ELSEVIER

Journal of Clinical Epidemiology ■ (2023) ■

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Item response theory assumptions were adequately met by the Oxford hip and knee scores

Conrad J. Harrison[a,*], Constantin Yves Plessen[b], Gregor Liegl[b], Jeremy N. Rodrigues[c,d], Shiraz A. Sabah[a], David J. Beard[a], Felix Fischer[b]

[a]*Surgical Intervention Trials Unit, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK*
[b]*Department of Psychosomatic Medicine, Center for Internal Medicine and Dermatology, Charité − Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin, Humboldt-Universität zu Berlin, Berlin Institute of Health, Berlin, Germany*
[c]*Clinical Trials Unit, University of Warwick, Coventry, UK*
[d]*Department of Plastic Surgery, Stoke Mandeville Hospital, Buckinghamshire Hospitals NHS Trust, Aylesbury, UK*

Accepted 19 April 2023; Published online xxxx

Abstract

Objectives: To develop item response theory (IRT) models for the Oxford hip and knee scores which convert patient responses into continuous scores with quantifiable precision and provide these as web applications for efficient score conversion.

Study Design and Setting: Data from the National Health Service patient-reported outcome measures program were used to test the assumptions of IRT (unidimensionality, monotonicity, local independence, and measurement invariance) before fitting models to preoperative response patterns obtained from patients undergoing primary elective hip or knee arthroplasty. The hip and knee datasets contained 321,147 and 355,249 patients, respectively.

Results: Scree plots, Kaiser criterion analyses, and confirmatory factor analyses confirmed unidimensionality and Mokken analysis confirmed monotonicity of both scales. In each scale, all item pairs shared a residual correlation of ≤ 0.20. At the test level, both scales showed measurement invariance by age and gender. Both scales provide precise measurement in preoperative settings but demonstrate poorer precision and ceiling effects in postoperative settings.

Conclusion: We provide IRT parameters and web applications that can convert Oxford Hip Score or Oxford Knee Score response sets into continuous measurements and quantify individual measurement error. These can be used in sensitivity analyses or to administer truncated and individualized computerized adaptive tests.

Keywords: Oxford hip score; Oxford knee score; Item response theory; Psychometrics; Validity; Arthroplasty

Funding: Conrad J. Harrison is funded by a National Institute for Health Research (NIHR) Doctoral Research Fellowship (NIHR300684) and the Rosetrees Clinical Trials Fellow. Jeremy N. Rodrigues is funded by a NIHR postdoctoral fellowship (PDF-2017-10-075). Shiraz A. Sabah is funded by a National Institute for Health Research (NIHR) Doctoral Research Fellowship (NIHR301771). This document presents independent research funded by the NIHR, the Rosetrees Trust, and the Oxford-Berlin Research Partnership. The views expressed are those of the authors and not necessarily those of the NHS, the Rosetrees Trust, the Oxford-Berlin Research Partnership, the NIHR, or the Department of Health and Social Care.

Declaration of interests: The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Conrad Harrison reports financial support was provided by National Institute for Health and Care Research. Conrad Harrison reports financial support was provided by Rosetrees Trust. Conrad Harrison reports financial support was provided by Oxford in Berlin. Shiraz Sabah reports financial support was provided by National Institute for Health and Care Research. Jeremy Rodrigues reports financial support was provided by National Institute for Health and Care Research. Conrad Harrison and Jeremy Rodrigues report a relationship with Methodology Oxford Ltd that includes board membership and consulting or advisory.

Author Contributions: C.H.: Conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, software, visualization, writing−original draft, and writing−review and editing. C.Y.P.: Project administration, methodology, resources, and writing−review and editing. G.L.: Project administration, methodology, resources, and writing−review and editing. J.R.: Funding acquisition, writing−review and editing, and supervision. S.A.S.: Data curation and writing−review and editing. D.J.B.: Funding acquisition, writing−review and editing, and supervision. F.F.: Funding acquisition, methodology, project administration, resources, software, writing−review and editing, and supervision.

* Corresponding author. The Botnar Research Centre, University of Oxford, Old Road, Oxford OX3 7LD, UK. Tel: +44-1865-227-374; fax: +44-1865-750-750.

E-mail address: Conrad.harrison@medsci.ox.ac.uk (C.J. Harrison).

https://doi.org/10.1016/j.jclinepi.2023.04.008

What is new?

Key findings

- We confirm modern test theory construct validity of the OHS and OKS in more than 670,000 patients.

What this adds to what was known?

- Our models produce continuous OHS and OKS scores and quantify potential measurement error.

What is the implication and what should change now?

- Researchers wishing to use item response theory parameters or scoring for the OHS and OKS may do so using this paper, or our supplementary web application.

1. Introduction

The Oxford Hip Score (OKS) and Oxford Knee Score (OKS) are widely used patient-reported outcome measures (PROMs). Both instruments have been used as primary outcome measures in high-profile randomized controlled trials [1,2], as clinical decision support tools [3], and as quality indicators in the UK National Health Service (NHS) PROMs program [4] and other arthroplasty registries [5]. The questionnaires were developed in 1996 (OHS) [6] and 1998 (OKS) [7] to measure the outcomes following hip and knee arthroplasty from the perspective of the patient. Each contains 12 equally weighted items with five response categories relating to severity and frequency of pain and disability (most items specifically attribute symptoms to the joint of interest). The recall period of both questionnaires is 4 weeks, and scores range from 0 to 48, with a higher value indicating a better clinical state.

The OHS and OKS were developed with classical test theory, a traditional psychometric approach which assumes a linear relationship between the observed score and the level of the underlying latent construct (hip or knee health) or true score. Although straightforward to apply, there are limitations to classical test theory [8,9]. First, all items in the scale (questionnaire) must usually be completed for valid score comparison. Second, measurement error is assumed to be constant across the measurement range and errors are assumed to cancel each other out on a population level. Third, although the scores derived by summing item responses are ordinal, they are typically treated as continuous and interval-scaled. In other words, the questions and their responses are treated as being weighted equally for analysis and interpretation despite this not necessarily being the case in the minds of patients as they answer them.

In recent years, there has been increasing interest in the application of item response theory (IRT), which uses probabilistic modelling to map specific response patterns (i.e., combinations of item responses) onto continuous scales [10,11]. This can provide more granular, continuous measurement with quantifiable uncertainty at the individual level. Metaphorically, this exchanges the ruler with large and unequally sized intervals for a ruler with many, tiny, equally sized intervals.

In IRT, all items function independently. This means that scores can be generated in the presence of missing responses, without imputation or exclusion. This can be applied deliberately, by only posing the most relevant items for an individual based on their responses to previous items. This is termed computerized adaptive testing (CAT) and can shorten and personalize assessments [12].

Researchers have previously attempted to fit OHS and OKS data to the Rasch model, a strict form of IRT model that assumes that the sum-score is a sufficient statistic for the latent score (which has interval scale properties) [13]. When this has been attempted, model fit to the unmodified questionnaires has been variable [14,15], and in some cases unconvincing [16]. Another approach is to use slightly more complex (and flexible) models, such as the graded response model (GRM) [17], to describe the relationship between item response patterns and latent constructs. This has been attempted in a recent paper which showed promising results, but there the authors used only a small proportion of available data to generate model parameters and made modifications to both questionnaires by collapsing several adjacent response options [18]. Models based on larger datasets and unmodified questionnaires may have more stable parameters, better generalizability, and leverage all available response options for more granular measurement [19].

Our first aim was to test the fit of NHS PROMs data to the GRM and establish IRT models that could derive continuous latent construct measurement from item response sets. Such models could be used by other researchers in future to quantify measurement error in clinical studies [manuscript under review with JCE] or to administer the OHS and OKS as computerized adaptive tests. If this was achieved, our second aim was to create a useable system where OHS/OKS response sets could be converted to this interval-scaled scoring. This would allow revised scoring of older datasets and might allow future datasets to benefit from improved scoring without needing specific psychometric programming experience in a study or clinical team. We planned to do this by creating an open-source web application.

2. Methods

All analyses were performed in R version 4.2.0. Code and data are available at: https://github.com/MrConradHarrison/IRT-modelling-for-the-OHS-and-OKS.

2.1. Data

We used publicly available NHS PROMs program data for this study. These were collected as part of a national audit across NHS England providers and include the demographics and PROM responses of patients undergoing elective primary hip or knee arthroplasty between April 1, 2012 and March 31, 2020. All patients undergoing hip or knee arthroplasty in NHS England are invited to complete the PROMs preoperatively and approximately 6 months postoperatively. This longitudinal, paired (preoperative and postoperative) dataset has been estimated to represent approximately 50% of procedures conducted during the period [20]. The data are deidentified, and ethics committee approval is not required for secondary analysis.

Hip and knee replacement procedures were assessed separately. We analyzed demographics and missing data patterns through descriptive statistics and excluded respondents with incomplete preoperative response sets list wise. We then used complete preoperative item response data to test the following key assumptions which underlie the IRT framework: unidimensionality, monotonicity, independence, and measurement invariance.

2.2. Unidimensionality

A set of items are described as unidimensional if they all measure the same, single, latent construct (or factor) in this case knee health or hip health. This is particularly relevant to the OHS and OKS, as other studies have suggested that they might each measure two correlated factors: pain and function [21,22]. If pain and function are experientially distinct constructs, positive changes in one could offset negative changes in the other when the scores of all items are combined. For example, a patient could experience improving function but worsening pain (two important changes), with a combined score that remains unchanged. The correlation between these factors has been estimated between 0.87 and 0.92 for the OKS [21] and 0.60 for the OHS [23].

For each PROM, we assessed unidimensionality using a scree plot, Kaiser criterion analysis (with a threshold of 1.0 eigenvalues) [24] and a confirmatory factor analysis (CFA) with polychoric correlation and a diagonally weighted least squares estimator in the *lavaan* package (version 0.6-11) [25]. The scree plot and Kaiser criterion analysis measure the variance in item responses explained by potential factors. The CFA tests how well our theoretical, unidimensional model explains the covariance in item response data. We used the following fit statistics and thresholds to indicate good model fit: root mean squared error of approximation < 0.06, standardized root means square residual ≤ 0.08, comparative fit index ≥ 0.95, and Tucker-Lewis index ≥ 0.95 [26].

2.3. Monotonicity

Monotonicity describes a nondecreasing relationship between item scores and latent construct levels. In other words, for any given item, if respondent x has a higher score than respondent y, the overall assessment score of respondent x must not be lower than that of respondent y. This can be assessed through Loevinger's H_i statistic, which compares the number of violations to this pattern (known as Guttman errors) to the number that would be expected in a set of unrelated items [27]. We took Loevinger's H_i values > 0.3 to indicate monotonicity [28].

2.4. Item independence

The local independence assumption states that two items are only related by the construct that they measure. We tested for this using Yen's Q3 residual correlation statistic, with a threshold of > 0.20 indicating undesirable local dependence between items [29]. A high residual correlation may suggest that the response to one item affects the response to the other, or that both items measure a second, unintended construct.

2.5. Measurement invariance

Measurement invariance describes a consistent relationship between item response patterns and latent construct levels across different population subgroups. For example, imagine an item that asks whether the respondent has difficulty using a toilet to urinate. For a given level of knee function, the response may differ between males and females as men may be more likely to stand up while urinating. In this case, the item would show differential item functioning (DIF) by gender.

We tested for DIF by gender (male vs. female) and age ($<$ 60 years or $\geq$ 60 years, as patients undergoing hip or knee arthroplasty less than the age of 60 years have substantially higher revision and dissatisfaction rates than those aged more than 60 years [30,31]). To do this, we used the logistic regression technique described by Choi et al. [32]. This method compares the fit of different logistic regression models that aim to predict item response based on the latent construct level. The addition of covariates (age or gender) should not improve model fit unless DIF exists. If the addition of a covariate (gender or age) improved the Nagelkerke pseudo-R^2 value of the model by $> 2\%$, we considered the item to exhibit DIF.

2.6. Graded response model

Using the *mirt* package (version 1.36.1) [33], we fitted GRMs to the complete preoperative item response sets in each dataset and used these to calculate IRT scores (specifically, expected a posteriori scores computed with a standard normal prior), for patients at both preoperative and postoperative time points. We compared these to test-

level and item-level information generated by the models, to illustrate how measurement precision varies with the level of hip or knee health.

We operationalized these models as an R Shiny web application that allows researchers to upload item response sets as a comma separated values (CSV) file, convert response sets to IRT scores, and download these together with the standard error of measurement for each respondent.

2.7. Cross-walk table

As an alternative to the response-pattern−specific IRT scores generated by the web application, we used the *mirt* package [33] to produce cross-walk tables that translate each of the 49 possible sum-scores on each instrument into expected a posteriori sum-scores and T-scores (mean 50, standard deviation 10), based on the GRMs. These serve as quick look-up tables to convert a (0-48) sum-score on either instrument into an IRT score. The expected a posteriori sum-score is the mean of each response-pattern−specific IRT score associated with a given sum-score [34]. For example, there are 12 possible response patterns that could achieve a sum-score of 1 on the OKS. Each of these response patterns is associated with its own response-pattern−specific IRT score (available through the web application). The expected a posteriori sum-score associated with the sum-score of 1 is the mean of these 12 response-pattern−specific IRT scores.

3. Results

3.1. Demographics, clinical characteristics, and ceiling effects

The demographics for each dataset are presented in Table 1. In both datasets, complete preoperative item response sets were available for 98.9% of individuals.

In respondents completing the OHS, $< 0.1\%$ achieved the ceiling score preoperatively, whereas 15.7% achieved the ceiling score postoperatively. In respondents completing the OKS, $< 0.1\%$ achieved the ceiling score preoperatively and 3.7% achieved the ceiling effect postoperatively.

Table 1. Sample sizes and demographics of the preoperative datasets used for item response theory analysis

	OHS	OKS
Sample size	321,147	355,249
Age band		
Under 60 years	42,145	34,536
Over 60 years	257,326	299,867
Missing	21,676	20,846
Gender		
Female	182,749	191,892
Male	116,648	142,404
Not specified	21,750	20,953

3.2. Unidimensionality

Scree plots, Kaiser criterion analyses suggested that both the OHS and OKS were unidimensional. This is illustrated in Figure 1.

The CFA provided further support for the assumption of unidimensionality, with both the OHS and OKS preoperative data demonstrating excellent fit to the one-factor model. The only fit statistic not to meet our prespecified threshold was the root mean squared error of approximation for the OHS (0.075, threshold < 0.060).

The fit statistics for each CFA are presented in Table 2, together with the thresholds that indicate good model fit. The results of CFA assumption tests and the models' standardized pattern coefficients are presented in the Supplementary Material.

3.3. Monotonicity

All items in each scale showed Loevinger's H_i statistics > 0.3, confirming monotonicity. These are presented with standard errors in the Supplementary Material.

3.4. Item independence

For the OHS, the Yen's Q3 residual correlation statistic between the items relating to 'washing' and 'dressing' was 0.20. For all other item pairs in the OHS, Yen's Q3 was $<$ 0.20. All item pairs in the OKS had a Yen's Q3 < 0.20.

3.5. Measurement invariance

The OHS items showed no DIF by age or gender. The OKS showed no DIF by age, but the item relating to 'kneeling' showed uniform DIF by gender, with an improvement in pseudo-R^2 of 6.17%. At any given latent construct level, men reported less difficulty kneeling down and getting up afterwards than women. When all items are administered together, the relationship between overall OKS score and latent construct level was very similar between genders (Fig. 2).

3.6. Graded response model

Having confirmed the assumptions of IRT, we fitted GRMs to both the OHS and OKS. These showed stable item parameters. Model parameters (together with 95% confidence intervals) are presented in Tables 3 and 4, and fit statistics are available in the Supplementary Material. Figure 3 demonstrates the relationship between sum-scores and scores derived from the IRT model.

The test-level information (which is closely related to measurement reliability and precision) was high across the ranges of the latent trait where most respondents are

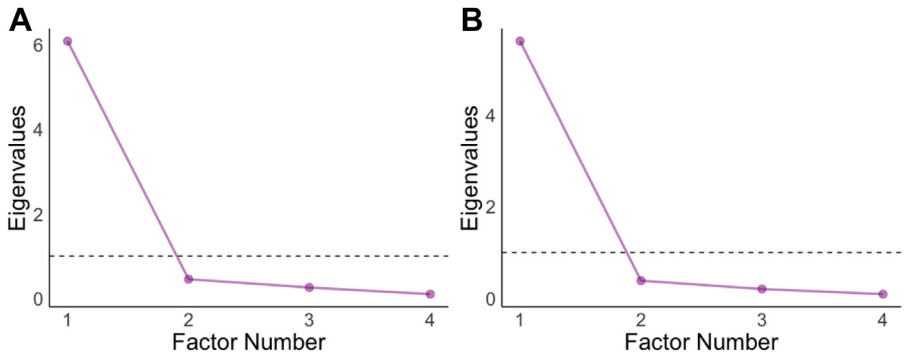

Fig. 1. Scree plots (minimum residual solution following oblimin rotation) for the Oxford Hip Score (panel A) and the Oxford Knee Score (panel B). The scree plots show a clear 'elbow' at the second factor, suggesting that most of the covariance in item responses is explained by the first factor. The horizontal dashed line shows the Kaiser criterion cutoff of 1 eigenvalue, with only the first factor accounting for more covariance than this limit. This strongly suggests unidimensionality.

located preoperatively. This means that overall both the OHS and OKS provide precise measurement in the preoperative setting. However, test-level information drops in the latent construct range where most respondents are located postoperatively. This means that particularly high OHS and OKS scores (demonstrated in the postoperative population) are less precise and less reliable than lower (i.e., preoperative) scores. This relationship is illustrated in Figure 4. A test-level information plot for the OHS is available in the Supplementary Material, along with item-level information plots for both scales. An information level $>$ 9.77 equates to a standard error of measurement $<$ 0.32, or a marginal reliability of $>$ 0.90, which is generally considered an excellent level of precision. An information level of 5.00 is equivalent to a marginal reliability of 0.80, which some consider acceptable for group-level measurement but not for individual-level measurement [35].

3.7. Web applications

The web application for converting item response data into IRT scores can be found at: https://conrad-harrison.shinyapps.io/IRTconverter/.

Users may upload item response data for either the OHS or OKS as a CSV file and convert these to continuous IRT scores. Data are not stored by the platform or viewable by other users.

Table 2. Confirmatory factor analysis results for a one-factor model

	RMSEA	SRMR	CFI	TLI
Threshold	<0.06	≤0.08	≥0.95	≥0.95
OHS	0.075 [0.075, 0.075]	0.051	0.990	0.987
OKS	0.060 [0.060, 0.060]	0.043	0.991	0.989

Abbreviations: OHS, Oxford Hip Score; OKS, Oxford Knee Score; RMSEA, root mean square error of approximation and 95% confidence intervals; SRMR, standardized root mean square residual; CFI, comparative fit index; TLI, Tucker-Lewis index.

The scores are presented as person location logits, which will range from approximately −4 to 4. Users may wish to scale these into other formats (e.g., to range from 0 to 100) [36] but it is usually reasonable to analyze logit scores without further scaling. Readers should be aware that scaling the logit scores into a continuous 0-48 format does not necessarily place them onto the same ordinal 0-48 scale achieved by summing the scores of each item.

Together with the IRT score, the web applications provide standard error of measurement values for each respondent. These can be interpreted as the standard deviation of plausible IRT scores that would result in the observed response set. In other words, 95% credible intervals can be presented for each score as *IRT score* $\pm$ 1.96 $\times$ *standard error of measurement* [37].

Missing data are handled directly by the IRT model. There is no need to impute or excluding missing item response data. In these cases, the score is measured from all available data and the uncertainty of the measurement is reflected in the standard error of measurement. Missing item responses can simply be left blank in the CSV file.

3.8. Cross-walk table

Table 5 is the cross-walk table for converting sum-scores into expected a posteriori sum-scores or T-scores. This can be used as a straightforward way to convert sum-scores to IRT scores, but provides less granular scoring than the response-pattern−specific scoring available through the web application.

4. Discussion

In this study, we found that the OHS and OKS fulfilled the assumptions of the GRM and developed models that allow specific response patterns to be mapped onto continuous scales. In future, the model parameters provided in this paper and our open-source web application can be used to:

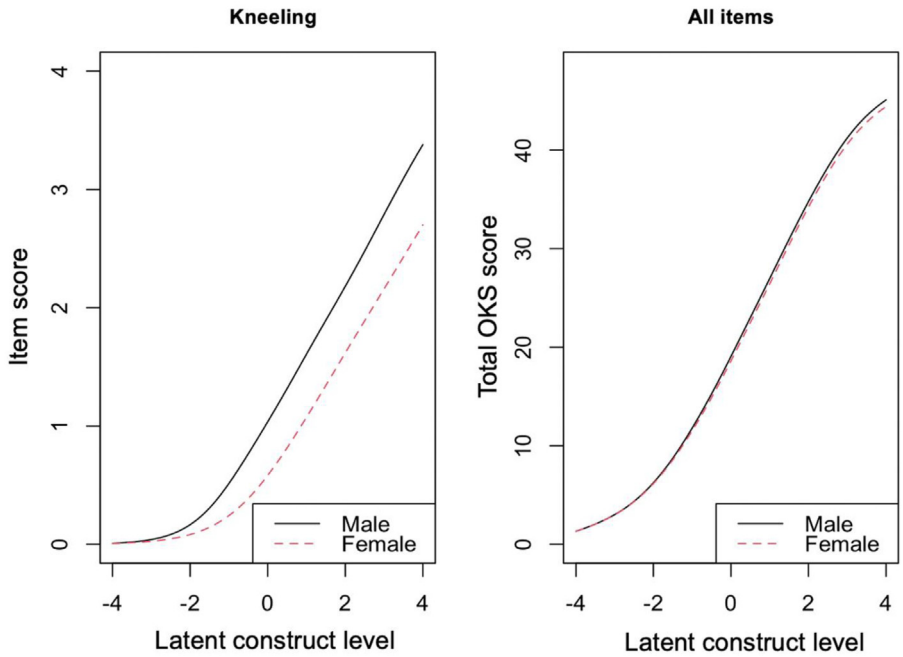

Fig. 2. Differential item functioning in the Oxford Knee Score. The left panel demonstrates the relationship between latent construct level (measured on a continuous logit scale) and the expected response to the 'kneeling' item. At any given latent construct level, men typically endorse a higher response (less trouble kneeling and getting up again) than women. The right panel demonstrates the relationship between latent construct level and total scale score. When all items are combined, there is no meaningful difference between the way men and women respond to the questionnaire.

- analyze OHS and OKS data with higher granularity, taking into account information on item characteristics;
- describe measurement precision at the individual level (e.g., for clinical decision support);
- quantify measurement error in clinical trials; and
- build computerized adaptive tests.

Although the classical test theory scoring of the OHS and OKS allow 49 different scores (including 0), our web applications will provide more than 244 million (5^{12}) different possible scores for each scale (or more than two billion when possible missing data patterns are included). This change in scoring is not necessarily sufficient to alter the conclusions of studies which use the OHS or OKS. Studies of other PROMs have failed to demonstrate superiority of IRT scoring to sum-scoring against external criteria [38,39], and in this study, we found a close correlation between EAP scores and sum-scores (Fig. 3). Nonetheless, this could be tested empirically in future. By demonstrating the agreement between IRT and classical test theory scoring, our study provides valuable reassurance that the foundation of previous research and policy remains sound, while also highlighting the potential benefits of using IRT in future studies.

Table 3. Item parameters for the Oxford Hip Score—graded response model: 95% confidence intervals are displayed in square brackets

Item	a	b_1	b_2	b_3	b_4
Walking	1.628 [1.617, 1.638]	−1.456 [−1.466, −1.447]	−0.683 [−0.690, −0.676]	0.697 [0.690, 0.704]	2.064 [2.051, 2.076]
Stairs	2.369 [2.354, 2.383]	−1.878 [−1.888, −1.868]	−0.472 [−0.478, −0.467]	0.928 [0.922, 0.935]	2.121 [2.110, 2.133]
Shopping	2.155 [2.142, 2.168]	−1.058 [−1.065, −1.051]	0.411 [−0.417, −0.406]	0.641 [0.635, 0.647]	1.594 [1.585, 1.603]
Dressing	1.550 [1.540, 1.560]	−1.522 [−1.532, −1.512]	0.048 [0.041, 0.054]	1.424 [1.414, 1.433]	2.867 [2.848, 2.886]
Transport	2.266 [2.251, 2.281]	−3.093 [−3.115, −3.071]	−0.555 [−0.561, −0.549]	1.305 [1.297, 1.312]	2.414 [2.400, 2.428]
Standing	2.212 [2.198, 2.225]	−2.184 [−2.196, −2.172]	−0.101 [−0.107, −0.096]	1.175 [1.168, 1.183]	2.716 [2.700, 2.733]
Work	2.761 [2.744, 2.778]	−1.085 [−1.092, −1.079]	0.354 [0.349, 0.360]	1.563 [1.555, 1.571]	2.716 [2.700, 2.733]
Sudden pain	1.272 [1.263, 1.281]	−0.970 [−0.980, −0.961]	0.239 [0.232, 0.246]	1.723 [1.711, 1.736]	2.278 [2.262, 2.295]
Limping	1.441 [1.430, 1.452]	0.184 [0.178, 0.191]	1.564 [1.552, 1.575]	2.344 [2.328, 2.361]	4.050 [4.014, 4.087]
Night Pain	1.224 [1.215, 1.233]	−0.290 [−0.298, 0.282]	0.981 [0.972, 0.991]	2.436 [2.418, 2.454]	3.147 [3.122, 3.171]
Pain	1.867 [1.854, 1.881]	0.106 [0.100, 0.112]	2.179 [2.166, 2.192]	3.190 [3.167, 3.214]	4.098 [4.055, 4.141]
Washing	1.711 [1.700, 1.721]	−3.009 [−3.029, −2.989]	−1.142 [−1.151, −1.134]	0.541 [0.534, 0.547]	1.586 [1.576, 1.596]

Table 4. Item parameters for the Oxford Knee Score—graded response model: 95% confidence intervals are displayed in square brackets

Item	a	b_1	b_2	b_3	b_4
Walking	1.330 [1.321, 1.339]	−1.920 [−1.932, −1.907]	−1.103 [−1.111, −1.094]	0.675 [0.667, 0.682]	2.323 [2.308, 2.339]
Standing	1.951 [1.939, 1.963]	−2.515 [−2.530, −2.501]	−0.118 [−0.123, −0.112]	1.299 [1.291, 1.306]	2.858 [2.840, 2.875]
Limping	1.280 [1.271, 1.290]	−0.299 [−0.306, −0.292]	1.256 [1.246, 1.266]	2.121 [2.106, 2.136]	4.010 [3.976, 4.043]
Kneeling	1.387 [1.377, 1.397]	−0.304 [−0.311, −0.298]	1.195 [1.186, 1.205]	2.949 [2.928, 2.969]	4.468 [4.427, 4.510]
Transport	1.887 [1.875, 1.899]	−3.756 [−3.787, −3.725]	−0.984 [−0.991, −0.977]	0.947 [0.940, 0.954]	2.050 [2.039, 2.062]
Work	2.548 [2.532, 2.563]	−1.406 [−1.413, −1.399]	0.154 [0.149, 0.159]	1.563 [1.556, 1.571]	2.712 [2.696, 2.728]
Stairs	2.108 [2.096, 2.121]	−2.119 [−2.130, −2.108]	−0.382 [−0.388, −0.377]	1.122 [1.115, 1.129]	2.383 [2.370, 2.396]
Give way	1.501 [1.491, 1.510]	−1.713 [−1.723, −1.702]	−0.270 [−0.276, −0.264]	0.651 [0.644, 0.657]	2.171 [2.158, 2.184]
Shopping	2.222 [2.209, 2.235]	−1.206 [−1.213, −1.199]	−0.443 [−0.449, −0.438]	0.690 [0.684, 0.695]	1.673 [1.664, 1.681]
Night Pain	1.221 [1.212, 1.230]	−0.811 [−0.819, −0.802]	0.484 [0.476, 0.491]	2.003 [1.989, 2.017]	2.624 [2.605, 2.642]
Pain	1.677 [1.665, 1.689]	0.019 [0.013, 0.025]	2.314 [2.300, 2.329]	3.415 [3.389, 3.441]	4.514 [4.464, 4.564]
Washing	1.475 [1.465, 1.485]	−4.283 [−4.321, −4.245]	−2.049 [−2.062, −2.036]	−0.327 [−0.334, −0.321]	0.734 [0.726, 0.741]

This work is potentially more impactful for individual-level scoring (e.g., when the OKS is used as a clinical decision aid on a per-patient basis [3]) than for group-level scoring, where positive and negative differences between classical test theory and IRT scoring (which are generally small, Fig. 3) are averaged out. Although it is possible that rescoring the OKS and OHS with IRT could alter between-group or within-group comparisons (such as those made in research studies), it is likely to have a bigger impact on between-patient or within-patient comparisons (such as those made in clinical practice). Using our web application, clinicians can now also estimate the potential measurement error around an individual's score (using the standard error of measurement or 95% credible intervals). This may be particularly useful for comparing repeated measures in an individual or for comparing an individual's score to those of other patients or clinically important thresholds [3].

The parameters we have presented could be used to build computerized adaptive tests that reduce the length of the OHS and OKS by selectively administering the most relevant items for an individual, based on the responses provided so far during the assessment [12]. CAT is most effective when used with large item banks, where it can provide more precise scoring than static short forms, in some cases from even fewer items [35,40]. However, recent research has shown that CAT can also reduce the length of PROM scales with similar lengths to the OHS and OKS [41–43] and this may be appealing in the context of clinical trials where several PROMs may be administered to respondents together. The publication of these IRT parameters complements previous efforts to reduce the burden of the OHS and OKS through CAT. These previous attempts have relied on either modifying the questionnaires [18] or the use of non-IRT techniques [44].

In the OKS, there was DIF by gender for the 'kneeling' item. But when all 12 items were combined, differential test functioning was negligible. This means overall scores for men and women can be interpreted in similar fashions.

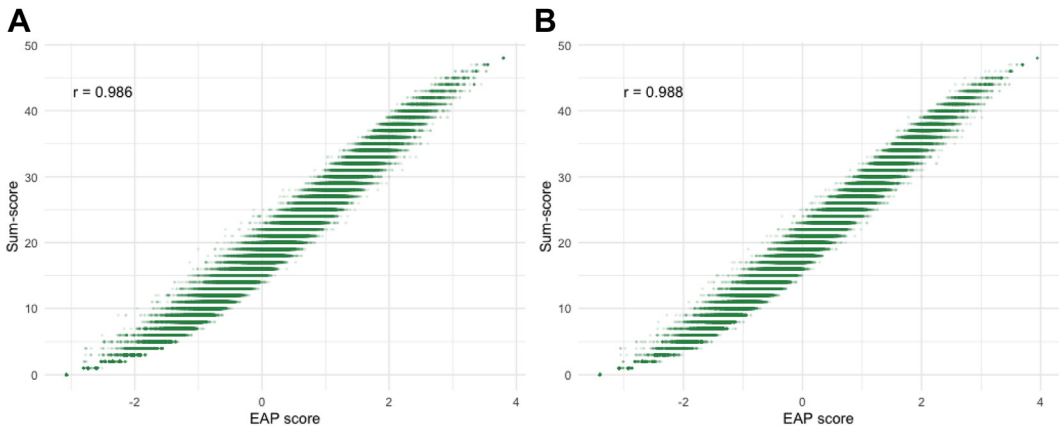

Fig. 3. Scatterplots comparing sum-scores and expected a posteriori (EAP) scores derived from the Oxford Hip Score (OHS, panel A) and Oxford Knee Score (OKS, panel B) graded response models. Pearson's correlation coefficient (r) is 0.986 for the OHS and 0.988 for the OKS. The points are arranged into 49 horizontal levels, representing the 49 possible sum-scores in each PROM. Overlying points appear more opaque.

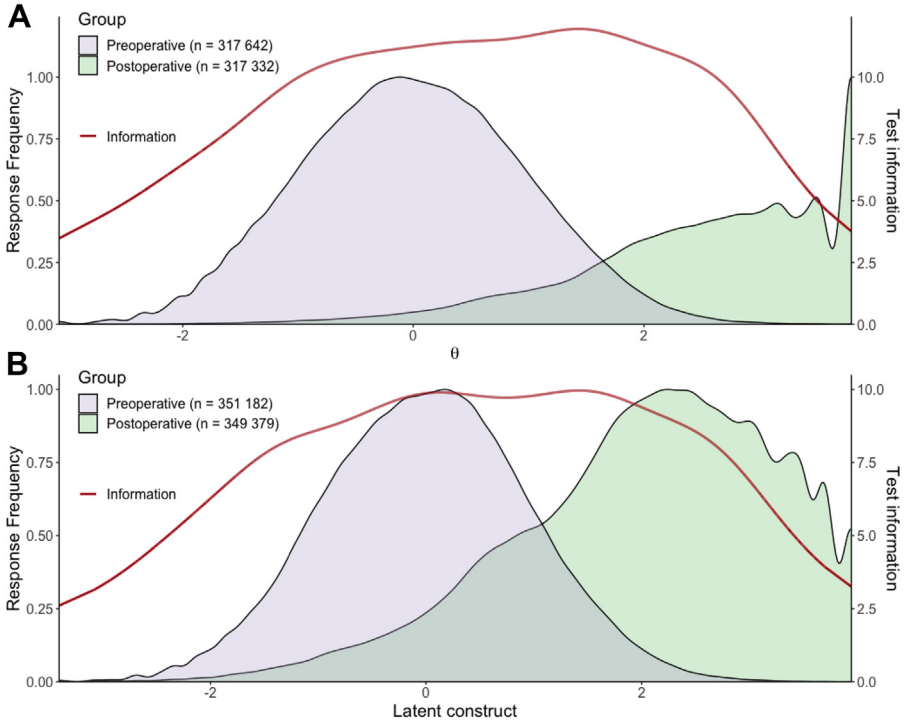

Fig. 4. Test-level information (precision) of the Oxford Hip Score (panel A) and Oxford Knee Score (panel B) across latent construct levels. The x-axis represents the latent construct (knee or hip health) measured on a continuous logit scale based on respondents' specific response patterns and the graded response model. The higher the latent trait level, the better the clinical state. The distribution of postoperative scores (shaded green) is higher (clinically better) than that of preoperative scores (shaded purple). The red line represents the level of information contained by the pattern of responses that achieve the latent construct score, which is closely related to the precision or reliability of the score. Scores at the extreme negative or positive ends of each scale provide less information for the model to calculate latent construct level. In other words, measurement is less precise at these levels. Test information is high for most preoperative response sets, but drops in the latent construct range where many postoperative respondents lie. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

However, the DIF in this item may be an important consideration for future computerized adaptive test development, as it could have a more significant impact in truncated assessments. If DIF is a concern, this item can simply be omitted when calculating IRT scores.

In recent years, there has been interest in the use of the OKS and OHS to measure pain and function constructs separately and in addition to a more global knee health construct [21,22]. This can be achieved by breaking each scale into two discrete subscales. Here, we have shown that the items in each instrument can be considered unidimensional (collectively measuring a single knee or hip health construct). One possible interpretation of this finding is that pain and function are closely related in osteoarthritis. This is consistent with previous factor analyses that showed cross-loading of items onto both pain and function constructs [21,22]. An alternative hypothesis might be that pain causes a reduction in function. The practical implication of this study for OKS and OHS users is that although pain and function can be measured individually, there may be little merit in doing so as these constructs are closely correlated. Using all items in each PROM together as a single scale will produce measurements with a lower standard error of measurement (higher precision) than will be achieved in individual pain and function subscales.

Although the OHS and OKS are well targeted (provide precise measurement) for preoperative patients, they both demonstrate ceiling effects postoperatively. This has been described previously [45,46], and in this study, we have demonstrated how this impacts on test-level information (and thus measurement precision). Respondents with high scores (e.g., an IRT score more than 2.5 logits or sum-score more than 40) will demonstrate higher standard errors of measurement (lower precision and reliability). In real-world terms, the Oxford scores would struggle to differentiate between someone who casually runs 5 km once a month and an elite athlete. This is more relevant for post-operative settings than preoperative settings. Although IRT may help to model the impact of ceiling effects on measurement precision, it does not resolve the ceiling effects themselves, which can be considered an issue with the content (wording) of the PROMs' items.

There are limitations to this work. Although we used very large datasets that produced stable model parameters, these models may not generalize to other populations (e.g., where significant cultural differences may affect the

Table 5. Expected a posteriori (EAP) sum-scores and T-scores (mean 50, SD 10) associated with each sum-score in the Oxford Hip and Knee Scores

Sum-score (either instrument)	EAP sum-score (Oxford Knee Score)	T-score (Oxford Knee Score)	EAP sum-score (Oxford Hip Score)	T-score (Oxford Hip Score)
0	−3.40	16	−3.07	19
1	−3.01	20	−2.66	23
2	−2.68	23	−2.34	27
3	−2.39	26	−2.07	29
4	−2.14	29	−1.84	32
5	−1.93	31	−1.64	33
6	−1.74	33	−1.47	35
7	−1.56	34	−1.30	37
8	−1.40	36	−1.15	39
9	−1.24	38	−1.01	40
10	−1.10	39	−0.88	41
11	−0.96	40	−0.75	43
12	−0.82	42	−0.62	44
13	−0.69	43	−0.50	45
14	−0.57	44	−0.38	46
15	−0.44	46	−0.26	47
16	−0.32	47	−0.15	49
17	−0.20	48	−0.03	50
18	−0.09	49	0.08	51
19	0.03	50	0.19	52
20	0.14	51	0.30	53
21	0.26	53	0.40	54
22	0.37	54	0.51	55
23	0.48	55	0.61	56
24	0.59	56	0.72	57
25	0.70	57	0.82	58
26	0.81	58	0.93	59
27	0.92	59	1.03	60
28	1.03	60	1.13	61
29	1.15	62	1.23	62
30	1.26	63	1.34	63
31	1.37	64	1.44	64
32	1.48	65	1.54	65
33	1.60	66	1.65	67
34	1.72	67	1.75	68
35	1.83	68	1.86	69
36	1.95	70	1.97	70
37	2.08	71	2.08	71
38	2.20	72	2.19	72
39	2.33	73	2.31	73
40	2.47	75	2.43	74
41	2.61	76	2.56	76
42	2.76	78	2.69	77
43	2.92	79	2.83	78
44	3.08	80	2.98	80
45	3.26	83	3.14	81

(*Continued*)

Table 5. Continued

Sum-score (either instrument)	EAP sum-score (Oxford Knee Score)	T-score (Oxford Knee Score)	EAP sum-score (Oxford Hip Score)	T-score (Oxford Hip Score)
46	3.46	85	3.32	83
47	3.67	87	3.53	85
48	3.94	89	3.80	88

This cross-walk table can be used as an alternative to the web application for converting sum-scores to either IRT scores or T-scores based on the IRT score. EAP sum-scores are the mean of EAP scores associated with any given sum-score. For example, there are 12 different possible response patterns that achieve a sum-score of 1, each with its own associated EAP score. The mean of these scores is the EAP sum-score.

relationship between latent construct level and item responses). In future, this could be examined through DIF analysis by country. All patients in this analysis were undergoing primary elective arthroplasty. Models may not generalize to very different conditions or treatments (e.g., major trauma or complex revision arthroplasty).

With our models, it is now possible to quantify measurement error in clinical trials that use the OHS or OKS, using techniques such as plausible value imputation [34]. Plausible value imputation is similar to multiple imputations, but instead of aiming to replace missing data, latent construct measurements for each respondent are randomly drawn from a distribution of plausible values. This distribution can be normally approximated with a mean equal to the expected a posteriori IRT score and a standard deviation equal to the standard error of measurement (both available through our web application). We have demonstrated this process in an accompanying paper (manuscript under review with JCE).

Future work should apply IRT scoring to OHS and OKS datasets. It will be particularly important to understand how this additional granularity affects the instruments' sensitivity and responsiveness and whether measurement error could have affected the results of landmark trials that have used these PROMs with classical test theory scoring. When doing this, trialists should be aware that interpretability statistics (such as minimal important difference and minimal important change) may vary with the scoring approach. Future work might also aim to define clinically important thresholds on this new, continuous, IRT scale.

Acknowledgments

We would like to acknowledge and thank Oxford-Berlin Research Partnership and Rosetrees Trust for supporting the Oxford-Berlin Partnership for Enhancing Measurement in Clinical Trials, as well as all patients who have contributed data to the NHS PROMs program and the open-source R community for making such software advances possible.

Supplementary Data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2023.04.008.

References

[1] Costa ML, Achten J, Parsons NR, Edlin RP, Foguet P, Prakash U, et al. Total hip arthroplasty versus resurfacing arthroplasty in the treatment of patients with arthritis of the hip joint: single centre, parallel group, assessor blinded, randomised controlled trial. BMJ 2012;344:e2147.

[2] Beard DJ, Davies LJ, Cook JA, MacLennan G, Price A, Kent S, et al. The clinical and cost-effectiveness of total versus partial knee replacement in patients with medial compartment osteoarthritis (TOPKAT): 5-year outcomes of a randomised controlled trial. Lancet 2019;394:746−56.

[3] Price A, Smith J, Dakin H, Kang S, Eibich P, Cook J, et al. The Arthroplasty Candidacy Help Engine tool to select candidates for hip and knee replacement surgery: development and economic modelling. Health Technol Assess 2019;23:1−216.

[4] NHS Digital. Patient reported outcome measures (PROMs). https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/patient-reported-outcome-measures-proms. Accessed October 3, 2021.

[5] Bohm ER, Kirby S, Trepman E, Hallstrom BR, Rolfson O, Wilkinson JM, et al. Collection and reporting of patient-reported outcome measures in arthroplasty registries: multinational survey and recommendations. Clin Orthop Relat Res 2021;479(10):2151−66.

[6] Dawson J, Fitzpatrick R, Carr A, Murray D. Questionnaire on the perceptions of patients about total hip replacement. J Bone Joint Surg Br 1996;78(2):185−90.

[7] Dawson J, Fitzpatrick R, Murray D, Carr A. Questionnaire on the perceptions of patients about total knee replacement. J Bone Joint Surg Br 1998;80-B(1):63−9.

[8] Rusch T, Lowry PB, Mair P, Treiblmaier H. Breaking free from the limitations of classical test theory: developing and measuring information systems scales using item response theory. Inf Manage 2017;54(2):189−203.

[9] Cappelleri JC, Jason Lundy J, Hays RD. Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. Clin Ther 2014;36:648−62.

[10] Harrison CJ, Sidey-Gibbons CJ. Modern psychometric measurement and computerized adaptive testing. In: Kassianos AP, editor. Handbook of Quality of Life in Cancer. Switzerland: Springer International Publishing; 2022:133−40.

[11] Edelen MO, Reeve BB. Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. Qual Life Res 2007;16:5−18.

[12] Harrison C, Loe BS, Lis P, Sidey-Gibbons C. Maximizing the potential of patient-reported assessments by using the open-source concerto platform with computerized adaptive testing and machine learning. J Med Internet Res 2020;22(10):e20950.

[13] Andrich D. Rating scales and Rasch measurement. Expert Rev Pharmacoecon Outcomes Res 2011;11(5):571−85.

[14] Conaghan PG, Emerton M, Tennant A. Internal construct validity of the Oxford knee scale: evidence from Rasch measurement. Arthritis Rheum 2007;57:1363−7.

[15] Fitzpatrick R, Norquist JM, Dawson J, Jenkinson C. Rasch scoring of outcomes of total hip replacement. J Clin Epidemiol 2003;56:68−74.

[16] Ko Y, Lo NN, Yeo SJ, Yang KY, Yeo W, Chong HC, et al. Rasch analysis of the Oxford knee score. Osteoarthritis Cartilage 2009;17:1163−9.

[17] Samejima F. Estimation of latent ability using a response pattern of graded scores. ETS Res Bull Ser 1968;1968(1):i−169.

[18] Evans JP, Gibbons C, Toms AD, Valderas JM. Use of computerised adaptive testing to reduce the number of items in patient-reported hip and knee outcome scores: an analysis of the NHS England National Patient-Reported Outcome Measures programme. BMJ Open 2022;12(7):e059415.

[19] García-Pérez MA. An analysis of (Dis)Ordered categories, thresholds, and crossings in difference and divide-by-total IRT models for ordered responses. Span J Psychol 2017;20:E10.

[20] Sabah SA, Alvand A, Beard DJ, Price AJ. Minimal important changes and differences were estimated for Oxford hip and knee scores following primary and revision arthroplasty. J Clin Epidemiol 2022;143:159−68.

[21] Harris K, Dawson J, Doll H, Field RE, Murray DW, Fitzpatrick R, et al. Can pain and function be distinguished in the Oxford Knee Score in a meaningful way? An exploratory and confirmatory factor analysis. Qual Life Res 2013;22:2561−8.

[22] Harris KK, Price AJ, Beard DJ, Fitzpatrick R, Jenkinson C, Dawson J. Can pain and function be distinguished in the Oxford Hip Score in a meaningful way?: an exploratory and confirmatory factor analysis. Bone Joint Res 2014;3(11):305−9.

[23] Norquist JM, Fitzpatrick R, Dawson J, Jenkinson C. Comparing alternative rasch-based methods vs raw scores in measuring change in health. Med Care 2004;42:25.

[24] Kanyongo GY. The influence of reliability on four rules for determining the number of components to retain. J Mod App Stat Meth 2005;5(2):332−43.

[25] Rosseel Y. Lavaan: an R package for structural equation modeling. J Stat Soft 2012;48(2):1−36.

[26] Schreiber JB, Nora A, Stage FK, Barlow EA, King J. Reporting structural equation modeling and confirmatory factor analysis results: a review. J Educ Res 2006;99(6):323−38.

[27] van Schuur WH. Mokken scale analysis: between the guttman scale and parametric item response theory. Polit Anal 2003;11(2):139−63.

[28] Sijtsma K, Molenaar I. Introduction to Nonparametric Item Response Theory. New Delhi, India: SAGE Publications, Inc.; 2002.

[29] Christensen KB, Makransky G, Horton M. Critical values for Yen's Q3: identification of local dependence in the Rasch model using residual correlations. Appl Psychol Meas 2017;41(3):178−94.

[30] Canovas F, Dagneaux L. Quality of life after total knee arthroplasty. Orthop Traumatol Surg Res 2018;104(1):S41−6.

[31] Bayliss LE, Culliford D, Monk AP, Glyn-Jones S, Prieto-Alhambra D, Judge A, et al. The effect of patient age at intervention on risk of implant revision after total replacement of the hip or knee: a population-based cohort study. Lancet 2017;389:1424−30.

[32] Choi SW, Gibbons LE, Crane PK. Lordif: an R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. J Stat Softw 2011;39(8):1−30.

[33] Chalmers RP. Mirt: a multidimensional item response theory package for the R environment. J Stat Soft 2012;48(6):1−29.

[34] Fischer HF, Rose M. Scoring depression on a common metric: a comparison of EAP estimation, plausible value imputation, and full bayesian IRT modeling. Multivariate Behav Res 2019;54(1):85−99.

[35] Gibbons C, Bower P, Lovell K, Valderas J, Skevington S. Electronic quality of life assessment using computer-adaptive testing. J Med Internet Res 2016;18(9):e240.

[36] Chapman R. Expected a posteriori scoring in PROMIS®. J Patient Rep Outcomes 2022;6(1):59.

[37] Martin M, Kosinski M, Bjorner JB, Ware JE, MacLean R, Li T. Item response theory methods can improve the measurement of physical function by combining the modified health assessment questionnaire and the SF-36 physical function scale. Qual Life Res 2007;16:647.

[38] Petersen MA, Groenvold M, Aaronson N, Brenne E, Fayers P, Nielsen JD, et al. Scoring based on item response theory did not alter the measurement ability of EORTC QLQ-C30 scales. J Clin Epidemiol 2005;58:902−8.

[39] Fischer F, Levis B, Falk C, Sun Y, Ioannidis JPA, Cuijpers P, et al. Comparison of different scoring methods based on latent variable models of the PHQ-9: an individual participant data meta-analysis. Psychol Med 2022;52:3472−83.

[40] Harrison C, Clelland AD, Davis TRC, Scammell BE, Zhang W, Russell P, et al. A comparative analysis of multidimensional computerized adaptive testing for the DASH and QuickDASH scores in Dupuytren's disease. J Hand Surg Eur 2022;47(7):750−4.

[41] Harrison CJ, Geerards D, Ottenhof MJ, Klassen AF, Riff KWYW, Swan MC, et al. Computerised adaptive testing accurately predicts CLEFT-Q scores by selecting fewer, more patient-focused questions. J Plast Reconstr Aesthet Surg 2019;72(11):1819−24.

[42] Ottenhof MJ, Geerards D, Harrison C, Klassen AF, Hoogbergen MM, van der Hulst RRWJ, et al. Applying computerized adaptive testing to the FACE-Q skin cancer module: individualizing patient-reported outcome measures in facial surgery. Plast Reconstr Surg 2021;148(4):863−9.

[43] Kamran R, Rodrigues JN, Dobbs TD, Wormald JCR, Trickett RW, Harrison CJ. Computerized adaptive testing of symptom severity: a registry-based study of 924 patients with trapeziometacarpal arthritis. J Hand Surg Eur 2022;47(9):893−8.

[44] Harrison CJ, Plummer OR, Dawson J, Jenkinson C, Hunt A, Rodrigues JN. Computerized adaptive testing for the Oxford Hip, Knee, Shoulder, and Elbow scores: accurate measurement from fewer, and more patient-focused, questions. Bone Joint Open 2022;3(10):786−94.

[45] Edwards TC, Guest B, Garner A, Logishetty K, Liddle AD, Cobb JP. The metabolic equivalent of task score: a useful metric for comparing high-functioning hip arthroplasty patients. Bone Joint Res 2022;11(5):317−26.

[46] Clement ND, Afzal I, Demetriou C, Deehan DJ, Field RE, Kader DF. The preoperative Oxford Knee Score is an independent predictor of achieving a postoperative ceiling score after total knee arthroplasty. Bone Joint J 2020;102-B(11):1519−26.