Journal of Clinical Epidemiology 158 (2023) 62–69

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Item response theory may account for unequal item weighting and individual-level measurement error in trials that use PROMs: a psychometric sensitivity analysis of the TOPKAT trial

Conrad J. Harrison[a,*], Constantin Yves Plessen[b], Gregor Liegl[b], Jeremy N. Rodrigues[c], Shiraz A. Sabah[a], Jonathan A. Cook[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, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Center for Internal Medicine and Dermatology, Charité — Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin, Berlin, Germany*
[c]*Clinical Trials Unit, University of Warwick, Coventry, UK*

Accepted 14 March 2023; Published online 24 March 2023

Abstract

Objectives: To apply item response theory as a framework for studying measurement error in superiority trials which use patient-reported outcome measures (PROMs).

Methods: We reanalyzed data from the The Total or Partial Knee Arthroplasty Trial, which compared the Oxford Knee Score (OKS) responses of patients undergoing partial or total knee replacement, using traditional sum-scoring, after accounting for OKS item characteristics with expected a posteriori (EAP) scoring, and after accounting for individual-level measurement error with plausible value imputation (PVI). We compared the marginalized mean scores of each group at baseline, 2 months, and yearly for 5 years. We used registry data to estimate the minimal important difference (MID) of OKS scores with sum-scoring and EAP scoring.

Results: With sum-scoring, we found statistically significant differences in mean OKS score at 2 months ($P = 0.030$) and 1 year ($P = 0.030$). EAP scores produced slightly different results, with statistically significant differences at 1 year ($P = 0.041$) and 3 years ($P = 0.043$). With PVI, there were no statistically significant differences.

Conclusion: Psychometric sensitivity analyses can be readily performed for superiority trials using PROMs and may aid the interpretation of results.

Keywords: Oxford knee score; Item response theory; Psychometrics; EAP; Plausible value imputation; Arthroplasty

1. Introduction

Pragmatic randomized controlled trials (RCTs) provide high-quality evidence that aligns closely to real-world practice. They are important for health service policy development [1] in initiatives like the National Institute for Health and Care Excellence (NICE) Technology Appraisal Programme in the United Kingdom, and the Comparative Effectiveness Research Initiative in the United

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 reports a relationship with Methodology Oxford Ltd that includes: board membership and consulting or advisory.

CRediT authorship contribution statement: CH: conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, software, visualization, writing—original draft, writing—review & editing. CYP: project administration, methodology, resources, writing—review & editing. GL: project administration, methodology, resources, writing—review & editing. R: funding acquisition, writing—review & editing, supervision. SAS: data curation, writing—review & editing. JAC: writing—review & editing, supervision. DJB: funding acquisition, writing—review & editing, supervision. FF: funding acquisition, methodology, project administration, resources, software, writing—review & editing, supervision.

* Corresponding author: The Botnar Research Centre, University of Oxford, Old Road, Oxford, OX3 7LD. Tel: +44-0-1865-227-374; fax: +44-0-1865-750-750.

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

https://doi.org/10.1016/j.jclinepi.2023.03.013
0895-4356/

What is new?

Key findings

- We propose item response theory as a framework for studying measurement error in RCTs that use PROMs.
- We demonstrate its use in a psychometric sensitivity analysis of the The Total or Partial Knee Arthroplasty Trial.
- This produced slightly different results to the original trial.

What this adds to what was known?

- Psychometric sensitivity analysis may have a role in demonstrating the robustness of trial findings.

What is the implication and what should change now?

- Future work is needed to establish the potential benefits of psychometric sensitivity analyses in clinical trials.

States [2]. Patient-reported outcome measures (PROMs) are often used as the primary endpoint in pragmatic RCTs. These instruments aim to measure health from the perspective of the patient [3–5]. However, interpretation of these instruments in RCTs and observational studies may be challenging.

Whilst trialists commonly account for uncertainty in the sampling of participants or the effect of a range of bias parameters, two specific characteristics of PROMs that may be associated with further uncertainty are often not addressed directly.

1. Ordinal PROM sum scores are often treated as continuous data. In the simplest unweighted summation form, it is implicitly assumed that the response option for each item is interval-scaled and contributes equally to the latent construct measurement. However, patients may not weight items equally when they complete PROMs. Furthermore, one aspect of interest may have a disproportionate number of associated items, leading to implicit weighting.
2. The scores obtained from PROMs will contain some measurement error. This is where a PROM response does not perfectly represent a patient's true state. Most trial analyses will aggregate scores within trial arms and accept that some portion of the observed score variance is caused by measurement error.

Item response theory (IRT) provides a framework for exploring the impact of individual item characteristics, and accounting for measurement error at the level of the individual. This may be useful for addressing these areas of uncertainty.

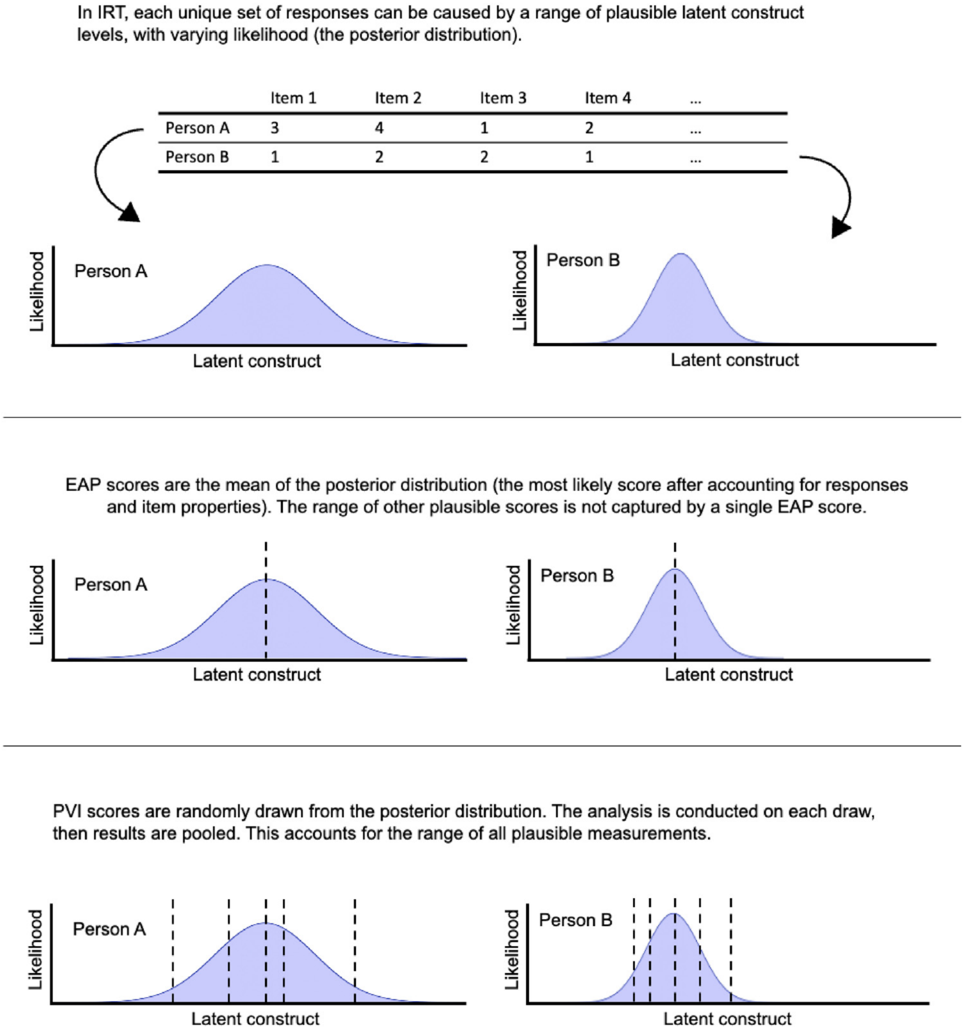
Item response theory models the relationship between item responses and the underlying latent construct [6,7]. In IRT, a given response pattern across a set of items can be caused by a distribution of latent construct levels, with varying likelihood. For example, there is a distribution of potential levels of knee function that might have given rise to a particular combination of item responses in a knee function PROM. For the purposes of trial analysis, a single point in this distribution is usually reported as the person's level of knee function. A popular approach to this, used by modern PROM systems such as the Patient-Reported Outcomes Measurement Information System (PROMIS), is expected a posteriori (EAP) scoring [8]. The EAP score is the weighted mean of all plausible latent construct measurements (Fig. 1).

Expected a posteriori scoring accounts for unequal weighting of items and their response options and provides continuous, rather than ordinal or interval, measurement (two respondents with a given sum score may have different EAP scores if their response pattern differs). But EAP scoring does not account for differing levels of measurement precision. Some measurements will be made with a high level of precision (the observed response pattern is unlikely to arise from different knee function levels) and other measurements will be made with a low level of precision (the observed response pattern could be explained by many other knee function levels). Measures typically exhibit high measurement precision at medium construct levels and low measurement precision at the extremes. Plausible value imputation (PVI) is a technique that accounts for these differences in measurement precision [6,9].

In PVI, multiple datasets are generated where latent construct (for example, knee function) measurements are randomly drawn from the distribution of plausible values (Fig. 1). The statistical analysis is then performed on each dataset independently, and results pooled. This is similar in principle to multiple imputation for missing data, but here we treat the latent construct level as "missing."

Simulation studies have suggested that IRT techniques such as EAP scoring and PVI might have a role in identifying and mitigating bias that is introduced to trial analyses when a PROM does not function as expected [10,11]. For example, when the assumption of equal item weighting is not supported by the observed measurement properties of the instrument, or when individual-level measurement error is high.

In this paper, we illustrate the use of EAP and PVI for psychometric sensitivity analyses by comparing these techniques to traditional sum-scoring using data from a landmark RCT that compared two types of knee replacement (the The Total or Partial Knee Arthroplasty Trial [TOPKAT] study) [3].



In IRT, each unique set of responses can be caused by a range of plausible latent construct levels, with varying likelihood (the posterior distribution).

	Item 1	Item 2	Item 3	Item 4	...
Person A	3	4	1	2	...
Person B	1	2	2	1	...

EAP scores are the mean of the posterior distribution (the most likely score after accounting for responses and item properties). The range of other plausible scores is not captured by a single EAP score.

PVI scores are randomly drawn from the posterior distribution. The analysis is conducted on each draw, then results are pooled. This accounts for the range of all plausible measurements.

Fig. 1. Illustration of expected a posteriori (EAP) scoring and plausible value imputation (PVI) in item response theory (IRT).

2. Methods

2.1. The TOPKAT trial

The TOPKAT trial, published in 2019, was a pragmatic, multicentre, superiority RCT that compared the clinical effectiveness of total vs. partial knee replacement in patients with medial compartment osteoarthritis [3]. Between January 18, 2010, and September 30, 2013, a total of 528 patients across 27 UK sites were randomly allocated in a 1:1 ratio to either receiving a partial or total knee replacement. The primary outcome measure was the Oxford Knee Score (OKS), calculated through sum-scoring, 5 years after randomization. The trial was powered to detect a 2.0 point minimal important difference (MID) in OKS sum score with 80% power at a 5% significance level, assuming a standard deviation of 10.0 points. The trial found no significant difference in OKS sum-score between arms (adjusted mean difference of 1.04 in favor of partial knee replacement, 95% confidence interval (CI) -0.42 to 2.50, $P = 0.159$), but recommended partial knee replacement, mainly for health economic reasons.

2.2. The Oxford Knee Score

The OKS was developed in 1998 specifically to measure knee replacement outcomes [12]. It contains 12 items, each with 5 response options (all scored 0 to 4). The official total sum-scores range from 0 to 48, with a higher score reflecting a better clinical state.

While the OKS was initially developed with sum-scoring, (the approach officially endorsed by the instrument's developers), an IRT model (specifically, a graded response model) describing the relationship between item responses and the latent construct (*knee health*) has recently been developed, based on the responses of over 350,000 patients undergoing elective primary knee replacement included in the NHS PROMs registry [manuscript under review with *JCE*]. The model development study demonstrated strong evidence that the assumptions of IRT (unidimensionality, monotonicity, measurement invariance, and local independence) were met by the OKS when used to measure knee replacement outcomes in the NHS. For any given OKS response pattern (with or without missing data), EAP scores and plausible latent construct distributions can now be generated with an online conversion tool [manuscript under review with *JCE*].

2.3. Sum-score analysis

First, we analyzed OKS responses in the TOPKAT trial using conventional methods. To do this, we used a mixed-effects linear model. The dependent variable was the OKS sum-score. The fixed effects were gender, age, and an interaction term comprising treatment allocation and time point. Patients' study numbers were included as random effects (random intercepts). Based on this model, marginal means were estimated for the OKS sum-score in both groups at baseline, 2 months, and 1, 2, 3, 4, and 5 years. We compared marginalized means pairwise by timepoint using the *emmeans R* package (version 1.7.5), and a Tukey adjustment for multiple comparisons [13], to identify any statistically significant differences in OKS sum-score between trial arms over the follow-up period. This differs from the original TOPKAT analysis, which compared OKS scores at 5 years, without a multiplicity adjustment.

The principal analysis of the original TOPKAT study controlled for baseline OKS score. In our EAP and PVI analyses (described below) we have altered the scoring of the OKS, and so to ensure a fair comparison, we did not include baseline OKS score in any of our models.

2.4. Expected a posteriori analysis

Next, we repeated the analysis using EAP scoring, based on the scale's IRT model and a standard normal prior. We generated these scores using the *mirt* package (version 1.36.1) in *R* [14]. Item response theory models produce EAP scores on a continuous logit (z-score) scale, which has no theoretical upper or lower limit, but is limited empirically by the minimally and maximally scoring OKS response sets [15]. The area of this logit scale measured by the OKS ranges from approximately −4 to 4. It is reasonable to analyse these logit scores without further scaling, but for the comparison with sum-scoring, we performed a linear transformation to align the mean and standard deviation of baseline sum-scores and EAP scores (across both treatment arms). This took the form

$$\widehat{\theta}^* = \frac{\widehat{\theta}}{\sigma} + \overline{x},$$

where $\widehat{\theta}^*$ is the transformed EAP score, $\widehat{\theta}$ is the untransformed EAP score (the z-score), σ is the standard deviation of all baseline sum-scores, and $\overline{x}$ is the mean of all baseline sum-scores. We made no changes to the mixed-effects model, other than replacing sum-scores with scaled EAP scores.

2.5. Plausible value imputation analysis

We then repeated the analysis using PVI. For each patient, at each time point, we randomly drew 25 plausible latent construct measurements (logit z-scores) from the normal approximation to the posterior distribution (Fig. 1). The mean of this distribution is the EAP score, and the standard deviation is the standard error of measurement. We obtained EAP scores and standard errors of measurement from the *mirt* package, but these are also available from the model's online calculator [manuscript under review with *JCE*].

We transformed plausible values in the same manner as EAP scores, performed the mixed-effects analysis on each set of draws, then pooled model parameters using Rubin's rule [16].

2.6. Minimal important difference

To understand what differences in OKS IRT scores should be considered meaningful, we calculated MID estimates using publicly available data from the NHS PROMs program [17]. This national audit collects preoperative and 6-month postoperative OKS scores for patients undergoing knee replacement in NHS England. We used data collected between 1st April 2012 and 31st March 2020, estimated to represent approximately 50% of procedures undertaken during this period [18]. At follow-up, patients were also asked an anchor question: "Overall, how are the problems now in the knee on which you had surgery, compared to before your operation?" The response options to this question were: "Much worse", "A little worse", "About the same", "A little better", and "Much better". We defined the MID as the difference in mean postoperative scores between the groups who responded "About the same" and "A little better". We first estimated the MID using sum-scores, and then re-estimated the MID using scaled EAP scores (for consistency, these were scaled using the linear transformation described above, with the mean and standard deviation of baseline sum-scores in the TOPKAT trial). We did not apply PVI when estimating the MID as, by definition, the pooled estimate of multiple posterior distribution draws will tend toward the EAP score as the number of draws increase.

3. Results

3.1. TOPKAT sensitivity analyses

The full results of the TOPKAT trial, including sample characteristics, missing data, and attrition and crossover rates, have been published in detail previously [3].

Table 1. Marginal mean OKS scores and Tukey-adjusted *P* values comparing treatment groups at each time point, with each scoring system; 95% confidence intervals (CIs) are presented in brackets.

	Baseline	2 months	1 year	2 years	3 years	4 years	5 years
Sum scoring							
Mean PKR score	18.7 [17.5, 19.9]	30.9 [29.7, 32.0]	36.6 [35.4, 37.8]	37.5 [36.3, 38.7]	37.5 [36.3, 38.7]	37.5 [36.2, 38.7]	37.4 [36.2, 38.6]
Mean TKR score	18.9 [17.8, 20.1]	29.0 [27.8, 30.2]	34.8 [33.6, 36.0]	36.2 [35.0, 37.4]	35.9 [34.7, 37.1]	36.2 [35.0, 37.4]	36.7 [35.4, 37.9]
P value	0.785	0.030	0.030	0.114	0.066	0.158	0.369
EAP scoring							
Mean PKR score	18.8 [17.7, 19.8]	28.6 [27.5, 29.7]	34.0 [33.0, 35.1]	34.9 [33.8, 36.0]	35.0 [33.9, 36.1]	34.9 [33.8, 36.1]	34.9 [33.8, 35.9]
Mean TKR score	18.9 [17.8, 19.9]	27.1 [26.0, 28.2]	32.5 [31.4, 33.5]	33.6 [32.6, 34.7]	33.4 [32.3, 34.5]	33.8 [32.7, 34.9]	34.1 [33.1, 35.2]
P value	0.902	0.058	0.041	0.103	0.043	0.153	0.370
PVI							
Mean PKR score	18.8 [17.7, 19.9]	28.6 [27.5, 29.8]	34.0 [32.9, 35.2]	35.0 [33.8, 36.2]	35.1 [33.8, 36.3]	35.0 [33.8, 36.2]	34.9 [33.6, 36.1]
Mean TKR score	18.9 [17.7, 20.0]	27.1 [25.9, 28.3]	32.5 [31.3, 33.7]	33.6 [32.4, 34.8]	33.4 [32.2, 34.6]	33.8 [32.6, 35.0]	34.1 [32.9, 35.4]
P value	0.939	0.059	0.063	0.112	0.059	0.162	0.424

PKR, partial knee replacement; TKR, total knee replacement; EAP, expected a posteriori; PVI, plausible value imputation.

Table 1 and Figure 2 present the marginal mean OKS scores for each treatment group, at each time point, calculated with each mixed-effects model. The mean baseline OKS sum-score across both trial arms (used for EAP scaling) was 18.9 and the standard deviation was 7.1.

At the 0.05 level, Tukey-adjusted *P* values suggested statistically significant differences in the mean sum-score between groups at 2 months and 1 year. When EAP scoring was applied, we found statistically significant differences at 1 and 3 years (Table 1 and Fig. 2).

When we used PVI to account for individual-level measurement error, there were no statistically significant differences between the groups at any time point. For each comparison in the PVI analysis, *P* values were larger, and CIs were broader, than in the EAP analysis. At the p 2-sided 5% significance level, the statistically significant differences between groups found at 2 months and 1 year (sum-scoring) or 1 year and 3 years (EAP scoring) but not once individual-level measurement error was accounted for through PVI.

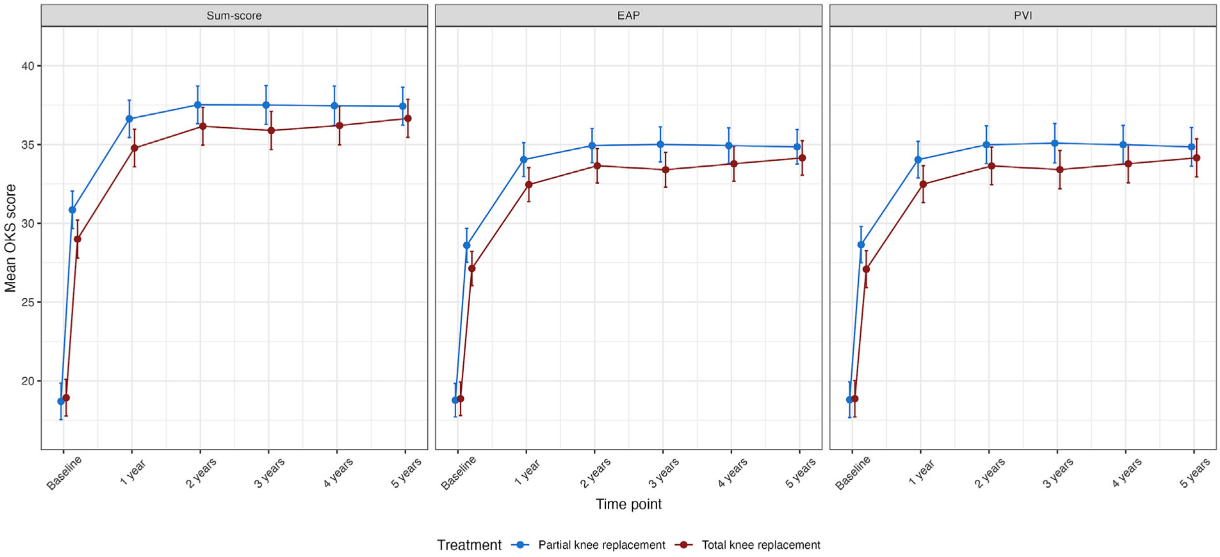


Fig. 2. Marginal mean OKS score and 95% confidence intervals (CIs) at each time point in the TOPKAT trial, calculated through sum-scoring, expected a posteriori (EAP) scoring and plausible value imputation (PVI).

Table 2. Distribution of postoperative OKS sum-scores and scaled EAP scores in the NHS PROMs program dataset, stratified by response to the question ''Overall, how are the problems now in the knee on which you had surgery, compared to before your operation?''

	Much worse	A little worse	About the same	A little better	Much better
Sample size	7,499	11,683	15,217	54,192	25,9440
Mean sum-score (standard deviation)	14.9 (8.2)	20.9 (7.8)	23.5 (8.0)	28.6 (7.6)	39.3 (6.7)
Mean scaled EAP score (standard deviation)	14.3 (7.6)	20.5 (6.6)	22.6 (6.7)	27.0 (6.4)	36.3 (6.5)

3.2. *Minimal important difference*

The NHS PROMs program dataset contained 348,031 complete postoperative OKS response sets, paired to an anchor question response. The median age band of the sample was 70 to 79 years, and 57% of respondents were female.

Table 2 demonstrates the distribution of OKS scores (sum-scores and scaled EAP scores) for respondents endorsing each anchor question response. The difference in mean postoperative sum-score between those responding ''About the same'' and ''A little better'' was 5.1, and the mean difference in scaled EAP score was 4.4. The variance in EAP scores was smaller than the variance in sum-scores for every category.

4. Discussion

This study has applied IRT-based EAP scoring and PVI to the primary outcome measure of an existing high-quality pragmatic RCT. As well as demonstrating that it is feasible to do this retrospectively, (and without any barrier to doing so prospectively in future), this study also demonstrates several reasons why this may be a useful additional analytical strategy for trialists to consider.

Theoretically, the use of IRT scoring in RCTs and the quantification of measurement error through PVI accommodates the naturally unequal weighting of PROM items and responses in the minds of patients, provides potentially more granular scoring by accounting for item response patterns (rather than simply the sum of item responses), and allows measurement error to be accounted for on the individual level. In this study, PVI resulted in broader CIs surrounding the adjusted mean group estimates, and larger *P* values in the pairwise comparison of OKS scores between trial arms. CI width can be considered an indicator of group measurement precision in the broader estimation sense. If we were to accept that this strategy truly captures individual-level measurement error, this finding might reduce our confidence in the apparent early differences between partial and total knee replacement which were identified in the original sum score analysis, and with EAP scoring.

The results of these psychometric sensitivity analyses were not dramatically different to the results of the original trial, and the use of PVI had a greater impact on the CIs surrounding the marginalized means than on the means themselves. This may reflect a high level of construct validity for the OKS, and simulation studies have shown similar results when PVI is applied to high-quality PROMs used in depression [9]. Nevertheless, other simulations have shown that with some PROMs, the use of sum-scoring can lead to substantial bias in the setting of longitudinal RCTs [10], and psychometric sensitivity analyses might have a greater impact in trials like these, which use PROMs with a poorer construct validity.

Trialists depend on PROM developers to produce accurate, reliable, and precise instruments that can identify meaningful differences between trial arms. In addition to the possibility of EAP and PVI scoring, a typical IRT analysis will test whether a PROM meets other prerequisites for high quality measurement, for example that items which measure experientially unrelated health constructs are not combined to produce composite scores, and that the measurement properties of items do not differ significantly between population subgroups. Ideally, the psychometric properties of a PROM, including IRT modeling and assumption testing, would be explored before that PROM was used in an RCT, but most trial funders do not explicitly mandate minimum construct validity standards that a PROM must meet before it is included in an RCT. In many cases, the most appropriate PROM will not have undergone IRT validation. In these circumstances, IRT assumption testing and modeling could be conducted retrospectively with all available trial data, and in future, there may be a role for these types of analyses as part of a trial's pilot study.

Our work may have implications for RCT sample size calculations. Accounting for measurement error with PVI is likely to decrease precision and increase sample size requirements. When this is not accounted for, trials may be at risk of underpowering. Standard error of measurement in PROM response sets is related to the number of items, the number of response options, the targeting of the PROM, and the proportion of missing responses. Shorter and poorly targeted questionnaires with fewer response categories and a high proportion of missing responses will produce less precise latent construct measurements than long questionnaires with many well-targeted response categories and a low proportion of missing responses [19].

In this study, we estimated the sum-score MID of the OKS as 5.1 and the transformed EAP score MID as 4.4. These should be compared cautiously as the scales on which they are measured are not the same. The limits of the OKS sum-score are 0 and 48, while the transformed EAP scores derived from the lowest and highest possible response patterns are −5.19 and 46.77. The EAP score graduations are continuous while the sum-score graduations are not equidistant. In these analyses, sum-score differences can be interpreted with reference to the sum-score MID and EAP differences with reference to the EAP-derived MID. In this study, sum-score analyses showed between-group differences smaller than the sum-score MID and EAP score analyses showed between-group differences smaller than the EAP MID. This has not changed the interpretation of the trial results−differences in the mean OKS score between groups did not exceed the estimated threshold for clinical importance. A responder analysis [20] could be undertaken to compare the proportion of patients reporting a meaningful improvement in each trial arm. While comparisons of group means illustrate the average health gains in each trial arm, responder analyses aim provide information about the likelihood of patients experiencing meaningful benefit from each intervention. For simplicity and consistency with the original TOPKAT publication, and due to limitations in responder analyses [21], we have not performed this here, but in future, psychometric sensitivity analyses may play a role in both approaches to measuring the comparative impact of an intervention.

It is possible that EAP scoring could result in smaller MID estimates, due to the additional granularity of continuous EAP scoring (with complete response sets, an individual can achieve 49 possible OKS sum-scores, but over 244 million different EAP scores) [manuscript under review with *JCE*]. Smaller MIDs could increase sample size requirements for RCTs and comparative observational studies, and the impact of IRT on MID estimates deserved further investigation.

There are limitations to this study, and to psychometric sensitivity analyses more generally. Firstly, IRT modeling is only appropriate for PROMs that meet four assumptions: unidimensionality, monotonicity, local independence, and measurement invariance [22]. Item response theory and factor analyses have demonstrated that the OKS items combine to produce a single monotonic scale (some authors have suggested two highly correlated constructs, representing pain and function [23], but the vast majority of item covariance is explainable by a single *knee health* score) [manuscript under review with *JCE*]. And while at any given level of knee health, men report less difficulty kneeling than women, the measurement properties of the 12 items combined do not meaningfully differ by age or gender [manuscript under review with *JCE*]. In such cases, where IRT assumptions have been met and model fit demonstrated, it is reasonable to apply EAP scoring and PVI. But in many cases, PROMs will not demonstrate these properties.

Secondly, PVI involves assumptions about the distribution of plausible measurements, based on the IRT model. It is possible that the PVI analysis conducted in this study has underestimated measurement error in the TOPKAT trial. By randomly drawing measurements from the response sets' posterior distributions, we have captured the measurement error described by the IRT model under the assumption that the model perfectly describes the relationship between item responses and latent construct levels. This does not necessarily capture the *true* error. An extension to PVI, which might capture the uncertainty of both the model and the measurement relative to the model is full Bayesian modeling [9]. In this framework, IRT model parameters themselves are drawn from a plausible distribution before they are used to compute the plausible distribution of latent construct measurements. At present, full Bayesian modeling is seldom applied to health measurement, but it may be highly appropriate for IRT modeling in RCTs where models are based on smaller sample sizes.

It is possible that PVI protects from type-1 error, or alternatively that it increases risk of type-2 error, and it is likely to be most impactful in superiority trials that demonstrate a small but statistically and clinically significant difference between arms. Concerns may arise in early phase trials, where "signal" of a potential effect for a new intervention is being sought and the intervention still requires optimization. There, the potential loss of relative precision from using PVI could prematurely halt a direction of meaningful scientific development. However, this scenario should not be the case for pragmatic trials, where real-world impact of an existing intervention is being evaluated. For this reason, we chose to explore PVI in a pragmatic way, rather than an explanatory RCT. Pragmatic RCTs are likely to influence health service policy change, and minimizing type-1 error when the trial results may lead to the reallocation of significant health service resource may be desirable. More work is required to understand when psychometric sensitivity analyses may be most useful, but we would suggest this is likely to be in situations where a PROM has not been developed with contemporary psychometric techniques, as this may have a relatively higher impact on the precision and interpretability of trial findings.

In this study, we use *P* values as a nominal marker and to align with the primary TOPKAT trial analysis. There are problems with the use of *P* values, which provide no information on the probability of the alternative hypothesis being true. In future, psychometric sensitivity analyses could be explored within alternative estimation frameworks [24].

Here, we retrospectively applied IRT and PVI to a completed trial. In theory, these techniques could be applied prospectively in the primary analysis of future RCTs, after researchers have confirmed that their PROM meets the assumptions of IRT [manuscript under review with *JCE*]. However, given the potential limitations discussed, using PVI as a secondary sensitivity analysis may be more appropriate. We would advocate further exploration into the role of psychometric sensitivity analyses for RCTs. Demonstrating that an effectiveness trial's superiority finding persists when subjected to sensitivity analyses based on alternative measurement assumptions should increase the impact of the study for all stakeholders concerned and support value for society.

5. Conclusion

This study has demonstrated the application of IRT to account for differences in item importance and individual-level measurement error using a trial dataset. Trialists should consider working with psychometricians when planning RCTs that use PROMs, and there may be benefit to performing psychometric sensitivity analyses on existing trial data. Funders, reviewers, and policy makers should be aware of measurement error in PROMs and interpret trial protocols and results with this in mind. Future work is needed to establish the potential benefits of psychometric sensitivity analyses in clinical trials.

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.

References

[1] Dal-Ré R, Janiaud P, Ioannidis JPA. Real-world evidence: how pragmatic are randomized controlled trials labeled as pragmatic? BMC Med 2018;16(1):49.

[2] Sorenson C, Drummond M, Chalkidou K. Comparative effectiveness research: the experience of the national Institute for health and clinical excellence. J Clin Oncol 2012;30:4267−74.

[3] 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(10200):746−56.

[4] 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.

[5] Beard DJ, Rees JL, Cook JA, Rombach I, Cooper C, Merritt N, et al. Arthroscopic subacromial decompression for subacromial shoulder pain (CSAW): a multicentre, pragmatic, parallel group, placebo-controlled, three-group, randomised surgical trial. Lancet 2018; 391(10118):329−38.

[6] Cai L, Choi K, Hansen M, Harrell L. Item response theory. Annu Rev Stat Appl 2016;3(1):297−321.

[7] 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.

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

[9] 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.

[10] Gorter R, Fox JP, Apeldoorn A, Twisk J. Measurement model choice influenced randomized controlled trial results. J Clin Epidemiol 2016;79:140−9.

[11] Gorter R, Fox JP, Riet GT, Heymans M, Twisk J. Latent growth modeling of IRT versus CTT measured longitudinal latent variables. Stat Methods Med Res 2020;29(4):962−86.

[12] 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.

[13] Lenth R, Burekner P, Herve M. Package "emmeans." https://cran.r-project.org/web/packages/emmeans/emmeans.pdf. Accessed October 23, 2022.

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

[15] Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. Med Care 2000;38:28−42.

[16] Marshall A, Altman DG, Holder RL, Royston P. Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. BMC Med Res Methodol 2009;9:57.

[17] 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.

[18] 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.

[19] Holman R, Glas CAW, de Haan RJ. Power analysis in randomized clinical trials based on item response theory. Controlled Clin Trials 2003;24:390−410.

[20] Guyatt GH, Juniper EF, Walter SD, Griffith LE, Goldstein RS. Interpreting treatment effects in randomised trials. BMJ 1998;316: 690−3.

[21] Ferreira GE, McLachlan AJ, Lin CWC, Zadro JR, Abdel-Shaheed C, O'Keeffe M, et al. Efficacy and safety of antidepressants for the treatment of back pain and osteoarthritis: systematic review and meta-analysis. BMJ 2021;1:13. m4825.

[22] Nguyen TH, Han HR, Kim MT, Chan KS. An introduction to item response theory for patient-reported outcome measurement. Patient 2014;7(1):23−35.

[23] 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.

[24] Halsey LG. The reign of the p -value is over: what alternative analyses could we employ to fill the power vacuum? Biol Lett 2019; 15(5):20190174.