

BMJ Open is committed to open peer review. As part of this commitment we make the peer review history of every article we publish publicly available.

When an article is published we post the peer reviewers' comments and the authors' responses online. We also post the versions of the paper that were used during peer review. These are the versions that the peer review comments apply to.

The versions of the paper that follow are the versions that were submitted during the peer review process. They are not the versions of record or the final published versions. They should not be cited or distributed as the published version of this manuscript.

BMJ Open is an open access journal and the full, final, typeset and author-corrected version of record of the manuscript is available on our site with no access controls, subscription charges or pay-per-view fees (<u>http://bmjopen.bmj.com</u>).

If you have any questions on BMJ Open's open peer review process please email <u>info.bmjopen@bmj.com</u>

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 <u>licence</u>.

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 <u>Creative Commons</u> 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.

terez oni

Erasmushogeschool . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies



2		
3	1	Computerised Adaptive Testing Dramatically Reduces the Number of Items in
4	-	Detions Departed Llin and Knop Outcome Server
5	2	Patient Reported Hip and Knee Outcome Scores
0 7 8	3	An analysis of the UK National PROMs programme
9 10	4	
10 11 12	5	Jonathan P Evans ^{1,2} , Christopher Gibbons ³ , Professor Andrew Toms ² , Professor Jose Valderas ¹
13 14	6	
15	7	1. Health Services and Policy Research, Exeter Collaboration for Academic Primary Care (APEx),
16	8	University of Exeter, Magdalen Campus, Smeall Building, Room JS02, Exeter, EX1 2LU, UK
17	9	2. Princess Elizabeth Orthopaedic Centre, Royal Devon and Exeter Hospital, Exeter, EX2 5DW,
18	10	UK
19 20	11	3. Division of Internal Medicine, Department of Symptom Research, The University of Texas
20	12	MD Anderson Cancer Center, Houston, TX, USA
22 23	13	
24 25	14	Corresponding Author
26 27	15	Jonathan P Evans
28 29	16 17	Address: Health Services and Policy Research, Exeter Collaboration for Academic Primary Care (APEx) University of Exeter Magdalen Campus, Smeall Building, Boom ISO2, Exeter, EX1, 2111, 11K
30	10	Empil: in pypes2@pyptor ps.uk
31 32	10	
33	19	
34 35 26	20	Contributions
36 37 29	21 22	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
30 39	•	
40 41	23 24	outcome measures
42	25	A Toms MB ChB ERCS (Ed) MSc Eng ERCS (Tr & Orth) is a Consultant Trauma and Orthonaedic
43 44	26	Surgeon and Honorary Clinical Professor
45 46	27	JM Valderas MPH, PhD is Professor of health services and policy research
47	28	
48 49 50	29	
50 51	30	
52 53	31	
54 55	32	
56 57	33	
58 59 60		

1

Érasmushogeschool . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

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

- 37 Abstract
- 38 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.

44 Methods

45 We calibrated the 2018/2019 PROMs data to an item response theory (IRT) model. We assessed IRT 46 model assumptions alongside reliability. We used parameters from the IRT model with data from 47 2017/2018 to simulate CAT assessments. Two simulations were run until a prespecified standard 48 error of measurement was met (SE = .32 and SE = .45). We compared the number of questions 49 required to meet each cut-off and assessed the correlation between the full-length and CAT 50 administration.

51 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.

59 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.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Erasmushogeschool . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

Erasmushogeschool . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

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

80 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-

BMJ Open

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

BMJ Open

2		
3	122	of the PRON
4 5		
6	123	scale.
7		
8	124	The purpose
9 10		
10	125	confers a re
12		
13		
14	126	
15		
17		
18		
19		
20		
21		
22		
24		
25		
26		
27		
29		
30		
31		
32		
33 34		
35		
36		
37		
38 30		
40		
41		
42		
43		
44 45		
46		
47		
48		
49 50		
51		
52		
53		
54 55		
55 56		
57		
58		
59		
60		

1

124 The purpose of this study is to assess whether the application of IRT and CAT to the OHS and OKS

to per terien ont

125 confers a reduction in questionnaire burden whilst maintaining precision.

1 2		
3 4 5	127	Methods
6	128	Data
/ 8 9	129	The OHS was developed in 1996, and the OKS in 1998 ^{17,18} Each PROM contains 12 items that assess
10 11	130	joint-specific symptoms over the last four weeks. Each item has five response options that grade the
12 13	131	severity of symptoms and functional limitations. Developed following interviews with joint
14 15	132	replacement patients they were found to be the best performing condition-specific instruments
16 17 18	133	available in a standardised comparison of the measurement properties. ¹⁹
19 20	134	All Individual item level pre-operative OHS and OKS scores were extracted from the 2018/2019 data
21 22 23	135	release for hip and knee replacements for IRT model development. A second sample for simulation
23 24 25	136	of the CAT was extracted from the 2017/2018 data release. ⁴ Raw PROMs data are released annually
26 27	137	on the NHS digital platform following pre and postoperative linking, health record linking and
28 29	138	validation and data cleaning. ⁴
30 31 32 33	139	Development of the IRT model
34 35 36	140	We assessed the number of missing responses at the item level and presented them as a percentage
37 38	141	difference. We assessed the IRT assumptions of unidimensionality, local independence and
39 40	142	monotonicity. ²⁰ To confirm that all items measure a single underlying construct, we assessed
41 42	143	unidimensionality using confirmatory factor analysis (CFA). Model fit for CFA was assessed through
43 44 45	144	root mean square error of approximation (RMSEA) with a borderline model fit set at ≤0.08 and good
43 46 47	145	fit ≤0.06, and comparative fit index (CFI) and Tucker-Lewis index (TLI) with borderline model fit set to
48 49	146	>0.90 and good fit >0.95 (R package `LAVAAN` version 0.5–23.1097). ²¹ We confirmed the
50 51	147	dimensional structure of each scale using Mokken scaling and assessed scalability (monotonicity) of
52 53 54	148	the items. This assesses whether the probability of scoring the item along its scale of symptom
55 56	149	severity increases with a higher level of the underlying construct. A Loevinger's H value of ≥0.3 per
57 58 59 60	150	item was deemed acceptable (R package `Mokken` version 2.8.4). ²²

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml	

Page 8 of 32

Erasmushogeschool . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

BMJ Open

> 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 categorical²⁴, 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.

⁴⁵ 1

169 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

2	
2	
ر ۸	
4	
5	
6	
7	
8	
9	
10	
11	
11	
12	
13	
14	
15	
16	
17	
18	
19	
20	
∠∪ ⊃1	
21	
22	
23	
24	
25	
26	
27	
20	
20	
29	
30	
31	
32	
33	
34	
35	
36	
27	
27	
38	
39	
40	
41	
42	
43	
44	
45	
16	
-10 //7	
4/	
48	
49	
50	
51	
52	
53	
54	
54	
22	
56	
57	
58	
59	
60	

176	simulation can provide information on the number of items needed to provide estimates of the
177	latent trait at predetermined levels of precision. Through the simulation, the items that provide the
178	highest level of information, and thereby the greatest utility in shortened versions, can be
179	determined.
180	We performed a CAT simulation using Firestar for R (version 1.3.2). ²⁷ Two separate simulations were
181	conducted for OHS and OKS with the 2017/18 dataset with predetermined stopping criteria
182	(precision) denoted as a SE of the latent trait estimate of <0.32 and <0.45. These SE values are
183	equivalent to a reliability coefficient of 0.90 and 0.80 respectively. Variables derived from the
184	simulation include the correlation (Intraclass Correlation Coefficient, (ICC)) between the latent trait
185	estimation of the full-length questionnaire and the CAT, and the mean and standard deviation,
186	median and interquartile range (IQR) items required to derive estimates of the latent trait at the two
187	levels of precision. The items selected by the CAT were reported by their percentage of use within
188	the simulation. Differences in the item use between full-length and CAT administration is presented
189	as a percentage difference. Time-saving between full-length and CAT administration were calculated
190	against the estimate that each item takes between 10 seconds and 75 seconds per item to
191	complete. ²⁸
192	All data analysis was conducted in R (RStudio Team (2020). RStudio: Integrated Development for R.
193	RStudio, PBC, Boston, MA).
194	Public and Patient Involvement Statement
195	Formal patient and public involvement was not undertaken for this analysis of public domain data.
196	The national PROMs programme and the data held within have themselves been evaluated via
197	public consultation. Response to this evaluation from a multiple stakeholders taskforce highlighted
198	the need to improve efficiency of data collection. ⁶

Erasmushogeschool . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

199 200	Data sharing Difficulty and discrimination parameters of the IRT model for both the OHS and OKS are available in
201	Appendix 1. All data are available from NHS digital and can be used in accordance with the open
202	government licence for public sector information.
203	

2		
3 4 5	204	Results
6 7 8	205	Dataset characteristics
9 10	206	Of the 40,172 preoperative OHS scores and 44,264 OKS scores in the 2018/19 data, 1,704 were
11 12 13	207	revision hip replacements and 1,162 revision knee replacements, which were excluded. Further
14 15	208	exclusion of incomplete questionnaires resulted in 37,995 OHS and 42,558 OKS observations.
16 17	209	Missing responses to items were found 4118 (0.90%) and 4803 (0.93%) times for the OHS and OKS.
18 19	210	Although the first two items had few missing responses (0.15% OHS, 0.10% OKS), the remaining ten
20 21 22	211	had substantially more (1.04% OHS, 1.1% OKS), indicating a possible patient preference for shorter
23 24 25	212	measures.
25 26 27	213	For hip replacements, 59.4% were undertaken in females, 93.8% were \geq 50 years, and 51.9% were
28 29	214	≥70 years old. 14.8% of respondents had assistance completing the questionnaires, median
30 31	215	symptom duration was one to five years. For knee replacements, 56.5% were undertaken in females,
32 33 34	216	97.4% were ≥50 years and 51.8% were ≥70 years old. 14.6% had assistance completing their
35 36	217	questionnaires, median symptom duration was one to five years. All demographic features of the
37 38 30	218	PROMS dataset were equivalent to that of the full National Joint Registry.
40 41 42	219	IRT model assumptions
43 44	220	The criterion of unidimensionality was met at a borderline level for both OHS and OKS with an
45 46	221	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
47 48 49	222	(OKS). Mokken scaling corroborated this finding of unidimensionality and produced overall
50 51	223	scalability coefficients (H) of 0.49 (range 0.41 – 0.58) (OHS) and 0.46 (0.38 – 0.55) (OKS). Local
52 53	224	independence of items was confirmed for both OHS and OKS with all item correlations below 0.02.
54 55 56	225	There were no misfitting items within the GRM model.
57 58	226	Following production of the IRT item characteristic curves disordered thresholds (where the curve
59 60	227	lies under the line created by an adjacent curve) were noted in items 5, 6, 9, 10 and 12 for the OHS

Erasmushogeschool . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

BMJ Open

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)

233 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

2		
3 4	252	frequently, items 12 (5.1%) was minimally utilised leaving two items that were not required in any
5 6	253	simulations (including amongst them items 6, and 8 whose response options had been identified as
7 8 9	254	disordered) and all others were used in less than 1.4% of simulations (fig 4)(Appendix 1).
10 11 12	255	The items utilised most frequently in estimating the level of the latent trait within the OHS were
13 14	256	item 8 (During the past 4 weeks After a meal (sat at a table), how painful has it been for you to
15 16	257	stand up from a chair because of your hip?) and within the OKS were item 9 (During the past 4
17 18	258	weeks How much has pain from your knee interfered with your usual work (including housework)?)
19 20 21	259	and item 11 (During the past 4 weeks Could you do the household shopping on your own?).
22 23	260	Out of a potential 438,192 items for the OHS scores, only 145,462 items were used by the CAT at
24 25 26	261	0.32 SE, and 82,980 at 0.45 SE. This represents a 100.3% and 136.3% difference. Taking the whole
20 27 28	262	2018/19 NJR dataset before exclusions of 95,977 total hip replacements, at 0.45 SE this represents a
29 30	263	potential time saving of 2583 – 19374 hours for pre-operative scores. Out of a potential 541,464
31 32	264	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,
33 34	265	representing a 100.3% and 136.3% difference For the entire NJR dataset in 2018/19, at 0.45 SE this
35 36 37	266	represents a potential saving of between 2832.2 – 21241.5 hours for collection of pre-operative
38 30	267	scores.
40		
41 42	268	
43		
44 45		
46		
47 48		
49		
50 51		
52		
53 54		
55		
56 57		
57 58		
59		
60		

Erasmushogeschool . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

269	Discu	ISSION

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,

Page 15 of 32

BMJ Open

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

Erasmushogeschool . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

Erasmushogeschool . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

BMJ Open

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.

329 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.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

2		
2 3 4	338	Contribution
5	339	Authors JPE, CG and JMV devised the project. JPE managed the data and analysed the results with
6	340	oversight from CG and JMV. AT contributed to the data management and interpretation. JPE wrote
/ 8	341	the manuscript. CG, AT and JMV edited the manuscript. All authors read and approved the final
9	342	manuscript.
10 11	343	Ethics approval
12	344	Formal ethics approval was not required. All data used in accordance with the open government
13 1/1	345	licence for public sector information. The National Archives. Open Government Licence for public
15	346	sector information. 2020. http://www.nationalarchives.gov.uk/doc/open-government-
16	347	licence/version/3/ (accessed Dec 21, 2020).
17 18	348	Transparency statement
19 20	240	The lead author (Ionathan Evans) affirms that the manuscript is an honest, accurate, and
20 21 22	350	transparent account and that no important aspects of the study have been omitted.
23 24	351	Role of the funding source
25	352	Jonathan Evans is in receipt of an NIHR Academic Clinical Lecturer award. The views expressed are
26	353	those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health
27	354	and Social Care. There was no involvement of the funder in study design, data collection, data
28 20	355	analysis, manuscript preparation or publication decisions. All authors had complete access to the
30 31	356	study data that support the publication.
32	357	Conflict of Interest
33	358	Jonathan Evans, Christopher Gibbons, Andrew Toms, and Jose Valderas declare that they have no
34		
25	359	conflict of interest
35 36	359	conflict of interest
35 36 37	359 360	conflict of interest Licence
35 36 37 38	359 360 361	conflict of interest Licence "The Corresponding Author has the right to grant on behalf of all authors and does grant on behalf
35 36 37 38 39	359 360 361 362	conflict of interest Licence "The Corresponding Author has the right to grant on behalf of all authors and does grant on behalf of all authors, an exclusive licence (or non exclusive for government employees) on a worldwide
35 36 37 38 39 40 41	359 360 361 362 363	conflict of interest Licence "The Corresponding Author has the right to grant on behalf of all authors and does grant on behalf of all authors, an exclusive licence (or non exclusive for government employees) on a worldwide basis to the BMJ Publishing Group Ltd ("BMJ"), and its Licencees to permit this article (if accepted) to
35 36 37 38 39 40 41 42	359 360 361 362 363 364	conflict of interest Licence "The Corresponding Author has the right to grant on behalf of all authors and does grant on behalf of all authors, an exclusive licence (or non exclusive for government employees) on a worldwide basis to the BMJ Publishing Group Ltd ("BMJ"), and its Licencees to permit this article (if accepted) to be published in The BMI's editions and any other BMJ products and to exploit all subsidiary rights, as
35 36 37 38 39 40 41 42 43	359 360 361 362 363 364 365	conflict of interest Licence "The Corresponding Author has the right to grant on behalf of all authors and does grant on behalf of all authors, an exclusive licence (or non exclusive for government employees) on a worldwide basis to the BMJ Publishing Group Ltd ("BMJ"), and its Licencees to permit this article (if accepted) to be published in The BMJ's editions and any other BMJ products and to exploit all subsidiary rights, as set out in our licence."
35 36 37 38 39 40 41 42 43 44	359 360 361 362 363 364 365	conflict of interest Licence "The Corresponding Author has the right to grant on behalf of all authors and does grant on behalf of all authors, an exclusive licence (or non exclusive for government employees) on a worldwide basis to the BMJ Publishing Group Ltd ("BMJ"), and its Licencees to permit this article (if accepted) to be published in The BMJ's editions and any other BMJ products and to exploit all subsidiary rights, as set out in our licence."
35 36 37 38 39 40 41 42 43 44 45 46	359 360 361 362 363 364 365 366	conflict of interest Licence "The Corresponding Author has the right to grant on behalf of all authors and does grant on behalf of all authors, an exclusive licence (or non exclusive for government employees) on a worldwide basis to the BMJ Publishing Group Ltd ("BMJ"), and its Licencees to permit this article (if accepted) to be published in The BMJ's editions and any other BMJ products and to exploit all subsidiary rights, as set out in our licence."
35 36 37 38 39 40 41 42 43 44 45 46 47 48	359 360 361 362 363 364 365 366 366	conflict of interest Licence "The Corresponding Author has the right to grant on behalf of all authors and does grant on behalf of all authors, an exclusive licence (or non exclusive for government employees) on a worldwide basis to the BMJ Publishing Group Ltd ("BMJ"), and its Licencees to permit this article (if accepted) to be published in The BMJ's editions and any other BMJ products and to exploit all subsidiary rights, as set out in our licence."
35 36 37 38 39 40 41 42 43 44 45 46 47 48 49	359 360 361 362 363 364 365 366 366 367 368	conflict of interest Licence "The Corresponding Author has the right to grant on behalf of all authors and does grant on behalf of all authors, an exclusive licence (or non exclusive for government employees) on a worldwide basis to the BMJ Publishing Group Ltd ("BMJ"), and its Licencees to permit this article (if accepted) to be published in The BMJ's editions and any other BMJ products and to exploit all subsidiary rights, as set out in our licence." Figure Legends Figure 1: Item response theory (IRT) item traces for the 12- items of the Oxford Hip Score (OHS)(a)
35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51	359 360 361 362 363 364 365 366 366 367 368 369	conflict of interest Licence "The Corresponding Author has the right to grant on behalf of all authors and does grant on behalf of all authors, an exclusive licence (or non exclusive for government employees) on a worldwide basis to the BMJ Publishing Group Ltd ("BMJ"), and its Licencees to permit this article (if accepted) to be published in The BMJ's editions and any other BMJ products and to exploit all subsidiary rights, as set out in our licence." 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)
35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52	359 360 361 362 363 364 365 366 366 367 368 369 370	conflict of interest Licence "The Corresponding Author has the right to grant on behalf of all authors and does grant on behalf of all authors, an exclusive licence (or non exclusive for government employees) on a worldwide basis to the BMJ Publishing Group Ltd ("BMJ"), and its Licencees to permit this article (if accepted) to be published in The BMJ's editions and any other BMJ products and to exploit all subsidiary rights, as set out in our licence." 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)
35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53	359 360 361 362 363 364 365 366 366 367 368 369 370 371	conflict of interest Licence "The Corresponding Author has the right to grant on behalf of all authors and does grant on behalf of all authors, an exclusive licence (or non exclusive for government employees) on a worldwide basis to the BMJ Publishing Group Ltd ("BMJ"), and its Licencees to permit this article (if accepted) to be published in The BMJ's editions and any other BMJ products and to exploit all subsidiary rights, as set out in our licence." 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
35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55	359 360 361 362 363 364 365 366 367 368 369 370 371 372	conflict of interest Licence "The Corresponding Author has the right to grant on behalf of all authors and does grant on behalf of all authors, an exclusive licence (or non exclusive for government employees) on a worldwide basis to the BMJ Publishing Group Ltd ("BMJ"), and its Licencees to permit this article (if accepted) to be published in The BMJ's editions and any other BMJ products and to exploit all subsidiary rights, as set out in our licence." 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.
35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56	359 360 361 362 363 364 365 366 367 368 369 370 371 372	conflict of interest Licence "The Corresponding Author has the right to grant on behalf of all authors and does grant on behalf of all authors, an exclusive licence (or non exclusive for government employees) on a worldwide basis to the BMJ Publishing Group Ltd ("BMJ"), and its Licencees to permit this article (if accepted) to be published in The BMJ's editions and any other BMJ products and to exploit all subsidiary rights, as set out in our licence." 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.
35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60	359 360 361 362 363 364 365 366 367 368 369 370 371 372 373 374	 conflict of interest Licence "The Corresponding Author has the right to grant on behalf of all authors and does grant on behalf of all authors, an exclusive licence (or non exclusive for government employees) on a worldwide basis to the BMJ Publishing Group Ltd ("BMJ"), and its Licencees to permit this article (if accepted) to be published in The BMJ's editions and any other BMJ products and to exploit all subsidiary rights, as set out in our licence." 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).

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

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).

to oper teries only

BMJ Open

1 2			
3 4 5	377	Refe	rences
6 7 8	378	1	Black N, Burke L, Forrest CB, et al. Patient-reported outcomes: pathways to better health,
9 10	379		better services, and better societies. Qual Life Res 2016; 25: 1103–12.
11 12 13	380	2	Health & Social Care Information Centre. National PROMs programme
14 15 16	381	3	Wilson I, Bohm E, Lübbeke A, et al. Orthopaedic registries with patient-reported outcome
17 18	382		measures. EFORT Open Rev 2019; 4 : 357–67.
19 20 21	383	4	NHS Digital. Finalised PROMs data release. Patient Reported Outcome Measures (PROMs) in
22 23	384		England for Hip and Knee Replacement Procedures (April 2018 to March 2019). 2020.
24 25	385		https://digital.nhs.uk/data-and-information/publications/statistical/patient-reported-
26 27 28	386		outcome-measures-proms/hip-and-knee-replacement-procedures-april-2019-to-march-2020
29 30	387		(accessed Dec 21, 2020).
31 32 33	388	5	Hutchings A, Neuburger J, Grosse Frie K, Black N, van der Meulen J. Factors associated with
34 35	389		non-response in routine use of patient reported outcome measures after elective surgery in
36 37 38	390		England. Health Qual Life Outcomes 2012; 10: 34.
39 40	391	6	Kyte D, Cockwell P, Lencioni M, et al. Reflections on the national patient-reported outcome
41 42 42	392		measures (PROMs) programme: Where do we go from here? <i>J R Soc Med</i> 2016; 109 : 441–5.
43 44 45	393	7	Rowland C, Walsh L, Harrop R, Roy B, Skevington SM. What Do U.K. Orthopedic Surgery
46 47	394		Patients Think About PROMs? Evaluating the Evaluation and Explaining Missing Data. Qual
48 49 50	395		Health Res 2019; 29 : 2057–69.
51 52	396	8	Gorter R, Fox J-P, Twisk JWR. Why item response theory should be used for longitudinal
53 54 55	397		questionnaire data analysis in medical research Data analysis, statistics and modelling. BMC
56 57	398		Med Res Methodol 2015; 15 . DOI:10.1186/s12874-015-0050-x.
58 59 60	399	9	Cappelleri JC, Lundy JJ, Hays RD. Overview of classical test theory and item response theory

Page 20 of 32

Érasmushogeschool . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

BMJ Open

3 4	400		for the quantitative assessment of items in developing patient-reported outcomes measures.
5 6 7	401		<i>Clin Ther</i> 2014; 36 : 648–62.
8 9	402	10	Sijtsma K. On the use, the misuse, and the very limited usefulness of Cronbach's alpha.
10 11 12	403		Psychometrika 2009; 74 : 107.
13 14 15	404	11	Cella D, Gershon R, Lai J-S, Choi S. The future of outcomes measurement: Item banking,
15 16 17	405		tailored short-forms, and computerized adaptive assessment. Qual Life Res 2007; 16: 133–41.
18 19 20	406	12	Ko Y, Lo NN, Yeo SJ, et al. Comparison of the responsiveness of the SF-36, the Oxford Knee
20 21 22	407		Score, and the Knee Society Clinical Rating System in patients undergoing total knee
23 24	408		replacement. <i>Qual Life Res</i> 2013; 22 : 2455–9.
25 26 27	409	13	Ko Y, Lo NN, Yeo SJ, et al. Rasch analysis of the Oxford Knee Score. Osteoarthr Cartil 2009; 17 :
28 29	410		1163–9.
30 31 32	411	14	Norquist JM, Fitzpatrick R, Dawson J, Jenkinson C. Comparing alternative Rasch-based
33 34	412		methods vs raw scores in measuring change in health. Med Care 2004; 42.
35 36 37	413		DOI:10.1097/01.mlr.0000103530.13056.88.
38 39 40	414	15	Fitzpatrick R, Norquist JM, Jenkinson C, et al. A comparison of Rasch with Likert scoring to
40 41 42	415		discriminate between patients' evaluations of total hip replacement surgery. Qual Life Res
43 44	416		2004; 13 : 331–8.
45 46 47	417	16	Fitzpatrick R, Norquist JM, Jenkinson C, et al. A comparison of Rasch with Likert scoring to
48 49	418		discriminate between patients' evaluations of total hip replacement surgery. Qual Life Res
50 51 52	419		2004; 13 : 331–8.
53 54	420	17	Dawson J, Fitzpatrick R, Carr A, Murray D. Questionnaire on the perceptions of patients about
55 56 57	421		total hip replacement. <i>J Bone Joint Surg Br</i> 1996; 78 : 185–90.
58 59 60	422	18	Dawson J, Fitzpatrick R, Murray D, Carr A. Questionnaire on the perceptions of patients about

1			
2 3 4 5	423		total knee replacement. <i>J Bone Joint Surg Br</i> 1998; 80 : 63–9.
5 6 7	424	19	Harris K, Dawson J, Gibbons E, et al. Systematic review of measurement properties of patient-
8 9	425		reported outcome measures used in patients undergoing hip and knee arthroplasty. Patient
10 11 12	426		Relat Outcome Meas 2016; Volume 7: 101–8.
13 14	427	20	Cella D, Yount S, Rothrock N, et al. The Patient-Reported Outcomes Measurement
15 16 17	428		Information System (PROMIS): Progress of an NIH roadmap cooperative group during its first
18 19	429		two years. <i>Med Care</i> 2007; 45 : S3–11.
20 21	430	21	Rosseel Y. Lavaan: An R package for structural equation modeling and more. Version 0.5–12
22 23 24	431		(BETA). <i>J Stat Softw</i> 2012; 48 : 1–36.
25 26 27	432	22	Van der Ark LA. Mokken scale analysis in R. <i>J Stat Softw</i> 2007; 20 : 1–19.
28 29 30	433	23	Yen WM. Scaling Performance Assessments: Strategies for Managing Local Item Dependence.
31 32	434		J Educ Meas 1993; 30 : 187–213.
33 34 35	435	24	Hays RD, Morales LS, Reise SP. Item Response Theory and Health Outcomes Measurement in
36 37	436		the 21st Century NIH Public Access
38 39 40	437	25	Green BF, Bock RD, Humphreys LG, Linn RL, Reckase MD. Technical guidelines for assessing
41 42 42	438		computerized adaptive tests. <i>J Educ Meas</i> 1984; 21 : 347–60.
43 44 45	439	26	Gibbons CJ. Turning the page on pen-and-paper questionnaires: combining ecological
46 47	440		momentary assessment and computer adaptive testing to transform psychological
48 49 50	441		assessment in the 21st Century. Front Psychol 2017; 7: 1933.
51 52	442	27	Choi SW. Firestar: Computerized adaptive testing simulation program for polytomous item
53 54 55	443		response theory models. Appl Psychol Meas 2009; 33 : 644.
56 57	444	28	McMurray R, Heaton J, Sloper P, Nettleton S. Measurement of patient perceptions of pain
58 59 60	445		and disability in relation to total hip replacement: the place of the Oxford hip score in mixed

1 ว			
2 3 4	446		methods. <i>BMJ Qual Saf</i> 1999; 8 : 228–33.
5 6 7	447	29	Cook KF, O'Malley KJ, Roddey TS. Dynamic assessment of health outcomes: time to let the
8 9 10	448		CAT out of the bag? <i>Health Serv Res</i> 2005; 40 : 1694–711.
10 11 12	449	30	Gagnier JJ, Huang H, Mullins M, et al. Measurement properties of patient-reported outcome
13 14 15	450		measures used in patients undergoing total hip arthroplasty: A systematic review. JBJS Rev
16 17	451		2018; 6 . DOI:10.2106/JBJS.RVW.17.00038.
18 19 20	452	31	Gagnier JJ, Mullins M, Huang H, et al. A Systematic Review of Measurement Properties of
21 22	453		Patient-Reported Outcome Measures Used in Patients Undergoing Total Knee Arthroplasty. J.
23 24 25	454		Arthroplasty. 2017; 32 : 1688-1697.e7.
26 27	455	32	Reeve BB, Wyrwich KW, Wu AW, et al. ISOQOL recommends minimum standards for patient-
28 29 20	456		reported outcome measures used in patient-centered outcome1. Reeve BB, Wyrwich KW, Wu
30 31 32	457		AW, Velikova G, Terwee CB, Snyder CF, et al. ISOQOL recommends minimum standards for
33 34	458		patient-reported outcome measures us. <i>Qual Life Res</i> 2013; 22 : 1889–905.
35 36 37	459	33	Krosnic J, Presser S. Question and Questionnaire Design. In "Handbook of Survey Research",
38 39	460		2nd edn. Elsevier, 2013.
40 41 42	461	34	Wylde V, Learmonth ID, Cavendish VJ. The Oxford hip score: the patient's perspective. Health
43 44	462		Qual Life Outcomes 2005; 3 : 1–8.
45 46 47	463	35	Brodke DJ, Hung M, Bozic KJ. Item response theory and computerized adaptive testing for
48 49	464		orthopaedic outcomes measures. JAAOS-Journal Am Acad Orthop Surg 2016; 24: 750–4.
50 51 52	465	36	Porter I, Gonçalves-Bradley D, Ricci-Cabello I, et al. Framework and guidance for
53 54	466		implementing patient-reported outcomes in clinical practice: evidence, challenges and
55 56 57	467		opportunities. <i>J Comp Eff Res</i> 2016; 5 : 507–19.
58 59 60	468	37	Gandek B, Roos EM, Franklin PD, Ware JE. Item selection for 12-item short forms of the Knee

2			
4	469		injury and Osteoarthritis Outcome Score (KOOS-12) and Hip disability and Osteoarthritis
5 6 7	470		Outcome Score (HOOS-12). Osteoarthr Cartil 2019; 27 : 746–53.
8 9	471	38	Harris KK, Price AJ, Beard DJ, Fitzpatrick R, Jenkinson C, Dawson J. Can pain and function be
10 11 12	472		distinguished in the Oxford Hip Score in a meaningful way?: An exploratory and confirmatory
12 13 14	473		factor analysis. Bone Jt Res 2014; 3: 305–9.
15 16	474	39	Harris K, Dawson J, Doll H, et al. Can pain and function be distinguished in the Oxford Knee
17	475		Score in a meaningful way? An exploratory and confirmatory factor analysis. Qual Life Res
19 20 21	476		2013; 22 : 2561–8.
22 23 24	477		
25 26			
27			
20 29			
30 31			
32			
33 34			
35 36			
37			
38 39			
40			
41 42			
43			
44 45			
46			
47 48			
49			
50 51			
52			
53 54			
55			
56 57			
58			
59 60			

BMJ Open







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)

BMJ Open: first published as 10.1136/bmjopen-2021-059415 on 20 July 2022. Downloaded from http://bmjopen.bmj.com/ on June 8, 2025 at Department GEZ-LTA Erasmushogeschool . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.



Scatter plot and correlation between the theta estimation (values of the latent trait) between the full 12item 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)



Comparison of CAT vs. External Theta Estimates



Scatter plot and correlation between the theta estimation (values of the latent trait) between the full 12item 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 12item 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)



External Theta

Scatter plot and correlation between the theta estimation (values of the latent trait) between the full 12item 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)

Page 30 of 32

BMJ Open: first published as 10.1136/bmjopen-2021-059415 on 20 July 2022. Downloaded from http://bmjopen.bmj.com/ on June 8, 2025 at Department GEZ-LTA Erasmushogeschool . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

BMJ Open





242x129mm (96 x 96 DPI)



BMJ Open

Appendix 1

				in	02		
	Item	Difficulty a	Discrimination	Discriminatio	Discrimination	Discrimination	Item
			b1	b2 d i	G p3	b4	RMSEA
1	During the past 4 weeks How would you describe the pain you usually had from your hip?	1.914691405	0.164325049	2.206224 2 88	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	-1.052648662	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	-0.496468 5 03	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	0.125804	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	0.659403	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	0.719789 66 6 6	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	-0.433126 2 6 2	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	-0.056100 4600 ar	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	1.633812 0 8	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	0.289697 61000	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	0.415519 9 61	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	1.068727	2.493862719	NA	0.011

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

p://bmjopen.bmj.com/ on June 8, 2025 at Department GEZ-LTA , Al training, and similar technologies.

136/bmjopen-2 d by copyright, Page 33 of 32

Appendix 1

Divis Open

je 33 of 32		BMJ Open c br						
		Appendix 1			pyright, in	nionen-2002		
		ltem	Difficulty a	Discrimination	Discrimination #2	Discrimination	Discrimination	Item
			Dimounty a	b1	g	b3	b4	RMSEA
1	L	During the past 4 weeks How would you describe the pain you usually have from your knee?	1.683615138	0.035557101	2.32820	3.433951121	4.561327448	0.005
ź	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	-2.01845 7 992	-0.292460352	0.770869766	0.016
11	3	During the past 4 weeks Have you had any trouble getting in and out of a car or using public transport	1.932656761	-3.747375182	-0.968340445	0.945305979	2.030550642	0.007
	1	because of your knee? (whichever you would tend to use)	1 207015021	1.005070526		2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2		0.010
	+	During the past 4 weeks For now long have you been able to walk before pain from	1.387915921	-1.095979526	0.04808	2.23788967	NA	0.010
)	During the past 4 weeks After a meai (sat at a table), now paintui has it been for you to stand up from a	1.973493643	-2.4/354046/	-0.1105/5589	5 1.302091221	2.835473629	0.003
6	5	During the past 4 weeks Have you been limping when walking, because of your knee?	1.263415959	-0.270638346	1.27261	4.061594039	NA	0.011
	7	During the past 4 weeks Could you kneel down and get up again afterwards?	1.413075377	-0.361382764	1.15070502	2.902993876	4.420398561	0.018
8	3	During the past 4 weeks Have you been troubled by pain from your knee in bed at night?	1.23865998	-0.757873886	0.54707 67 469	2.065465542	NA	0.005
g)	During the past 4 weeks How much has pain from your knee interfered with your usual work (including housework)?	2.563072755	-1.375193192	0.17746 4 80 0	1.561747127	2.724240742	0.008
1	LO	During the past 4 weeks Have you felt that your knee might suddenly 'give way' or let you down?	1.507070288	-1.693738699	-0.25965 3 505	0.672066751	2.202410973	0.008
1	11	During the past 4 weeks Could you do the household shopping on your own?	2.235209642	-1.264960065	-0.48315	0.662939289	1.641645648	0.019
1	12	2 During the past 4 weeks Could you walk down one flight of stairs?	2.135163585	-2.14417636	-0.39877 4 015	1.11031124	2.341481914	0.011
		Table 2 : Oxford Knee Score items with associated IRT derived difficulty and	l discriminat	ion parameter	rs.	hmiopen hmi com/ on June 8 2025 at Departm		
						hent o		

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

//bmjopen.bmj.com/ on June 8, 2025 at Department GEZ-LTA

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

SCHOLARONE[™] Manuscripts



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 <u>licence</u>.

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 <u>Creative Commons</u> 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.

terez oni

Erasmushogeschool . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies



3 4	1	Use of computerised adaptive testing to reduce the number of items in
5	2	patient-reported hip and knee outcome scores: an analysis of the NHS
6	-	England national Dations Departed Outcome Macaurea programme
7	3	England national Patient Reported Outcome Measures programme
8 9	4	
10 11	5	Jonathan P Evans ^{1,2} , Christopher Gibbons ³ , Professor Andrew Toms ² , Professor Jose Valderas ^{1,4}
12 13	6	
14	7	1 Health Services and Policy Research, Exeter Collaboration for Academic Primary Care (APEx)
15	8	University of Exeter Magdalen Campus Smeall Building Room ISO2 Exeter EX1 2111 LIK
16	9	2 Princess Elizabeth Orthonaedic Centre Royal Devon and Exeter Hosnital Exeter EX2 5DW
17	10	
18	11	2 Division of Internal Medicine, Department of Symptom Research. The University of Texas
20 21	12	MD Anderson Cancer Center, Houston, TX, USA
22	13	4 Yong Loo Lin School of Medicine, National University of Singapore, 1F Kent Ridge Road,
23	14	NUHS Tower Block – Level 9 Singapore 119228
24	15	
25	16	
26	10	
27 28	17	Correspondence to:
29 30	18	Jonathan P Evans
31	19	Address: Health Services and Policy Research, Exeter Collaboration for Academic Primary Care
32 33	20	(APEx), University of Exeter, Magdalen Campus, Smeall Building, Room JS02, Exeter, EX1 2LU, UK
34 35	21	Email: j.p.evans2@exeter.ac.uk
36	22	
37		
38	23	Abstract
40 41	24	Objective
42	25	Over 160,000 participants per year complete the 12-item Oxford Hip and Knee Scores (OHS/OKS) as
43	26	part of the NHS England Patient Reported Outcome Measures (PROMs) programme. We used a
44	27	modern computational approach known as computerised adaptive testing (CAT) to simulate
45	27	inductric computational approach, known as computerised adaptive testing (CAT), to simulate
46	28	individually-tailored OHS and OKS assessment, with the goal of reducing the number of questions a
47 48	29	patient must complete without compromising measurement accuracy.
49 50	30	Methods
51	31	We calibrated the 2018/2019 PROMs data to an item response theory (IRT) model. We assessed IRT
52	32	model assumptions alongside reliability. We used parameters from the IRT model with data from
53	22	2017/2018 to simulate CAT assessments. Two simulations were run until a prospecified standard
54	22 24	2017/2010 to simulate CAT assessments. Two simulations were run until a prespective stalluaru
55	34	error of measurement was met (SE = .32 and SE = .45). We compared the number of questions
50 57	35	required to meet each cut-off and assessed the correlation between the full-length and CAT
58	36	administration.
59 60	37	Results

1 ว		
2	20	We conducted IPT analysis using 40.422 OHS and 44.714 OKS observations. The OHS and OKS were
4	20 20	both unidimensional (Post Moan Square Error of Approximation (PMSEA), 08 and 07 respectively)
5	39 40	and marginal reliability 01 and 00. The CAT, with a precision limit of SE = 22 and SE = 45 required a
6 7	40	and marginal reliability .91 and .90. The CAT, with a precision limit of SE $=$.52 and SE $=$.45 required a
8	41	median of 4 items (IQR 1) and 2 items (IQR 1) respectively for the OHS, and median of 4 items (IQR 2)
9	42	and 2 items (IQR 0) for the OKS. This represents a potential 82% reduction in PROM length. In the
10	43	context of 160,000 yearly assessments, these methodologies could result in the omission of some
11 12	44	1,280,000 redundant questions per year which equates to 40,000 hours of patient time.
12	45	Conclusion
14		
15	46	The application of IRT to the OHS and OKS produces an efficient and substantially reduced CAT. We
16 17	47	have demonstrated a path to reduce the burden and potentially increase the compliance for these
18	48	ubiquitous outcome measures without compromising measurement accuracy.
19	19	
20	75	
21	50	
23	51	
24	о- г	
25 26	52	Strengths and limitations of this study
27	53	- Our study is the first application of computerised adaptive testing on the worlds largest repository
28 29	54	of patient reported outcome measures.
30 31	55	- Over 35,000 responses were used in each modelling and simulation group.
32 33	56	- The Oxford Hip and Knee scores are very widely used at an international level.
34 35	57	- This secondary database analysis requires validation in a prospectively collected cohort.
36	58	- The available datasets are limited due to attrition that is attributed to the linking of patient
37	59	reported outcome measure data to health records.
38	L	
39 40	60	
40 41	61	
42		
43	~ ~	Introduction
44	62	Introduction
45 46		
47		
48	63	The ability to assess a patient's perspective about their health is central to holistic clinical decision
49 50	64	making medical research and health policy construction ¹ For hin and knee replacement surgery
50 51	04	making, medical research, and health policy construction. Tor hip and knee replacement surgery,
52	65	patients often complete questionnaires called patient reported outcome measures (PROMs) before
54	66	and after their expertion. Since 2000, ever 160,000 patients pervises underseins a his and his a
55	00	and after their operation. Since 2009, over 160,000 patients per year undergoing a hip or a Knee
56 57	67	replacement complete PROMs as part of the NHS England's PROMs Programme. ²
58		

Page 4 of 33

Erasmushogeschool . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

BMJ Open

> 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

BMJ Open

94	limitation of CTT is that it can only identify items that are not related to the construct being
95	measured, it does not identify items which are redundant (e.g., too similar to others) and can
96	incentivise the inclusion of redundant items. ¹⁰ This flexibility is leveraged by a computational
97	technique known as computerised adaptive testing (CAT). ¹¹ A CAT method iteratively select the most
98	informative and relevant items for a particular individual, thereby individualising the assessment to
99	the patient, often resulting in reducing assessment length whilst maintaining acceptable levels of
100	accuracy. Importantly, IRT analyses can be retrospectively applied to legacy PROMs that were
101	initially designed using CTT. Rasch analysis of the OHS and OKS has previously been undertaken, with
102	all studies demonstrating improvements in precision and group discrimination. ^{12–16} The development
103	of OHS/OKS CAT could improve the efficiency of administration and reduce the administrative
104	burden of the PROMs programme while offering the opportunity to implement a CAT at an
105	unprecedented scale.
106	The purpose of this study is to assess whether the application of IRT and CAT to the OHS and OKS
107	confers a reduction in questionnaire burden whilst maintaining precision.
108	

Erasmushogeschool . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

Erasmushogeschool . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

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

BMJ Open

2	
ך ע	
4	
5	
6	
7	
8	
9	
10	
11	
12	
13	
14	
15	
16	
10	
17	
18	
19	
20	
21	
22	
23	
24	
25	
26	
27	
28	
20	
29	
50 21	
31	
32	
33	
34	
35	
36	
37	
38	
39	
40	
41	
רד ⊿ר	
42	
45	
44	
45	
46	
47	
48	
49	
50	
51	
52	
53	
54	
55	
56	
50	
57	
20	
59	
60	

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

140 Following confirmation of IRT model fit assumptions a Graded Response Model (GRM), which is appropriate when item responses can be categorised as ordered categorical²⁴, was fit to the item 141 response data (R package `mirt` version 3.3.2). This model yields two-item parameters, the item 142 143 difficulty (a) which is a representation of the level of information about the underlying construct 144 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 145 thresholds, where the response category does not accord with the latent trait score, reordering of 146 147 adjacent response options was undertaken. Item and model fit was assessed using the RMSEA, TLI, 148 CFI and Standardized Root Mean Square Residual (SRMSR). Reliability in the IRT model was 149 estimated as marginal reliability where the overall reliability of the test was based on the average 150 conditional standard errors.²⁵ This overall index of precision can be compared to the classical 151 internal consistency (Cronbach's alpha) reliability estimate for CTT, where scores >0.8 indicate 152 excellent reliability.

0 153 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

Erasmushogeschool . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

> 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.

BMJ Open

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 12item questionnaire.28 All data analysis was conducted in R (RStudio Team (2020). RStudio: Integrated Development for R.

178 RStudio, PBC, Boston, MA).

179 Public and Patient Involvement

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

2		
4	182	public consultation. Response to this evaluation from a multiple stakeholders taskforce highlighted
5 6	183	the need to improve efficiency of data collection. ⁶
7		
8 9	184	Data availability statement
10 11	185	Difficulty and discrimination parameters of the IRT model for both the OHS and OKS are available in
12 13	186	Appendix 1. All data are available from NHS digital and can be used in accordance with the open
14 15 16	187	government licence for public sector information.
$\begin{array}{c} 15\\ 16\\ 17\\ 18\\ 19\\ 20\\ 21\\ 22\\ 23\\ 24\\ 25\\ 26\\ 27\\ 28\\ 29\\ 30\\ 31\\ 32\\ 33\\ 34\\ 35\\ 36\\ 37\\ 38\\ 9\\ 40\\ 41\\ 42\\ 43\\ 44\\ 56\\ 47\\ 48\\ 49\\ 50\\ 51\\ 52\\ 53\\ 54\\ 55\\ 56\end{array}$	187	government licence for public sector information.
57 58		
59 60		

2 3	
4	
5 6	
7 8	
9 10	
11	
12 13	
14 15	
16	
18	
19 20	
21 22	
23	
24 25	
26 27	
28 29	
30	
31 32	
33 34	
35 36	
37	
38 39	
40 41	
42 43	
44	
45 46	
47 48	
49 50	
51	
52 53	
54 55	
56 57	
58	
59 60	

188 Results

189 Dataset characteristics

190 Of the 40,172 preoperative OHS scores and 44,264 OKS scores in the 2018/19 data, 1,704 were 191 revision hip replacements and 1,162 revision knee replacements, which were excluded. Further 192 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. 193 194 Although the first two items had few missing responses (0.15% OHS, 0.10% OKS), the remaining ten 195 had substantially more (1.04% OHS, 1.1% OKS), indicating a possible patient preference for shorter 196 measures. 197 For hip replacements, 59.4% were undertaken in females, 93.8% were ≥50 years, and 51.9% were 198 ≥70 years old. 14.8% of respondents had assistance completing the questionnaires, median

199 symptom duration was one to five years. For knee replacements, 56.5% were undertaken in females,

200 97.4% were ≥50 years and 51.8% were ≥70 years old. 14.6% had assistance completing their

201 questionnaires, median symptom duration was one to five years. All demographic features of the

202 PROMS dataset were equivalent to that of the full National Joint Registry.

203 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 for the curve)
 Ites under the line created by an adjacent curve) were noted in items 5, 6, 9, 10 and 12 for the OHS

BMJ Open

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).

218 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

Erasmushogeschool . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

Erasmushogeschool . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies

BMJ Open

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.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

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,

1	
2	
4	258
5	250
6 7	259
8	260
9	
10 11	261
12	262
13	262
14 15	263
16	
17	264
18 10	• • •
20	265
21	266
22	200
25 24	267
25	
26 27	268
27	
29	269
30	
31	270
33	271
34 25	
35 36	272
37	
38	273
39 40	274
41	_/ .
42	275
43 44	
45	276
46	277
47 48	_,,
49	278
50	
51 52	279
53	280
54	200
55 56	281
57	
58	282
59 60	7 00
	203

261	and psychometrically robust. In this study, we have applied a modern psychometric approach to the
262	one of the world's largest repositories of orthopaedic arthroplasty PROMs. The OHS and OKS
263	conformed to IRT assumptions by demonstrating unidimensionality, monotonicity, and local
264	independence. CAT simulations demonstrated the possibility to dramatically reduce the length of
265	these 12-item PROMs to as little as two items at a high level of precision. In large-scale data
266	collection, the potential time saving from the deployment of a CAT is equivalent to more than a
267	million redundant questionnaire items per year and more than 4 years of collective patient time
268	annually.
269	Although Rasch analysis, a type of IRT methodology, has previously been applied to both the OHS
270	and OKS, ^{13,15} no research to date has explored the possibility of using these methods to reduce the
271	burden of assessment using CAT. CAT achieved the goal of minimising the burden of a PROM by only
272	delivering the most relevant and informative items required to measure a patient's level of hip or
273	knee pain and function. ²⁹ The simulations performed within this study were able to reduce the
274	number of items required by 67% for the OHS and 65% for the OKS at 0.32 SE (equivalent to 90%
275	precision) and as much as 81% for both PROMs at 0.45 SE (80% precision). As a comparator, the

For peer review on	y - http://bmjopen.bm	j.com/site/about/gui	delines.xhtml

were required to estimate patients' pain and function dramatically reducing the burden on a patient.

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)

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

of 0.82 - 0.94.^{30,31} Although this is excellent, delivery of the full test does not demonstrate

Erasmushogeschool . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

BMJ Open

> 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

BMJ Open

2	
3	
4	
5	
6	
7	
8	
9	
10	
11	
12	
13	
14	
15	
10	
1/	
10	
20 20	
20	
22	
23	
24	
25	
26	
27	
28	
29	
30	
31	
32	
33	
34	
35	
36	
3/	
38 20	
<u>79</u>	
40 41	
42	
43	
44	
45	
46	
47	
48	
49	
50	
51	
52	
53	
54	
55	
50	
5/ 50	
20	
22	

310 injury and Osteoarthritis Outcome Scores (HOOS and KOOS) that have also undergone IRT analysis,³⁷ 311 or contemporarily designed PROMs such as the PROMIS physical function and pain interference 312 scores. This attribute can have a profound effect on the translatability of research findings. Lastly, 313 IRT level analysis also opens up future assessment of differential item functioning (DIF). Here an 314 exploration of the extent to the item may be measing different abilities dependent on variables such 315 as age, gender, comorbidity profile or operation type could be undertaken. 316 The authors recognise limitations inherent to this study. We recognise that the dimensionality of 317 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 318 319 commonly applied scoring method utilises the total score for this very common PROM, it was 320 deemed appropriate to maintain a unidimensional model. The authors do recognise that by 321 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 322 to allow the presentation of IRT and CTT based scoring. We also recognise that the IRT parameters 323 324 were derived on pre-operative data, and therefore further analysis of post-operative data would be 325 required, of particular importance would be an assessment ceiling effect under this revised scoring 326 metric. We would recommend using the IRT deried parameters and the availability of the full 327 question bank in the post operative population, rather than a specifically reduced short-form 328 version. To improve the interpritability of the score, we would also recommend IRT derived minimal 329 important difference calculation for the OHS and OKS. By contextualising the differences in scorethat 330 would be deemed relevant to patients, this would inform the utility of this method in trial design and

331 as a potential adjunct to communication and decision making. The strength of this analysis is the 332 very large sample size for the IRT model construction. The CAT simulation requires validation on 333 patients with both qualitative and quantitative analysis of validity and acceptability. The significant

- 334 limitation to the practical application of IRT and CAT is the availability of a computer and an
- 335 appropriate interface, and we recognise that currently this national programme collects this data 60

Erasmushogeschool . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

BMJ Open

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.

342 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.

3 351 **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,

16 361 accurate, and transparent account and that no important aspects of the study have been omitted.

1718 362 Ethics approval

Formal ethics approval was not required. All data used in accordance with the open government
 364 licence for public sector information. The National Archives. Open Government Licence for public
 365 sector information. 2020. http://www.nationalarchives.gov.uk/doc/open-government-

23
 24 366 licence/version/3/ (accessed Dec 21, 2020).

²⁵ 367 **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.

34 373 Competing interests

3536 374 We declare that we have no competing interests.

38 375 Licence

The Corresponding Author has the right to grant on behalf of all authors and does grant on behalf of
all authors, an exclusive licence (or non exclusive for government employees) on a worldwide basis
to the BMJ Publishing Group Ltd ("BMJ"), and its Licencees to permit this article (if accepted) to be
published in The BMJ's editions and any other BMJ products and to exploit all subsidiary rights, as
set out in our licence.

Erasmushogeschool . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

2			
3 4 5	381	Refe	erences
6 7 8	382	1	Black N, Burke L, Forrest CB, et al. Patient-reported outcomes: pathways to better health,
9 10	383		better services, and better societies. Qual Life Res 2016; 25: 1103–12.
11 12 13	384	2	Health & Social Care Information Centre. National PROMs programme
14 15 16	385	3	Wilson I, Bohm E, Lübbeke A, et al. Orthopaedic registries with patient-reported outcome
17 18	386		measures. EFORT Open Rev 2019; 4 : 357–67.
20 21	387	4	NHS Digital. Finalised PROMs data release. Patient Reported Outcome Measures (PROMs) in
22 23	388		England for Hip and Knee Replacement Procedures (April 2018 to March 2019). 2020.
24 25	389		https://digital.nhs.uk/data-and-information/publications/statistical/patient-reported-
26 27 28	390		outcome-measures-proms/hip-and-knee-replacement-procedures-april-2019-to-march-2020
29 30	391		(accessed Dec 21, 2020).
31 32 33	392	5	Hutchings A, Neuburger J, Grosse Frie K, Black N, van der Meulen J. Factors associated with
34 35	393		non-response in routine use of patient reported outcome measures after elective surgery in
36 37 38	394		England. Health Qual Life Outcomes 2012; 10 : 34.
39 40	395	6	Kyte D, Cockwell P, Lencioni M, et al. Reflections on the national patient-reported outcome
41 42 43	396		measures (PROMs) programme: Where do we go from here? <i>J R Soc Med</i> 2016; 109 : 441–5.
44 45	397	7	Rowland C, Walsh L, Harrop R, Roy B, Skevington SM. What Do U.K. Orthopedic Surgery
46 47	398		Patients Think About PROMs? Evaluating the Evaluation and Explaining Missing Data. Qual
48 49 50	399		Health Res 2019; 29 : 2057–69.
51 52 53	400	8	Gorter R, Fox J-P, Twisk JWR. Why item response theory should be used for longitudinal
54 55	401		questionnaire data analysis in medical research Data analysis, statistics and modelling. BMC
56 57	402		Med Res Methodol 2015; 15 . DOI:10.1186/s12874-015-0050-x.

58 59

60

403 9 Cappelleri JC, Lundy JJ, Hays RD. Overview of classical test theory and item response theory

Page 19 of 33

1

2 3			
4	404		for the quantitative assessment of items in developing patient-reported outcomes measures.
5 6 7	405		<i>Clin Ther</i> 2014; 36 : 648–62.
8 9	406	10	Sijtsma K. On the use, the misuse, and the very limited usefulness of Cronbach's alpha.
10 11 12	407		Psychometrika 2009; 74 : 107.
13 14	408	11	Cella D, Gershon R, Lai J-S, Choi S. The future of outcomes measurement: Item banking,
16 17	409		tailored short-forms, and computerized adaptive assessment. <i>Qual Life Res</i> 2007; 16 : 133–41.
18 19 20	410	12	Ko Y, Lo NN, Yeo SJ, et al. Comparison of the responsiveness of the SF-36, the Oxford Knee
20 21 22	411		Score, and the Knee Society Clinical Rating System in patients undergoing total knee
23 24 25	412		replacement. <i>Qual Life Res</i> 2013; 22 : 2455–9.
26 27	413	13	Ko Y, Lo NN, Yeo SJ, et al. Rasch analysis of the Oxford Knee Score. Osteoarthr Cartil 2009; 17 :
28 29 30	414		1163–9.
31 32	415	14	Norquist JM, Fitzpatrick R, Dawson J, Jenkinson C. Comparing alternative Rasch-based
33 34 35	416		methods vs raw scores in measuring change in health. Med Care 2004; 42.
35 36 37	417		DOI:10.1097/01.mlr.0000103530.13056.88.
38 39	418	15	Fitzpatrick R, Norquist JM, Jenkinson C, et al. A comparison of Rasch with Likert scoring to
40 41 42	419		discriminate between patients' evaluations of total hip replacement surgery. Qual Life Res
43 44 45	420		2004; 13 : 331–8.
46 47	421	16	Fitzpatrick R, Norquist JM, Jenkinson C, et al. A comparison of Rasch with Likert scoring to
47 48 49 50 51 52	422		discriminate between patients' evaluations of total hip replacement surgery. Qual Life Res
	423		2004; 13 : 331–8.
53 54	424	17	Dawson J, Fitzpatrick R, Carr A, Murray D. Questionnaire on the perceptions of patients about
55 56 57	425		total hip replacement. J Bone Joint Surg Br 1996; 78: 185–90.
58 59 60	426	18	Dawson J, Fitzpatrick R, Murray D, Carr A. Questionnaire on the perceptions of patients about

Page 20 of 33

BMJ Open

2			
2 3 4 5	427		total knee replacement. <i>J Bone Joint Surg Br</i> 1998; 80 : 63–9.
5 6 7	428	19	Harris K, Dawson J, Gibbons E, et al. Systematic review of measurement properties of patient-
8 9	429		reported outcome measures used in patients undergoing hip and knee arthroplasty. Patient
10 11 12	430		Relat Outcome Meas 2016; Volume 7: 101–8.
13 14	431	20	Cella D, Yount S, Rothrock N, et al. The Patient-Reported Outcomes Measurement
15 16 17	432		Information System (PROMIS): Progress of an NIH roadmap cooperative group during its first
18 19 20	433		two years. <i>Med Care</i> 2007; 45 : S3–11.
20 21 22	434	21	Rosseel Y. Lavaan: An R package for structural equation modeling and more. Version 0.5–12
23 24	435		(BETA). J Stat Softw 2012; 48 : 1–36.
25 26 27	436	22	Van der Ark LA. Mokken scale analysis in R. <i>J Stat Softw</i> 2007; 20 : 1–19.
28 29 30	437	23	Yen WM. Scaling Performance Assessments: Strategies for Managing Local Item Dependence.
31 32 33	438		J Educ Meas 1993; 30 : 187–213.
34 35	439	24	Hays RD, Morales LS, Reise SP. Item Response Theory and Health Outcomes Measurement in
36 37 38	440		the 21st Century NIH Public Access
39 40	441	25	Green BF, Bock RD, Humphreys LG, Linn RL, Reckase MD. Technical guidelines for assessing
41 42 43	442		computerized adaptive tests. <i>J Educ Meas</i> 1984; 21 : 347–60.
44 45	443	26	Gibbons CJ. Turning the page on pen-and-paper questionnaires: combining ecological
46 47 48	444		momentary assessment and computer adaptive testing to transform psychological
48 49 50	445		assessment in the 21st Century. Front Psychol 2017; 7: 1933.
51 52 53	446	27	Choi SW. Firestar: Computerized adaptive testing simulation program for polytomous item
54 55	447		response theory models. Appl Psychol Meas 2009; 33 : 644.
50 57 58	448	28	McMurray R, Heaton J, Sloper P, Nettleton S. Measurement of patient perceptions of pain
59 60	449		and disability in relation to total hip replacement: the place of the Oxford hip score in mixed

BMJ Open

1 2			
2 3 4 5	450		methods. <i>BMJ Qual Saf</i> 1999; 8 : 228–33.
6 7	451	29	Cook KF, O'Malley KJ, Roddey TS. Dynamic assessment of health outcomes: time to let the
8 9 10	452		CAT out of the bag? <i>Health Serv Res</i> 2005; 40 : 1694–711.
10 11 12	453	30	Gagnier JJ, Huang H, Mullins M, et al. Measurement properties of patient-reported outcome
13 14	454		measures used in patients undergoing total hip arthroplasty: A systematic review. JBJS Rev
15 16 17	455		2018; 6 . DOI:10.2106/JBJS.RVW.17.00038.
18 19 20	456	31	Gagnier JJ, Mullins M, Huang H, et al. A Systematic Review of Measurement Properties of
20 21 22	457		Patient-Reported Outcome Measures Used in Patients Undergoing Total Knee Arthroplasty. J.
23 24	458		Arthroplasty. 2017; 32 : 1688-1697.e7.
25 26 27	459	32	Reeve BB, Wyrwich KW, Wu AW, et al. ISOQOL recommends minimum standards for patient-
28 29	460		reported outcome measures used in patient-centered outcome1. Reeve BB, Wyrwich KW, Wu
30 31	461		AW, Velikova G, Terwee CB, Snyder CF, et al. ISOQOL recommends minimum standards for
32 33 34	462		patient-reported outcome measures us. <i>Qual Life Res</i> 2013; 22 : 1889–905.
35 36	463	33	Krosnic J, Presser S. Question and Questionnaire Design. In "Handbook of Survey Research",
37 38 39	464		2nd edn. Elsevier, 2013.
40 41 42	465	34	Wylde V, Learmonth ID, Cavendish VJ. The Oxford hip score: the patient's perspective. <i>Health</i>
42 43 44	466		Qual Life Outcomes 2005; 3 : 1–8.
45 46 47	467	35	Brodke DJ, Hung M, Bozic KJ. Item response theory and computerized adaptive testing for
48 49	468		orthopaedic outcomes measures. JAAOS-Journal Am Acad Orthop Surg 2016; 24: 750–4.
50 51 52	469	36	Porter I, Gonçalves-Bradley D, Ricci-Cabello I, et al. Framework and guidance for
53 54	470		implementing patient-reported outcomes in clinical practice: evidence, challenges and
55 56 57	471		opportunities. <i>J Comp Eff Res</i> 2016; 5 : 507–19.
58 59 60	472	37	Gandek B, Roos EM, Franklin PD, Ware JE. Item selection for 12-item short forms of the Knee

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open

1

2		
3	473	injury and Osteoarthritis Outcome Score (KOOS-12) and Hip disability and Osteoarthritis
4		
5 6	474	Outcome Score (HOOS-12). Osteoarthr Cartil 2019; 27: 746–53.
7		
8	175	38 Harris KK Price AI Beard DI Eitzpatrick R Jenkinson C Dawson I Can pain and function be
9	475	so harns kk, thee As, beard bs, thepatrick k, senkinson e, bawsons, ear pair and function be
10	476	distinguished in the Oxford Hip Score in a meaningful way?: An exploratory and confirmatory
12	-	
13	477	factor analysis. <i>Bone Jt Res</i> 2014; 3 : 305–9.
14		
15	470	20 Harris K. Dawson L. Doll II. et al. Can pain and function he distinguished in the Oxford Knop
16 17	470	59 Harris K, Dawson J, Don H, <i>et di.</i> Can pain and function be distinguished in the Oxford Knee
18	479	Score in a meaningful way? An exploratory and confirmatory factor analysis. <i>Ougl Life Res</i>
19	175	
20	480	2013; 22 : 2561–8.
21		
22 23	401	
24	481	
25		
26	482	
27		
29	400	Figure Titles
30	483	Figure Titles
31	484	Figure 1: Item response theory (IRT) item traces for the 12 items of the Oxford Hip Score (OHS)(a)
32	485	and Oxford Knee Score (OKS)(b)
34	486	Figure 2: Scatter plot and correlation between the theta estimation (values of the latent trait)
35	487	between the full 12-item administration and the computerised adaptive test (CAT) for the Oxford
36	488	His Score (OHS) (a & b) and Oxford Knee Score (OKS) (c & d) at 0.32 standard error (SE) and 0.45 SE
37	180	Figure 3: Bar chart showing the number of items used per participant at 0.32 standard error (SE)
38 39	490	and 0.45 SE for the OHS (a, b) and OKS (c, d) computerised adaptive test (CAT)
40	150	
41	491	Figure 4: Bar chart showing the proportional use of each item at 0.35 standard error (SE) and 0.45
42	492	SE for the OHS (a, b) and OKS (c, d) computerised adaptive test (CAT)
43		
44 45		
46		
47		
48		
49 50		
51		
52		
53		
54 55		
56		
57		
58		
59 60		
00		

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml







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)



BMJ Open: first published as 10.1136/bmjopen-2021-059415 on 20 July 2022. Downloaded from http://bmjopen.bmj.com/ on June 8, 2025 at Department GEZ-LTA Erasmushogeschool . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

59

60

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)

Page	27 o	f 33 BMJ Oper	ı		136/bm d by cop			
1 2		Appendix 1			jopen-202 yright, in			
3 4		Item	Discrimination a	Difficulty b1	Diffeulty 102	Difficulty b3	Difficulty b4	ltem RMSEA
5	1	During the past 4 weeks How would you describe the pain you usually had from your hip?	1.914691405	0.164325049	2.2 622 888	3.209345224	4.186925039	0.011
6	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	15 - 1.05 264 2 662	0.609238754	1.637069089	0.003
7 8	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	0.435 0.435 0.435	1.340835438	2.431036075	0.014
9 10	4	During the past 4 weeks Have you been able to put on a pair of socks, stockings or tights?	1.547750826	- 1.475337619		1.50632625	2.924283288	0.009
11	5	During the past 4 weeks Could you do the household shopping on your own?	2.376360472	- 0.348672227	0.62 89 89 89 89 89 89 89 89 89 89	1.549282667	NA	0.008
12 13	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	0.749986766 ex100	2.086871457	NA	0.015
14 15	7	During the past 4 weeks Have you been able to climb a flight of stairs?	2.360767142	- 1.889430738	0.4 0.4 0.4 0.4 0.4 0.4 0.4 0.4 0.4 0.4	0.981278344	2.144165736	0.007
16	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	0.036465	1.205103327	2.706143566	0.012
1/	9	During the past 4 weeks Have you been limping when walking, because of your hip?	1.42746619	0.251525777	1.633812085	4.171461153	NA	0.007
18 19	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	0.22396972408	1.709666854	NA	0.011
20 21	11	During the past 4 weeks How much has pain from your hip interfered with your usual work (including housework)?	2.775690212	- 1.028278183	0.415516961	1.607031821	2.776328526	0.006
22	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 6 872 7 884	2.493862719	NA	0.011
23				•	3 2	•	· · · · · · · · · · · · · · · · · · ·	

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

ing, and similar technologies.

pen.bmj.com/ on June 8, 2025 at Department GEZ-LTA

 BMJ Open

Appendix 1

			ding	5941		
Item	Discrimination a	Difficulty b1	Difficulty b2 Q	Difficulty b3	Difficulty b4	ltem RMSEA
During the past 4 weeks How would you describe the pain you usually have from your knee?	1.683615138	0.035557101	2.32820	3.433951121	4.561327448	0.005
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
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		0.945305979	2.030550642	0.007
During the past 4 weeks For how long have you been able to walk before pain from	1.387915921	-1.095979526	0.64868 2 🖗	2.23788967	NA	0.010
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	-0.11057 × 80	1.302091221	2.835473629	0.003
During the past 4 weeks Have you been limping when walking, because of your knee?	1.263415959	-0.270638346	1.27261	4.061594039	NA	0.011
During the past 4 weeks Could you kneel down and get up again afterwards?	1.413075377	-0.361382764	1.15070 🔂 💆	2.902993876	4.420398561	0.018
During the past 4 weeks Have you been troubled by pain from your knee in bed at night?	1.23865998	-0.757873886	0.54707 1749	2.065465542	NA	0.005
During the past 4 weeks How much has pain from your knee interfered with your usual work (including housework)?	2.563072755	-1.375193192	0.17746 34 86 9	1.561747127	2.724240742	0.008
During the past 4 weeks Have you felt that your knee might suddenly 'give way' or let you down?	1.507070288	-1.693738699	-0.25965	0.672066751	2.202410973	0.008
During the past 4 weeks Could you do the household shopping on your own?	2.235209642	-1.264960065	-0.48315 ಶ 55	0.662939289	1.641645648	0.019
2 During the past 4 weeks Could you walk down one flight of stairs?	2.135163585	-2.14417636	-0.39877 # 015	1.11031124	2.341481914	0.011
			ining			

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

jopen.bmj.com/ on June 8, 2025 at Department GEZ-LTA ining, and similar technologies.

136/bmjopen-2021-0 d by copyright, includ

	Item No.	STROBE items	Location in manuscript where items are reported	RECORD items including for	Location in manuscript where items ar reported
Title and abstrac	t		D 2		D 2
	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 shame of the databases used should be included. RECORD 1.2: If applications included. RECORD 1.2: If applications included within which the study to be 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		1			
Background rationale	2	Explain the scientific background and rationale for the investigation being reported	Pages 3-4	similar tec	
Objectives	3	State specific objectives, including any prespecified hypotheses	Pages 4-5	Ine 8, 2028	
Methods				at	
Study Design	4	Present key elements of study design early in the paper	Pages 6-8	Depart	
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	Page 6	ment GEZ-L	

Participants	6	(a) Cohort study - Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up <i>Case-control study</i> - Give the eligibility criteria, and the sources and methods of case ascertainment and control	Page 6	RECORD 6.1: The methods of study population selection (such as codes or algorithms used to identify subjects) should be listed in details. possible, an explanation should be provided. RECORD 6.2: Any validation studies of the codes or algorithms used to	Page 6
		 selection. Give the rationale for the choice of cases and controls <i>Cross-sectional study</i> - 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 <i>Case-control study</i> - For matched studies, give matching criteria and the number of controls per case 	Pr rev	select the population should be referenced. If validation with the for this study and not published elsewhere, detailed methods and results should be provided. RECORD 6.3: If the study of volved linkage of databases, consider use of a flow diagram or other graphical display to demonstrate the data finkage process, including the number of individuals with linked that teach stage.	
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 lest of codes and algorithms used to chassefy exposures, outcomes, conformeders, 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	at Department GEZ-L1	

			ono open	1 by	
Bias	9	Describe any efforts to address potential sources of bias	NA	/ copyri	
Study size	10	Explain how the study size was arrived at	Page 6	en-202 ght, inc	
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen, and why	Pages 6 – 8	1-059415 on 20 د uding for uses	
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 <i>Case-control study</i> - If applicable, explain how matching of cases and controls was addressed <i>Cross-sectional study</i> - If applicable, describe analytical methods taking account of sampling strategy (e) Describe any sensitivity analyses 	Pages 6 – 8	July 2022. Downloaded from http://bmjopen.bmj.com/ on June 8, 2025 Erasmushogeschool . related to text and data mining, Al training, and similar technologies	
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

			BMJ Open	ed by	
				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, 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				ateo	
Participants	13	 (a) Report the numbers of individuals at each stage of the study (<i>e.g.</i>, 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 detail the selection of the persons belowed in the study (<i>i.e.</i> , study population delection) including filtering based be detail quality, data availability detailinkage. The selection of include 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 (<i>e.g.</i>, 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) <i>Cohort study</i> - summarise follow-up time (<i>e.g.</i>, average and total amount) 	Page 9	n.bmj.com/ on June 8, 2025 at Depar g, and similar technologies.	
Outcome data	15	Cohort study - Report numbersof outcome events or summarymeasures over timeCase-control study - Reportnumbers in each exposure	Page 9	tment GEZ-LTA	

e 33 of 33			BMJ Open	0.1136 cted by	
		category, or summary measures of exposure <i>Cross-sectional study</i> - Report numbers of outcome events or summary measures		/bmjopen-2021-0; copyright, incluc	
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	59415 on 20 July 2022. Downloaded from http: Erasmushogeschool . ding for uses related to text and data mining, /	
Other analyses	17	Report other analyses done— e.g., analyses of subgroups and interactions, and sensitivity analyses	NA	Al training, ar	
Discussion					
Key results	18	Summarise key results with reference to study objectives	Page 12	om/ on	
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 a swer 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	GEZ-LTA	

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

			BMJ Open	10.1136/ cted by		Page
		limitations, multiplicity of analyses, results from similar studies, and other relevant evidence		bmjopen-202 copyright, inc		
Generalisability	21	Discuss the generalisability (external validity) of the study results	Page 13	1-059415 (sluding fo		
Other Information	on			r ng		
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	20 July 2022. Do Erasmusho es related to te:		
Accessibility of protocol, raw data, and programming code			2	RECORD 22.1: Authors in the study protocol, raw datagos programming code.	d access such as r	Page 8
ommittee. The R n press. Checklist is protec	Eportin	g of studies Conducted using Observ	CC BY) license.	llected health Data (RECORD) Spen.bmj.com/ on June 8, 2025	itement. <i>PI</i>	LoS Medicine 2015
		For poor roviou only b	tu//bmionon.bmi.com/c	iat Department GEZ-LTA		
		For peer review only - ht	tp://bmjopen.bmj.com/s	ite/about/guidelines.xhtml		

age 34 of 33