms, Andrew ; Royal Devon and Exeter NHS Foundation Trust Valderas, Jose; University of Exeter Medical School, Health Services and Policy research Group ; National University Singapore Yong Loo Lin School of Medicine, Department of Family Medicine
Primary Subject Heading:	Health services research
Secondary Subject Heading:	Surgery, Patient-centred medicine
Keywords:	Quality in health care < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, Adult orthopaedics < ORTHOPAEDIC & TRAUMA SURGERY, Hip < ORTHOPAEDIC & TRAUMA SURGERY, Knee < ORTHOPAEDIC & TRAUMA SURGERY

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

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

Jonathan P Evans[1,2], Christopher Gibbons[3], Professor Andrew Toms[2], Professor Jose Valderas[1,4]

1. Health Services and Policy Research, Exeter Collaboration for Academic Primary Care (APEx), University of Exeter, Magdalen Campus, Smeall Building, Room JS02, Exeter, EX1 2LU, UK
2. Princess Elizabeth Orthopaedic Centre, Royal Devon and Exeter Hospital, Exeter, EX2 5DW, UK
3. Division of Internal Medicine, Department of Symptom Research, The University of Texas MD Anderson Cancer Center, Houston, TX, USA
4. Yong Loo Lin School of Medicine, National University of Singapore, 1E Kent Ridge Road, NUHS Tower Block – Level 9 Singapore 119228

Correspondence to:

Jonathan P Evans

Address: Health Services and Policy Research, Exeter Collaboration for Academic Primary Care (APEx), University of Exeter, Magdalen Campus, Smeall Building, Room JS02, Exeter, EX1 2LU, UK

Email: j.p.evans2@exeter.ac.uk

Abstract

Objective

Over 160,000 participants per year complete the 12-item Oxford Hip and Knee Scores (OHS/OKS) as part of the NHS England Patient Reported Outcome Measures (PROMs) programme. We used a modern computational approach, known as computerised adaptive testing (CAT), to simulate individually-tailored OHS and OKS assessment, with the goal of reducing the number of questions a patient must complete without compromising measurement accuracy.

Methods

We calibrated the 2018/2019 PROMs data to an item response theory (IRT) model. We assessed IRT model assumptions alongside reliability. We used parameters from the IRT model with data from 2017/2018 to simulate CAT assessments. Two simulations were run until a prespecified standard error of measurement was met (SE = .32 and SE = .45). We compared the number of questions required to meet each cut-off and assessed the correlation between the full-length and CAT administration.

Results

We conducted IRT analysis using 40,432 OHS and 44,714 OKS observations. The OHS and OKS were both unidimensional (Root Mean Square Error of Approximation (RMSEA) .08 and .07 respectively) and marginal reliability .91 and .90. The CAT, with a precision limit of SE = .32 and SE = .45 required a median of 4 items (IQR 1) and 2 items (IQR 1) respectively for the OHS, and median of 4 items (IQR 2) and 2 items (IQR 0) for the OKS. This represents a potential 82% reduction in PROM length. In the context of 160,000 yearly assessments, these methodologies could result in the omission of some 1,280,000 redundant questions per year which equates to 40,000 hours of patient time.

Conclusion

The application of IRT to the OHS and OKS produces an efficient and substantially reduced CAT. We have demonstrated a path to reduce the burden and potentially increase the compliance for these ubiquitous outcome measures without compromising measurement accuracy.

Strengths and limitations of this study

- Our study is the first application of computerised adaptive testing on the worlds largest repository of patient reported outcome measures.
- Over 35,000 responses were used in each modelling and simulation group.
- The Oxford Hip and Knee scores are very widely used at an international level.
- This secondary database analysis requires validation in a prospectively collected cohort.
- The available datasets are limited due to attrition that is attributed to the linking of patient reported outcome measure data to health records.

Introduction

The ability to assess a patient's perspective about their health is central to holistic clinical decision making, medical research, and health policy construction.[1] For hip and knee replacement surgery, patients often complete questionnaires called patient reported outcome measures (PROMs) before and after their operation. Since 2009, over 160,000 patients per year undergoing a hip or a knee replacement complete PROMs as part of the NHS England's PROMs Programme.[2]

The PROMs used as part of this programme include the Oxford Hip Score (OHS) and Oxford Knee Score (OKS) which are filled in using pen and paper. Outside of the UK, they are also collected routinely as part of arthroplasty registries in Australia, New Zealand, Canada and the Netherlands.[3] The completion rates across England for the 2018/2019 pre-operative OHS and OKS were 85.7% and 86.1% respectively,[4] however at the hospital trust level, the completion rate varies from 30% to 100%.[5] Attrition is evident when obtaining completed post-operative PROMs (70% completion), further reduction in the data is caused by the process of transcribing the scores to a digital platform and linking with health records which reduces the number of usable records to below 50%.[4] It has been recognised that PROM questionnaires collected using paper and pen for the England PROMs programme are resource-intensive, inefficient for providers and burdensome for patients.[6] The time required to complete orthopaedic PROMs is seen as a key barrier by patients, and the risk of non-completion is highest in those from the most deprived quintile of socioeconomic status and those with poorer general health.[5,7]

Patient-reported outcome measures are composed of a series of questions (items) that ask patients about aspects of their health. These are scored with a structured format to give an estimate of a continuous construct known as a latent trait (i.e. a variable that is not directly observable).[8] Latent traits in orthopaedics typically include pain and physical function. The OHS and OKS were developed using a methodological process called Classical Test Theory (CTT), whereby fixed-length questionnaires were given an overall score, without weighting or standardisation, which estimates the latent trait. More recently, PROMs developed using advanced psychometric techniques have emerged. Influenced by state-of-the-science psychological tests, modelling approaches including Rasch analysis and Item Response Theory (IRT) focus on the individual item within the scale, in contrast to CTT methods, which focus on the total score of all the items together.[9] The ability to calibrate each item individually dramatically increases the versatility of the resulting PROM. Within the IRT paradigm, valid measurement can be obtained using any number of questions from the scale whereas under CTT each item must be administered for the score to be deemed valid. Another

limitation of CTT is that it can only identify items that are not related to the construct being measured, it does not identify items which are redundant (e.g., too similar to others) and can incentivise the inclusion of redundant items.[10] This flexibility is leveraged by a computational technique known as computerised adaptive testing (CAT).[11] A CAT method iteratively select the most informative and relevant items for a particular individual, thereby individualising the assessment to the patient, often resulting in reducing assessment length whilst maintaining acceptable levels of accuracy. Importantly, IRT analyses can be retrospectively applied to legacy PROMs that were initially designed using CTT. Rasch analysis of the OHS and OKS has previously been undertaken, with all studies demonstrating improvements in precision and group discrimination.[12–16] The development of OHS/OKS CAT could improve the efficiency of administration and reduce the administrative burden of the PROMs programme while offering the opportunity to implement a CAT at an unprecedented scale.

The purpose of this study is to assess whether the application of IRT and CAT to the OHS and OKS confers a reduction in questionnaire burden whilst maintaining precision.

Methods

Data

The OHS was developed in 1996, and the OKS in 1998 [17,18] Each PROM contains 12 items that assess joint-specific symptoms over the last four weeks. Each item has five response options that grade the severity of symptoms and functional limitations. Developed following interviews with joint replacement patients they were found to be the best performing condition-specific instruments available in a standardised comparison of the measurement properties.[19]

All Individual item level pre-operative OHS and OKS scores were extracted from the 2018/2019 data release for hip and knee replacements for IRT model development. A second sample for simulation of the CAT was extracted from the 2017/2018 data release.[4] Raw PROMs data are released annually on the NHS digital platform following pre and postoperative linking, health record linking and validation and data cleaning.[4]

Development of the IRT model

We assessed the number of missing responses at the item level and presented them as a percentage difference. We assessed the IRT assumptions of unidimensionality, local independence and monotonicity.[20] To confirm that all items measure a single underlying construct, we assessed unidimensionality using confirmatory factor analysis (CFA). Model fit for CFA was assessed through root mean square error of approximation (RMSEA) with a borderline model fit set at ≤0.08 and good fit ≤0.06, and comparative fit index (CFI) and Tucker-Lewis index (TLI) with borderline model fit set to >0.90 and good fit >0.95 (R package `LAVAAN` version 0.5–23.1097).[21] We confirmed the dimensional structure of each scale using Mokken scaling and assessed scalability (monotonicity) of the items. This assesses whether the probability of scoring the item along its scale of symptom severity increases with a higher level of the underlying construct. A Loevinger's H value of ≥0.3 per item was deemed acceptable (R package `Mokken` version 2.8.4).[22]

Assessment of local independence of items was undertaken to ensure that all items only relate to the dominant construct being measured, and not to a further independent construct. This was assessed by examining the residual covariance between item responses. A high residual covariance may indicate that items are unintentionally measuring another construct, or that they are very similar to each other and potentially redundant. 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 above 0.2 demonstrating locally dependent items.[23]

Following confirmation of IRT model fit assumptions a Graded Response Model (GRM), which is appropriate when item responses can be categorised as ordered categorical[24], was fit to the item response data (R package `mirt` version 3.3.2). This model yields two-item parameters, the item difficulty (a) which is a representation of the level of information about the underlying construct each item provides, and the discrimination (b) thresholds which locate the response categories and their transitions along a contiguous scale. If the item-characteristic curves revealed disordered thresholds, where the response category does not accord with the latent trait score, reordering of adjacent response options was undertaken. Item and model fit was assessed using the RMSEA, TLI, CFI and Standardized Root Mean Square Residual (SRMSR). Reliability in the IRT model was estimated as marginal reliability where the overall reliability of the test was based on the average conditional standard errors.[25] This overall index of precision can be compared to the classical internal consistency (Cronbach's alpha) reliability estimate for CTT, where scores >0.8 indicate excellent reliability.

Computerised adaptive testing simulation

The production of item thresholds and difficulty information from the IRT models allows the construction of a CAT. The administration of a CAT utilises algorithms, which match participants to the most informative items within a PROM and once an acceptable level of precision is reached, denoted by the reliability (standard error (SE)), of the latent trait estimate, no further items are

required.[26] Within a CAT simulation, the estimate of the latent trait from the full-length PROM can be compared to the delivery of shortened versions where particular items are selected. This simulation can provide information on the number of items needed to provide estimates of the latent trait at predetermined levels of precision. Through the simulation, the items that provide the highest level of information, and thereby the greatest utility in shortened versions, can be determined.

We performed a CAT simulation using Firestar for R (version 1.3.2).[27] Two separate simulations were conducted for OHS and OKS with the 2017/18 dataset with predetermined stopping criteria (precision) denoted as a SE of the latent trait estimate of <0.32 and <0.45. These SE values are equivalent to a reliability coefficient of 0.90 and 0.80 respectively. Variables derived from the simulation include the correlation (Intraclass Correlation Coefficient, (ICC)) between the latent trait estimation of the full-length questionnaire and the CAT, and the mean and standard deviation, median and interquartile range (IQR) 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 is presented as a percentage difference. Time-saving between full-length and CAT administration were calculated against the estimate that each item takes between 10 seconds and 75 seconds per item to complete, a time extrapolated from published reports of total completion time of two to 15 minutes for the 12-item questionnaire.[28]

All data analysis was conducted in R (RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA).

Public and Patient Involvement

Formal patient and public involvement was not undertaken for this analysis of public domain data. The national PROMs programme and the data held within have themselves been evaluated via

public consultation. Response to this evaluation from a multiple stakeholders taskforce highlighted the need to improve efficiency of data collection.[6]

Data availability statement

Difficulty and discrimination parameters of the IRT model for both the OHS and OKS are available in Appendix 1. All data are available from NHS digital and can be used in accordance with the open government licence for public sector information.

Results

Dataset characteristics

Of the 40,172 preoperative OHS scores and 44,264 OKS scores in the 2018/19 data, 1,704 were revision hip replacements and 1,162 revision knee replacements, which were excluded. Further exclusion of incomplete questionnaires resulted in 37,995 OHS and 42,558 OKS observations. Missing responses to items were found 4118 (0.90%) and 4803 (0.93%) times for the OHS and OKS. Although the first two items had few missing responses (0.15% OHS, 0.10% OKS), the remaining ten had substantially more (1.04% OHS, 1.1% OKS), indicating a possible patient preference for shorter measures.

For hip replacements, 59.4% were undertaken in females, 93.8% were ≥50 years, and 51.9% were ≥70 years old. 14.8% of respondents had assistance completing the questionnaires, median symptom duration was one to five years. For knee replacements, 56.5% were undertaken in females, 97.4% were ≥50 years and 51.8% were ≥70 years old. 14.6% had assistance completing their questionnaires, median symptom duration was one to five years. All demographic features of the PROMS dataset were equivalent to that of the full National Joint Registry.

IRT model assumptions

The criterion of unidimensionality was met at a borderline level for both OHS and OKS with an RMSEA of 0.08 (OHS) and 0.07 (OKS), CFI of 0.93 (OHS) and 0.94 (OKS), TFI of 0.91 (OHS) and 0.93 (OKS). Mokken scaling corroborated this finding of unidimensionality and produced overall scalability coefficients (H) of 0.49 (range 0.41 – 0.58) (OHS) and 0.46 (0.38 – 0.55) (OKS). Local independence of items was confirmed for both OHS and OKS with all item correlations below 0.02. There were no misfitting items within the GRM model.

Following production of the IRT item characteristic curves disordered thresholds (where the curve lies under the line created by an adjacent curve) were noted in items 5, 6, 9, 10 and 12 for the OHS

and items 4, 6 and 8 for the OKS. Items with disordered thresholds were rescored, giving them the same score as the adjacent item whose area it lay within (Fig 1a & 1b). Item level RMSEA was good for both scores with all items RMSEA <0.02 (Appendix 1). The marginal reliability of the model was 0.91 for OHS and 0.90 for OKS. Overall model fit was boarderline for the OHS (RMSEA 0.09, SRMSR 0.05, TLI 0.82 and CFI 0.90) and borderline to good for OKS (RMSEA 0.06, SRMSR 0.04, TLI 0.91 and CFI 0.94).

CAT simulation

We conducted a CAT simulation using the derived IRT parameters and utilising the preoperative OHS and OKS item responses from the 2017/2018 data release as the testing set. For the OHS 36,516 participants scores were included, and for the OKS 45,122. Incomplete records (i.e. less than 12 item scores) were included as the IRT method accounts for missing data, using all available responses to gain the best estimate of the latent trait. At the standard error threshold of 0.32 (corresponding to a reliability of 0.9) the ICC between full-length and CAT latent trait estimates was r = 0.96 (OHS) and r = 0.96 (OKS) (fig 2). For the OHS CAT, the mean number of items required was 3.98 (SD 1.26) with a median of 4 (IQR 1)). For the OKS CAT, the mean number of items required was 4.22 (SD 1.32) with a median of 4 (IQR 2) (fig 3).

With a precision SE threshold of 0.45 (corresponding to a reliability of 0.8), the concordance between full-length and CAT simulations decreased marginally to r = 0.90 (OHS) and r = 0.91 (OKS) (fig 2). The OHS CAT required a mean of 2.27 (SD 0.45) items, median of 2 (IQR 1). The OKS CAT required 2.13 (SD 0.45) items, median of 2 (IQR 0) (fig 3).

For the OHS, the simulation selected item 8 as the starting item for all participants, unless item 8 was not scored. For the OKS, item 9 was used as the staring item. Overall, when all items are collated, the items used most frequently within the 0.35 SE CAT were items 8 (24.9%) and 11 (21.3%) for OHS, and items 9 (23.5%), 11 (23.4%) and 12 (16.5%) for the OKS, all other items were used less

than 16% of the time. At 0.45 SE, OHS items 3 (19.9%), 8 (43.7%) and 11 (32.9.0%) were used most frequently within the simulations, four items were not used in any simulations and all other items were used less than 3% of the time. For the OKS items, 9 (46.1%) and 11 (45.7.0%) were used most frequently, item 12 (5.1%) was minimally utilised leaving two items that were not required in any simulations (including amongst them items 6, and 8 whose response options had been identified as disordered) and all others were used in less than 1.4% of simulations (fig 4)(Appendix 1).

The items utilised most frequently in estimating the level of the latent trait, and selected as the starting item within the simulations, were item 8 for the OHS (*During the past 4 weeks… After a meal (sat at a table), how painful has it been for you to stand up from a chair because of your hip?*) and within the OKS were item 9 (*During the past 4 weeks… How much has pain from your knee interfered with your usual work (including housework)?*). Item 11 was also consistently utilised as the second item for the OKS simulations (*During the past 4 weeks… Could you do the household shopping on your own?*).

Out of a potential 438,192 items for the OHS scores, only 145,462 items were used by the CAT at 0.32 SE, and 82,980 at 0.45 SE. This represents a 100.3% and 136.3% difference. Taking the whole 2018/19 NJR dataset before exclusions of 95,977 total hip replacements, at 0.45 SE this represents a potential time saving of 2583 – 19374 hours for pre-operative scores. Out of a potential 541,464 items for the OKS scores, only 190,410 items were used by the CAT at 0.32 SE, and 96,922 at 0.45 SE, representing a 100.3% and 136.3% difference. For the entire NJR dataset in 2018/19, at 0.45 SE this represents a potential saving of between 2832.2 – 21241.5 hours for collection of pre-operative scores.

Discussion

The use of PROMs in the outcome assessment of hip and knee replacements is widely accepted. The best PROMs for patients, researchers and clinicians are easy to understand, free from redundancy, and psychometrically robust. In this study, we have applied a modern psychometric approach to the one of the world's largest repositories of orthopaedic arthroplasty PROMs. The OHS and OKS conformed to IRT assumptions by demonstrating unidimensionality, monotonicity, and local independence. CAT simulations demonstrated the possibility to dramatically reduce the length of these 12-item PROMs to as little as two items at a high level of precision. In large-scale data collection, the potential time saving from the deployment of a CAT is equivalent to more than a million redundant questionnaire items per year and more than 4 years of collective patient time annually.

Although Rasch analysis, a type of IRT methodology, has previously been applied to both the OHS and OKS,[13,15] no research to date has explored the possibility of using these methods to reduce the burden of assessment using CAT. CAT achieved the goal of minimising the burden of a PROM by only delivering the most relevant and informative items required to measure a patient's level of hip or knee pain and function.[29] The simulations performed within this study were able to reduce the number of items required by 67% for the OHS and 65% for the OKS at 0.32 SE (equivalent to 90% precision) and as much as 81% for both PROMs at 0.45 SE (80% precision). As a comparator, the reliability, (taken as a proxy marker of precision) of the 12-item OHS and OKS delivered, (using the classical test theory derived scoring system) has a test-retest Intraclass Correlation Coefficient (ICC) of 0.82 – 0.94.[30,31] Although this is excellent, delivery of the full test does not demonstrate superiority over a CAT administration. A minimal reliability threshold of 0.70 is commonly accepted for PROMs, such as those used in the NHS England PROMS programme. The standard error at this reliability level is 0.55 of a standard deviation, which is roughly equivalent to a reliability of .70.[32] Similarly, a SE of 0.45, (equivalent to .80 reliability), in this simulation, a median of only two items were required to estimate patients' pain and function dramatically reducing the burden on a patient.

Interestingly, although the overall completion rate of the PROMs was high, within this sample, the non-completion of items substantially increased after the first two items, and then remained stable for the remaining 10 items. Whether this is related to the item structure or order, or indeed whether this is related to the mode in which the OHS and OKS are delivered as part of the National PROMs programme is uncertain. The OHS and OKS are asked as part of a battery of tests within the National PROMs programme, overall 27 questions are asked within an 8 page booklet. Beyond the OHS and OKS, the questions include the 3-level EuroQOL 5-Dimension PROM, co-morbidity profiles, surgical history, symptom duration and demographic profiles. Within this question set some repetition exists, and the non-completion or partial completion may relate to the size of this dataset. Minimising respondent fatigue by simply asking two items is likely to improve completion rates both at the start of data collection and longitudinally, thereby optimising the utility of this valuable data.[33] The value and reliability of PROMs is vastly improved by regular administration over time, the ability to conduct this with targeted highly condensed PROMs that retain their ability to precisely estimate the latent trait is only possible through IRT analysis and CAT administration.[26] Furthermore, the two items used most frequently in the CAT deployment for OHS (Items 8 and 11), have been judged by patients as having the most clarity and fewest limitations.[34] Of interest, within both questionnaires, the pre-operative items pertaining to function rather than pain were selected by the simulation as most valuable.

The use of IRT-derived PROMs is becoming increasingly prevalent in efforts to advance high-value care and improve shared decision-making.[35] The ability to score on a simple continuum (eg 0-100) and derive population norms (eg a score of 50), vastly improves patient comprehension of their score. Patients understanding of the relevance of their PROM score improves their compliance with future assessment and optimises the use of a PROM as a decision aid.[36] The use of this latent trait continuum that is independent of the PROM also allows comparison of the OHS and OKS scores with other scores assessing the same trait. Therefore, so-called “cross-walks” can be derived to compare the scores derived from the OHS and OKS with other hip scores such as the Hip disability and Knee

injury and Osteoarthritis Outcome Scores (HOOS and KOOS) that have also undergone IRT analysis,[37] or contemporarily designed PROMs such as the PROMIS physical function and pain interference scores. This attribute can have a profound effect on the translatability of research findings. Lastly, IRT level analysis also opens up future assessment of differential item functioning (DIF). Here an exploration of the extent to the item may be measing different abilities dependent on variables such as age, gender, comorbidity profile or operation type could be undertaken.

The authors recognise limitations inherent to this study. We recognise that the dimensionality of both the OHS and OKS could be contested on the basis of the borderline results. It has previously been identified that both one-factor and two-factor models fit these scores.[38,39] As the most commonly applied scoring method utilises the total score for this very common PROM, it was deemed appropriate to maintain a unidimensional model. The authors do recognise that by proposing an alternative method of scoring, there is a risk of loosing legacy knowledge relating to the Oxford scores, to ameliorate this risk we would recommend the provision of a conversion matrix to allow the presentation of IRT and CTT based scoring. We also recognise that the IRT parameters were derived on pre-operative data, and therefore further analysis of post-operative data would be required, of particular importance would be an assessment ceiling effect under this revised scoring metric. We would recommend using the IRT deried parameters and the availability of the full question bank in the post operative population, rather than a specifically reduced short-form version. To improve the interpritability of the score, we would also recommend IRT derived minimal important difference calculation for the OHS and OKS. By contextualising the differences in scorethat would be deemed relevant to patients, this would inform the utility of this method in trial design and as a potential adjunct to communication and decision making. The strength of this analysis is the very large sample size for the IRT model construction. The CAT simulation requires validation on patients with both qualitative and quantitative analysis of validity and acceptability. The significant limitation to the practical application of IRT and CAT is the availability of a computer and an appropriate interface, and we recognise that currently this national programme collects this data

through pen and paper completion and postal communication. However, the utilisation of the PROMIS system in the USA highlights that these barriers can be overcome, furthermore, the increasing ubiquity of tablet and smartphone interfaces and the often-underappreciated technological literacy of this patient population suggest that this problem is far from insurmountable. Both during and in the post-pandemic era, remote medicine is becoming the norm; refined PROMs collection has a vital role to play in this process.

Conclusion

The collection of hip and knee outcome measures for the NHS England National PROMs programme has been criticised as remote from patient care. By applying modern psychometric analysis to the world's largest repository of hip and knee patients PROMs, we have demonstrated up to an 80% reduction in the number of items required to estimate the patient-specific impact of joint disease without compromising precision. Widespread adoption of this system has the potential to reduce participant burden and increase completion rates, thereby maximising the reliability and utility of longitudinal data.

Contributors

JP Evans (JPE) MBChB, MSc, MD(res), FRCS (Tr and Orth) is an NIHR Academic Clinical Lecturer in health services research and is a senior fellow in trauma and orthopaedics. C Gibbons (CG) PhD is Deputy Chair and Associate Professor in symptom research and patient reported outcome measures. A Toms(AT) MB ChB FRCS (Ed) MSc Eng FRCS (Tr & Orth) is a Consultant Trauma and Orthopaedic Surgeon and Honorary Clinical Professor. JM Valderas (JMV) MPH, PhD is Professor of health services and policy research. JPE, CG and JMV devised the project. JPE managed the data and analysed the results with oversight from CG and JMV. AT contributed to the data management and interpretation. JPE wrote the manuscript. CG, AT and JMV edited the manuscript. All authors read and approved the final manuscript. The lead author (Jonathan P Evans) affirms that the manuscript is an honest, accurate, and transparent account and that no important aspects of the study have been omitted.

Ethics approval

Formal ethics approval was not required. All data used in accordance with the open government licence for public sector information. The National Archives. Open Government Licence for public sector information. 2020. http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/ (accessed Dec 21, 2020).

Funding

JPE is in receipt of an NIHR Academic Clinical Lecturer award. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. There was no involvement of the funder in study design, data collection, data analysis, manuscript preparation or publication decisions. All authors had complete access to the study data that support the publication.

Competing interests

We declare that we have no competing interests.

References

1 Black N, Burke L, Forrest CB, *et al.* Patient-reported outcomes: pathways to better health, better services, and better societies. *Qual Life Res* 2016; **25**: 1103–12.

2 Health & Social Care Information Centre. National PROMs programme. .

3 Wilson I, Bohm E, Lübbeke A, *et al.* Orthopaedic registries with patient-reported outcome measures. *EFORT Open Rev* 2019; **4**: 357–67.

4 NHS Digital. Finalised PROMs data release. Patient Reported Outcome Measures (PROMs) in England for Hip and Knee Replacement Procedures (April 2018 to March 2019). 2020. https://digital.nhs.uk/data-and-information/publications/statistical/patient-reported-outcome-measures-proms/hip-and-knee-replacement-procedures-april-2019-to-march-2020 (accessed Dec 21, 2020).

5 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.

6 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**: 441–5.

7 Rowland C, Walsh L, Harrop R, Roy B, Skevington SM. What Do U.K. Orthopedic Surgery Patients Think About PROMs? Evaluating the Evaluation and Explaining Missing Data. *Qual Health Res* 2019; **29**: 2057–69.

8 Gorter R, Fox J-P, Twisk JWR. Why item response theory should be used for longitudinal questionnaire data analysis in medical research Data analysis, statistics and modelling. *BMC Med Res Methodol* 2015; **15**. DOI:10.1186/s12874-015-0050-x.

9 Cappelleri JC, Lundy JJ, Hays RD. Overview of classical test theory and item response theory

for the quantitative assessment of items in developing patient-reported outcomes measures. *Clin Ther* 2014; **36**: 648–62.

10 Sijtsma K. On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika* 2009; **74**: 107.

11 Cella D, Gershon R, Lai J-S, Choi S. The future of outcomes measurement: Item banking, tailored short-forms, and computerized adaptive assessment. *Qual Life Res* 2007; **16**: 133–41.

12 Ko Y, Lo NN, Yeo SJ, *et al.* Comparison of the responsiveness of the SF-36, the Oxford Knee Score, and the Knee Society Clinical Rating System in patients undergoing total knee replacement. *Qual Life Res* 2013; **22**: 2455–9.

13 Ko Y, Lo NN, Yeo SJ, *et al.* Rasch analysis of the Oxford Knee Score. *Osteoarthr Cartil* 2009; **17**: 1163–9.

14 Norquist JM, Fitzpatrick R, Dawson J, Jenkinson C. Comparing alternative Rasch-based methods vs raw scores in measuring change in health. *Med Care* 2004; **42**. DOI:10.1097/01.mlr.0000103530.13056.88.

15 Fitzpatrick R, Norquist JM, Jenkinson C, *et al.* A comparison of Rasch with Likert scoring to discriminate between patients' evaluations of total hip replacement surgery. *Qual Life Res* 2004; **13**: 331–8.

16 Fitzpatrick R, Norquist JM, Jenkinson C, *et al.* A comparison of Rasch with Likert scoring to discriminate between patients' evaluations of total hip replacement surgery. *Qual Life Res* 2004; **13**: 331–8.

17 Dawson J, Fitzpatrick R, Carr A, Murray D. Questionnaire on the perceptions of patients about total hip replacement. *J Bone Joint Surg Br* 1996; **78**: 185–90.

18 Dawson J, Fitzpatrick R, Murray D, Carr A. Questionnaire on the perceptions of patients about

total knee replacement. *J Bone Joint Surg Br* 1998; **80**: 63–9.

19 Harris K, Dawson J, Gibbons E, *et al.* Systematic review of measurement properties of patient-reported outcome measures used in patients undergoing hip and knee arthroplasty. *Patient Relat Outcome Meas* 2016; **Volume 7**: 101–8.

20 Cella D, Yount S, Rothrock N, *et al.* The Patient-Reported Outcomes Measurement Information System (PROMIS): Progress of an NIH roadmap cooperative group during its first two years. *Med Care* 2007; **45**: S3–11.

21 Rosseel Y. Lavaan: An R package for structural equation modeling and more. Version 0.5–12 (BETA). *J Stat Softw* 2012; **48**: 1–36.

22 Van der Ark LA. Mokken scale analysis in R. *J Stat Softw* 2007; **20**: 1–19.

23 Yen WM. Scaling Performance Assessments: Strategies for Managing Local Item Dependence. *J Educ Meas* 1993; **30**: 187–213.

24 Hays RD, Morales LS, Reise SP. Item Response Theory and Health Outcomes Measurement in the 21st Century NIH Public Access. .

25 Green BF, Bock RD, Humphreys LG, Linn RL, Reckase MD. Technical guidelines for assessing computerized adaptive tests. *J Educ Meas* 1984; **21**: 347–60.

26 Gibbons CJ. Turning the page on pen-and-paper questionnaires: combining ecological momentary assessment and computer adaptive testing to transform psychological assessment in the 21st Century. *Front Psychol* 2017; **7**: 1933.

27 Choi SW. Firestar: Computerized adaptive testing simulation program for polytomous item response theory models. *Appl Psychol Meas* 2009; **33**: 644.

28 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. *BMJ Qual Saf* 1999; **8**: 228–33.

29 Cook KF, O'Malley KJ, Roddey TS. Dynamic assessment of health outcomes: time to let the CAT out of the bag? *Health Serv Res* 2005; **40**: 1694–711.

30 Gagnier JJ, Huang H, Mullins M, *et al.* Measurement properties of patient-reported outcome measures used in patients undergoing total hip arthroplasty: A systematic review. *JBJS Rev* 2018; **6**. DOI:10.2106/JBJS.RVW.17.00038.

31 Gagnier JJ, Mullins M, Huang H, *et al.* A Systematic Review of Measurement Properties of Patient-Reported Outcome Measures Used in Patients Undergoing Total Knee Arthroplasty. J. Arthroplasty. 2017; **32**: 1688-1697.e7.

32 Reeve BB, Wyrwich KW, Wu AW, *et al.* ISOQOL recommends minimum standards for patient-reported outcome measures used in patient-centered outcome1. Reeve BB, Wyrwich KW, Wu AW, Velikova G, Terwee CB, Snyder CF, et al. ISOQOL recommends minimum standards for patient-reported outcome measures us. *Qual Life Res* 2013; **22**: 1889–905.

33 Krosnic J, Presser S. Question and Questionnaire Design. In "Handbook of Survey Research", 2nd edn. Elsevier, 2013.

34 Wylde V, Learmonth ID, Cavendish VJ. The Oxford hip score: the patient's perspective. *Health Qual Life Outcomes* 2005; **3**: 1–8.

35 Brodke DJ, Hung M, Bozic KJ. Item response theory and computerized adaptive testing for orthopaedic outcomes measures. *JAAOS-Journal Am Acad Orthop Surg* 2016; **24**: 750–4.

36 Porter I, Gonçalves-Bradley D, Ricci-Cabello I, *et al.* Framework and guidance for implementing patient-reported outcomes in clinical practice: evidence, challenges and opportunities. *J Comp Eff Res* 2016; **5**: 507–19.

37 Gandek B, Roos EM, Franklin PD, Ware JE. Item selection for 12-item short forms of the Knee

injury and Osteoarthritis Outcome Score (KOOS-12) and Hip disability and Osteoarthritis Outcome Score (HOOS-12). *Osteoarthr Cartil* 2019; **27**: 746–53.

38 Harris KK, Price AJ, Beard DJ, Fitzpatrick R, Jenkinson C, Dawson J. Can pain and function be distinguished in the Oxford Hip Score in a meaningful way?: An exploratory and confirmatory factor analysis. *Bone Jt Res* 2014; **3**: 305–9.

39 Harris K, Dawson J, Doll H, *et al.* Can pain and function be distinguished in the Oxford Knee Score in a meaningful way? An exploratory and confirmatory factor analysis. *Qual Life Res* 2013; **22**: 2561–8.

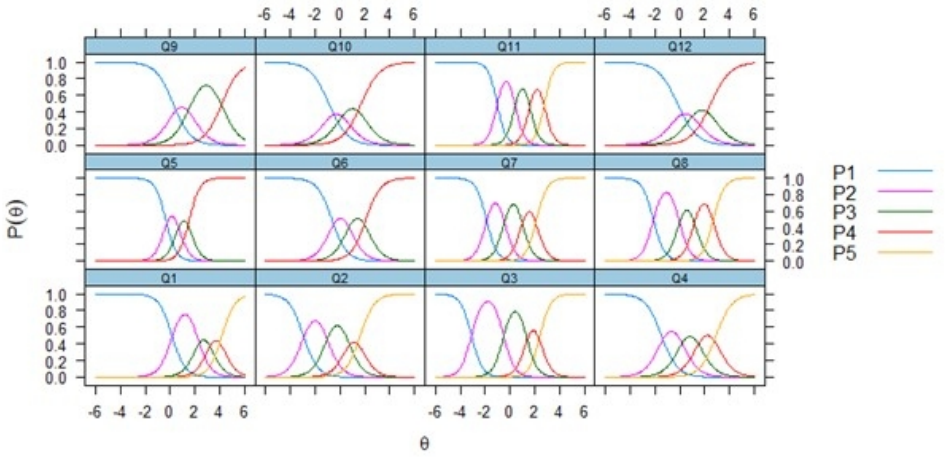
Figure Titles

Figure 1: Item response theory (IRT) item traces for the 12 items of the Oxford Hip Score (OHS)(a) and Oxford Knee Score (OKS)(b)

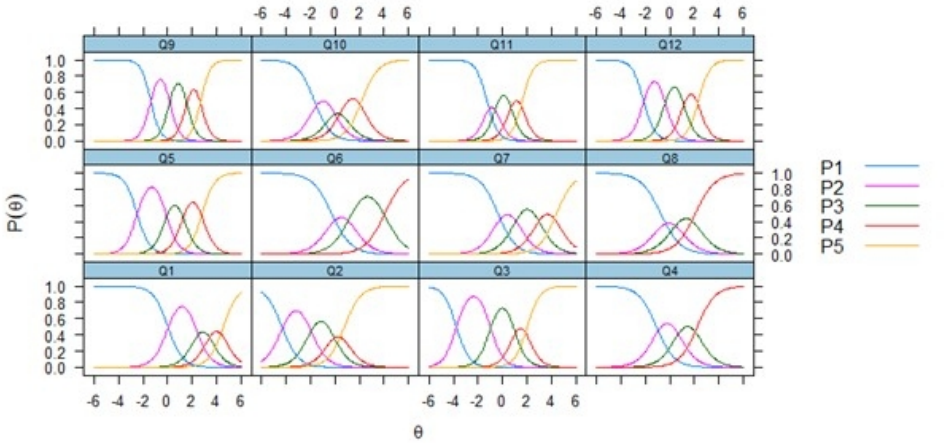
Figure 2: Scatter plot and correlation between the theta estimation (values of the latent trait) between the full 12-item administration and the computerised adaptive test (CAT) for the Oxford His Score (OHS) (a & b) and Oxford Knee Score (OKS) (c & d) at 0.32 standard error (SE) and 0.45 SE

Figure 3: Bar chart showing the number of items used per participant at 0.32 standard error (SE) and 0.45 SE for the OHS (a, b) and OKS (c, d) computerised adaptive test (CAT)

Figure 4: Bar chart showing the proportional use of each item at 0.35 standard error (SE) and 0.45 SE for the OHS (a, b) and OKS (c, d) computerised adaptive test (CAT)



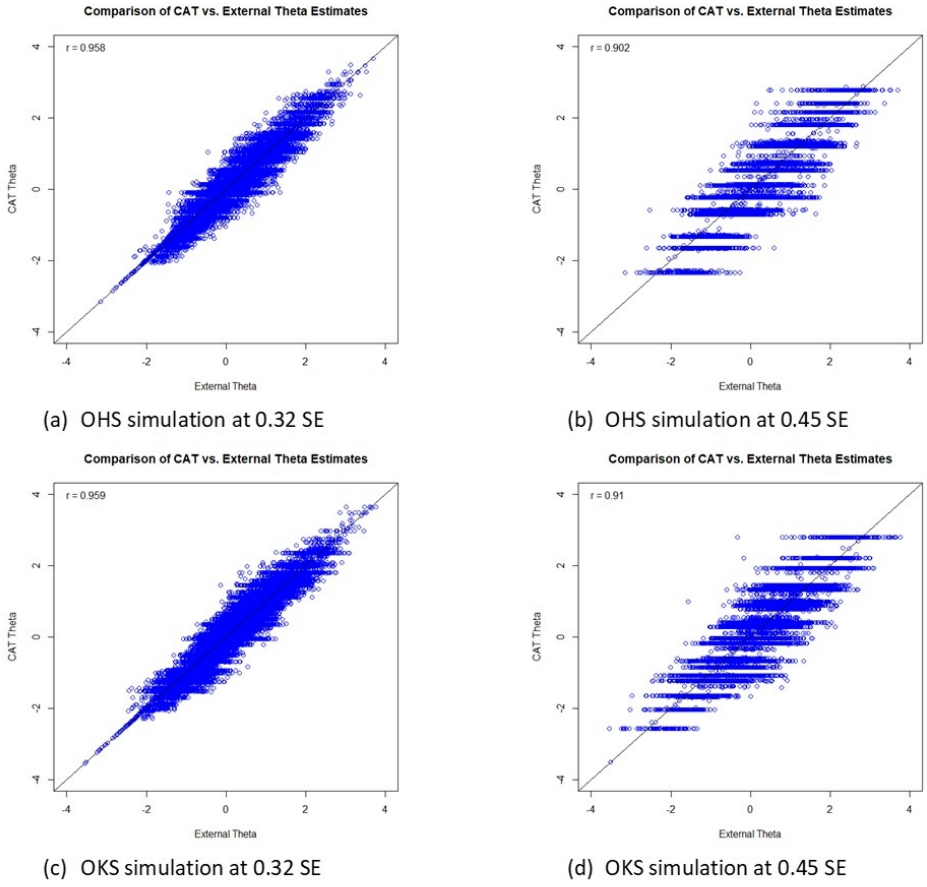
(a) OHS IRT Item trace



(b) OKS IRT Item trace

Figure 1: Item response theory (IRT) item traces for the 12- items of the Oxford Hip Score (OHS)(a) and Oxford Knee Score (OKS)(b)

106x117mm (150 x 150 DPI)



(a) OHS simulation at 0.32 SE

(b) OHS simulation at 0.45 SE

(c) OKS simulation at 0.32 SE

(d) OKS simulation at 0.45 SE

Figure 2: Scatter plot and correlation between the theta estimation (values of the latent trait) between the full 12-item administration and the Computerised Adaptive Test (CAT) for the Oxford His Score (OHS) (a & b) and Oxford Knee Score (OKS) (c & d) at 0.32 Standard Error (SE) and 0.45 SE.

166x165mm (150 x 150 DPI)

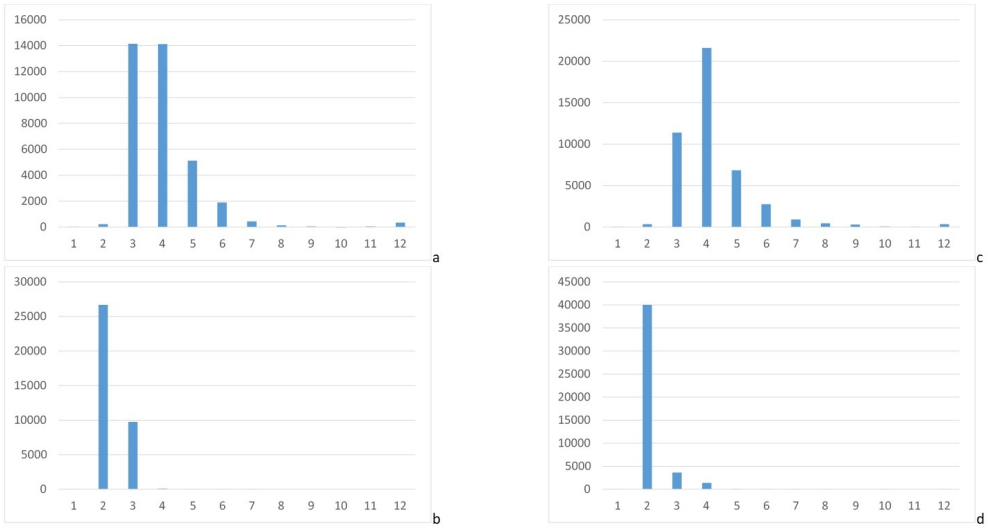


Figure 3: Bar chart showing the number of items used per participant at 0.32 Standard Error (SE) and 0.45 SE for the OHS (a, b) and OKS (c, d) Computerised Adaptive Test (CAT).

246x136mm (150 x 150 DPI)

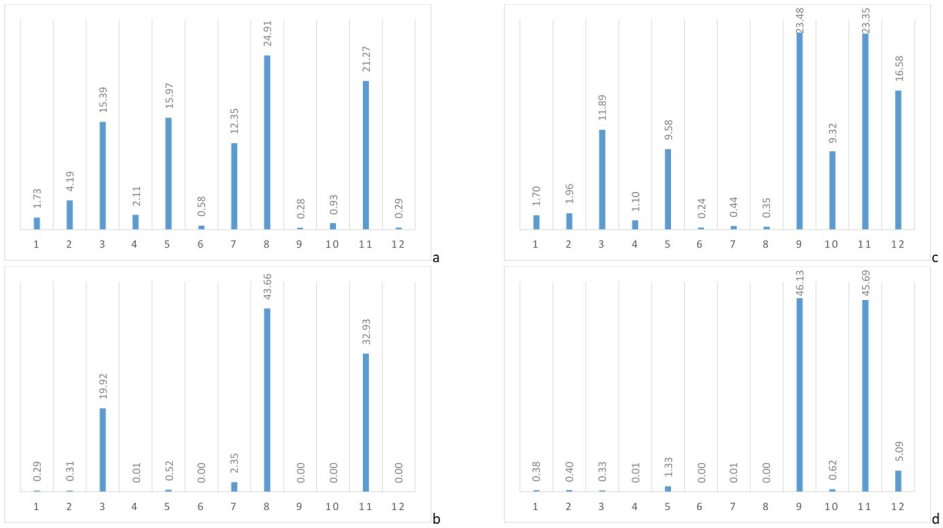


Figure 4: Bar chart showing the proportional use of each item at 0.35 Standard Error (SE) and 0.45 SE for the OHS (a, b) and OKS (c, d) Computerised Adaptive Test (CAT).

246x136mm (150 x 150 DPI)

Appendix 1

	Item	Discrimination a	Difficulty b1	Difficulty b2	Difficulty b3	Difficulty b4	Item RMSEA
1	During the past 4 weeks..... How would you describe the pain you usually had from your hip?	1.914691405	0.164325049	2.286224388	3.209345224	4.186925039	0.011
2	During the past 4 weeks..... Have you had any trouble with washing and drying yourself (all over) because of your hip?	1.740856269	-2.940926073	-1.052645662	0.609238754	1.637069089	0.003
3	During the past 4 weeks..... Have you had any trouble getting in and out of a car or using public transport because of your hip? (whichever you tend to use)	2.321961482	-3.093433431	-0.486468803	1.340835438	2.431036075	0.014
4	During the past 4 weeks..... Have you been able to put on a pair of socks, stockings or tights?	1.547750826	-1.475337619	0.158801902	1.50632625	2.924283288	0.009
5	During the past 4 weeks..... Could you do the household shopping on your own?	2.376360472	-0.348672227	0.6[illegible]89	1.549282667	NA	0.008
6	During the past 4 weeks..... For how long have you been able to walk before pain from your hip becomes severe? (with or without a stick)	1.668616909	-0.642568044	0.719[illegible]66	2.086871457	NA	0.015
7	During the past 4 weeks..... Have you been able to climb a flight of stairs?	2.360767142	-1.889430738	-0.436[illegible]962	0.981278344	2.144165736	0.007
8	During the past 4 weeks..... After a meal (sat at a table), how painful has it been for you to stand up from a chair because of your hip?	2.258019654	-2.12187152	-0.[illegible]6100465	1.205103327	2.706143566	0.012
9	During the past 4 weeks..... Have you been limping when walking, because of your hip?	1.42746619	0.251525777	1.6[illegible]3812085	4.171461153	NA	0.007
10	During the past 4 weeks..... Have you had any sudden, severe pain - 'shooting', 'stabbing' or 'spasms' - from the affected hip?	1.324519295	-0.892084132	0.2[illegible]9697408	1.709666854	NA	0.011
11	During the past 4 weeks..... How much has pain from your hip interfered with your usual work (including housework)?	2.775690212	-1.028278183	0.4[illegible]5519961	1.607031821	2.776328526	0.006
12	During the past 4 weeks..... Have you been troubled by pain from your hip in bed at night?	1.260879482	-0.194492803	1.0[illegible]872884	2.493862719	NA	0.011

Table 1 : Oxford Hip Score items with associated IRT derived difficulty and discrimination parameters.

Appendix 1

	Item	Discrimination a	Difficulty b1	Difficulty b2	Difficulty b3	Difficulty b4	Item RMSEA
1	During the past 4 weeks..... How would you describe the pain you usually have from your knee?	1.683615138	0.035557101	2.328205621	3.433951121	4.561327448	0.005
2	During the past 4 weeks..... Have you had any trouble with washing and drying yourself (all over) because of your knee?	1.492252738	-4.326258935	-2.018457092	-0.292460352	0.770869766	0.016
3	During the past 4 weeks..... Have you had any trouble getting in and out of a car or using public transport because of your knee? (whichever you would tend to use)	1.932656761	-3.747375182	-0.968344144	0.945305979	2.030550642	0.007
4	During the past 4 weeks..... For how long have you been able to walk before pain from	1.387915921	-1.095979526	0.648687459	2.23788967	NA	0.010
5	During the past 4 weeks..... After a meal (sat at a table), how painful has it been for you to stand up from a chair because of your knee?	1.973493643	-2.473540467	-0.110575121	1.302091221	2.835473629	0.003
6	During the past 4 weeks..... Have you been limping when walking, because of your knee?	1.263415959	-0.270638346	1.272611	4.061594039	NA	0.011
7	During the past 4 weeks..... Could you kneel down and get up again afterwards?	1.413075377	-0.361382764	1.150705202	2.902993876	4.420398561	0.018
8	During the past 4 weeks..... Have you been troubled by pain from your knee in bed at night?	1.23865998	-0.757873886	0.547077749	2.065465542	NA	0.005
9	During the past 4 weeks..... How much has pain from your knee interfered with your usual work (including housework)?	2.563072755	-1.375193192	0.177461486	1.561747127	2.724240742	0.008
10	During the past 4 weeks..... Have you felt that your knee might suddenly 'give way' or let you down?	1.507070288	-1.693738699	-0.259656505	0.672066751	2.202410973	0.008
11	During the past 4 weeks..... Could you do the household shopping on your own?	2.235209642	-1.264960065	-0.483151955	0.662939289	1.641645648	0.019
12	2 During the past 4 weeks..... Could you walk down one flight of stairs?	2.135163585	-2.14417636	-0.398774015	1.11031124	2.341481914	0.011

Table 2 : Oxford Knee Score items with associated IRT derived difficulty and discrimination parameters.

The RECORD statement – checklist of items, extended from the STROBE statement, that should be reported in observational studies using routinely collected health data.

	Item No.	STROBE items	Location in manuscript where items are reported	RECORD items	Location in manuscript where items are reported
Title and abstract					
	1	(a) Indicate the study's design with a commonly used term in the title or the abstract (b) Provide in the abstract an informative and balanced summary of what was done and what was found	Page 2	RECORD 1.1: The type of data used should be specified in the title or abstract. When possible, the name of the databases used should be included. RECORD 1.2: If applicable, the geographic region and timeframe within which the study took place should be reported in the title or abstract. RECORD 1.3: If linkage between databases was conducted for the study, this should be clearly stated in the title or abstract.	Page 2
Introduction					
Background rationale	2	Explain the scientific background and rationale for the investigation being reported	Pages 3-4		
Objectives	3	State specific objectives, including any prespecified hypotheses	Pages 4-5		
Methods					
Study Design	4	Present key elements of study design early in the paper	Pages 6-8		
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	Page 6		

Participants	6	*(a) Cohort study* - Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up *Case-control study* - Give the eligibility criteria, and the sources and methods of case ascertainment and control selection. Give the rationale for the choice of cases and controls *Cross-sectional study* - Give the eligibility criteria, and the sources and methods of selection of participants *(b) Cohort study* - For matched studies, give matching criteria and number of exposed and unexposed *Case-control study* - For matched studies, give matching criteria and the number of controls per case	Page 6	RECORD 6.1: The methods of study population selection (such as codes or algorithms used to identify subjects) should be listed in detail. If this is not possible, an explanation should be provided. RECORD 6.2: Any validation studies of the codes or algorithms used to select the population should be referenced. If validation was conducted for this study and not published elsewhere, detailed methods and results should be provided. RECORD 6.3: If the study involved linkage of databases, consider use of a flow diagram or other graphical display to demonstrate the data linkage process, including the number of individuals with linked data at each stage.	Page 6
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable.	Pages 6-7	RECORD 7.1: A complete list of codes and algorithms used to classify exposures, outcomes, confounders, and effect modifiers should be provided. If these cannot be reported, an explanation should be provided.	Page 6
Data sources/ measurement	8	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group	Page 6		

Bias	9	Describe any efforts to address potential sources of bias	NA		
Study size	10	Explain how the study size was arrived at	Page 6		
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen, and why	Pages 6 – 8		
Statistical methods	12	(a) Describe all statistical methods, including those used to control for confounding (b) Describe any methods used to examine subgroups and interactions (c) Explain how missing data were addressed (d) *Cohort study* - If applicable, explain how loss to follow-up was addressed *Case-control study* - If applicable, explain how matching of cases and controls was addressed *Cross-sectional study* - If applicable, describe analytical methods taking account of sampling strategy (e) Describe any sensitivity analyses	Pages 6 – 8		
Data access and cleaning methods		..		RECORD 12.1: Authors should describe the extent to which the investigators had access to the database population used to create the study population.	Pages 3 and 6

				RECORD 12.2: Authors should provide information on the data cleaning methods used in the study.	
Linkage		..		RECORD 12.3: State whether the study included person-level, institutional-level, or other data linkage across two or more databases. The methods of linkage and methods of linkage quality evaluation should be provided.	Non-linked
Results					
Participants	13	(a) Report the numbers of individuals at each stage of the study (*e.g.*, numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed) (b) Give reasons for non-participation at each stage. (c) Consider use of a flow diagram	Page 9	RECORD 13.1: Describe in detail the selection of the persons included in the study (*i.e.,* study population selection) including filtering based on data quality, data availability and linkage. The selection of included persons can be described in the text and/or by means of the study flow diagram.	No filtering applied
Descriptive data	14	(a) Give characteristics of study participants (*e.g.*, demographic, clinical, social) and information on exposures and potential confounders (b) Indicate the number of participants with missing data for each variable of interest (c) *Cohort study* - summarise follow-up time (*e.g.*, average and total amount)	Page 9		
Outcome data	15	*Cohort study* - Report numbers of outcome events or summary measures over time *Case-control study* - Report numbers in each exposure	Page 9		

10.1136/bmjopen-2021-053815 on 20 July 2022. Downloaded from

		category, or summary measures of exposure *Cross-sectional study* - Report numbers of outcome events or summary measures			
Main results	16	(a) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (e.g., 95% confidence interval). Make clear which confounders were adjusted for and why they were included (b) Report category boundaries when continuous variables were categorized (c) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period	Pages 8 - 11		
Other analyses	17	Report other analyses done—e.g., analyses of subgroups and interactions, and sensitivity analyses	NA		
Discussion					
Key results	18	Summarise key results with reference to study objectives	Page 12		
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias	Page 13-14	RECORD 19.1: Discuss the implications of using data that were not created or collected to answer the specific research question(s). Include discussion of misclassification bias, unmeasured confounding, missing data, and changing eligibility over time, as they pertain to the study being reported.	Page 13-14
Interpretation	20	Give a cautious overall interpretation of results considering objectives,	Page 12		

		limitations, multiplicity of analyses, results from similar studies, and other relevant evidence			
Generalisability	21	Discuss the generalisability (external validity) of the study results	Page 13		
Other Information					
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based	Page 15		
Accessibility of protocol, raw data, and programming code		..		RECORD 22.1: Authors should provide information on how to access any supplemental information such as the study protocol, raw data, or programming code.	Page 8

*Reference: Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, Sørensen HT, von Elm E, Langan SM, the RECORD Working Committee. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. *PLoS Medicine* 2015; in press.