Journal of Plastic, Reconstructive & Aesthetic Surgery 104 (2025) 434–439

Evaluating the precision of computer adaptive testing in longitudinal hand surgery analyses: A psychometric approach

JS Teunissen [a,b,c], BEPA van der Heijden [a,d], JN Rodrigues [e,f,*], F Issa [b], CJ Harrison [g], Contributors of the BSSH UK Hand Registry [1]

[a] *Department of Plastic, Reconstructive and Hand Surgery, Radboud Institute of Health Research, Radboud University Medical Centre, Geert Grooteplein Zuid 10, 6525 GA Nijmegen, the Netherlands*
[b] *Nuffield Department of Surgical Sciences, University of Oxford, Oxford OX3 9DU, UK*
[c] *Department of Plastic, Reconstructive and Hand Surgery, Erasmus MC, Dr. Molewaterplein 40, 3015 GD Rotterdam, the Netherlands*
[d] *Department of Plastic, Reconstructive and Hand Surgery, Jeroen Bosch Hospital, Henri Dunantstraat 1, 5223 GZ 's-Hertogenbosch, the Netherlands*
[e] *Warwick Clinical Trials Unit, University of Warwick, Coventry CV4 7AL, UK*
[f] *Department of Plastic, Reconstructive, and Hand Surgery, Stoke Mandeville Hospital, Mandeville Rd, Aylesbury HP21 8AL, UK*
[g] *Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford OX3 7LD, UK*

Received 8 October 2024; Accepted 10 February 2025

KEYWORDS
PROMs;
PEM;
Cubital tunnel syndrome;
Item response theory;
Longitudinal analysis;
Sensitivity

Summary Patient-reported outcome measures (PROMs) are essential in hand surgery for assessing patient health but it can be time-consuming. Computerized adaptive testing (CAT) offers a more efficient alternative by reducing the number of questions asked. This study sourced the data of 268 patients undergoing cubital tunnel release from the UK Hand Registry to evaluate whether CAT's inherent imprecision affects longitudinal research conclusions. Mean patient evaluation measure (PEM) scores at baseline, 2 months and 6 months from the traditional full-length assessment (10 questions) were compared with the simulated scores assuming that the CAT version (median of 2 questions) was used. Both methods showed significant improvements in PEM scores post-surgery ($p < 0.01$), with minimal differences between the mean scores and overlapping confidence intervals. These findings confirm that CAT replicates full-

* Corresponding author at: Warwick Clinical Trials Unit, University of Warwick, Coventry CV4 7AL, UK.
E-mail address: j.rodrigues@warwick.ac.uk (J. Rodrigues).
[1] See Appendix A Contributors of the BSSH UK Hand Registry.

https://doi.org/10.1016/j.bjps.2025.02.013

length PROM results while significantly reducing patient burden, thereby supporting its use in clinical and research settings for hand surgery.
Level of evidence: III

Patient-reported outcome measures (PROMs) are questionnaires designed to measure a patient's health status and are being increasingly adopted in hand surgery.[1]

Computerized adaptive testing (CAT) is a psychometric technique that can reduce the length of a PROM by selectively administering only the most relevant questions to individual respondents. Its efficiency and reduced participant burden make it an attractive alternative to traditional full-length assessments, particularly in clinical settings where time is limited. In the field of hand surgery, CAT versions of the patient evaluation measure (PEM) have been produced and validated under conditions such as thumb base osteoarthritis and cubital tunnel syndrome (CuTS).[2,3]

In CAT, there is a theoretical trade-off between the number of questions asked and score precision. The fewer questions asked, the more imprecise the score. This balance is often achieved through stopping rules that force the CAT to continue asking questions until a prespecified level of precision is met. One concern with this is whether the imprecision introduced by CAT could lead to real-world error in the results of an observational study or randomised controlled trial (RCT).

Plausible value imputation (PVI) is a cutting-edge psychometric technique that can examine the impact of imprecision in measurement on research conclusions.[4] This individual-level imprecision is not typically accounted for in classical PROM analyses.

In this study, we revisited data from the UK Hand Registry on patients undergoing cubital tunnel release. We previously demonstrated that patients undergoing cubital tunnel release showed significant improvement in hand function using PEM sum scoring.[5] In this simulation study, we repeated the analysis assuming that CAT was used, accounted for imprecision using PVI and compared these results to the original findings. We aimed to understand whether the imprecision introduced by CAT could make a meaningful difference to the results or conclusions of such observational analysis.

Methods

Study design and setting

This study made secondary use of data from our previously published study on the outcomes of surgical decompression for primary CuTS.[5] These data originated from the UK Hand Registry, which is a voluntary national registry, and the data were used to evaluate surgical outcomes for hand and wrist conditions. Patients who agreed to enter their details to the registry were asked to complete the PEM before surgery (either in-situ decompression or decompression with subcutaneous transpositions) and at 2 months and 6 months after surgery.

Patient evaluation measure

The PEM was originally developed in 1995 to measure hand function.[6] It is the main PROM of the UK Hand Registry and was used in several high-profile randomized controlled trials. Between 2012 and 2017, the UKHR captured the original ten-item version of the PEM.[6] This was changed to the updated 11-item version in 2017,[7] which has an additional question concerning the duration of pain. As this item was missing for most patients in the registry, we chose to use the complete response sets of the original 10-item for the analysis. Each question has 7 answer options (all scored from 1 to 7). The total sum-scores ranges from 10-70 with higher scores indicating more symptoms (worse hand function). Our analysis did not include parts one or 3 of the PEM questionnaire as these parts measure the care process and are more akin to a patient-reported experience measure, and therefore was not considered a hand function measure.

Item response theory and CAT

In a previous study, an item response theory (IRT) model was developed based on the PEM responses of 522 patients with CuTS from the UK Hand Registry database.[3] This study confirmed that the assumptions of IRT were met by PEM to measure hand function in the UK population; details have been published before.[3] In the same study, we showed that CAT could predict full-length PEM scores from a median of 2 questions with a precision threshold ≤ 0.3, consistent with patient-reported outcome measurement information system (PROMIS).[3,8,9] However, the impact of this level of imprecision on longitudinal hand surgery analyses is not known and is examined in this study.

Statistical analysis

The original analysis evaluated PEM sum scores at baseline and at 2 and 6 months following cubital tunnel release using linear mixed models (LMM) that were adjusted for age and sex. In this study, IRT scores were used instead of sum scores. These IRT scores can potentially be more accurate and precise than sum scores; more details on this concept are presented in the Supplementary file (Appendix A).

PVI

A schematic explanation of the analyses is provided in Figure 1.

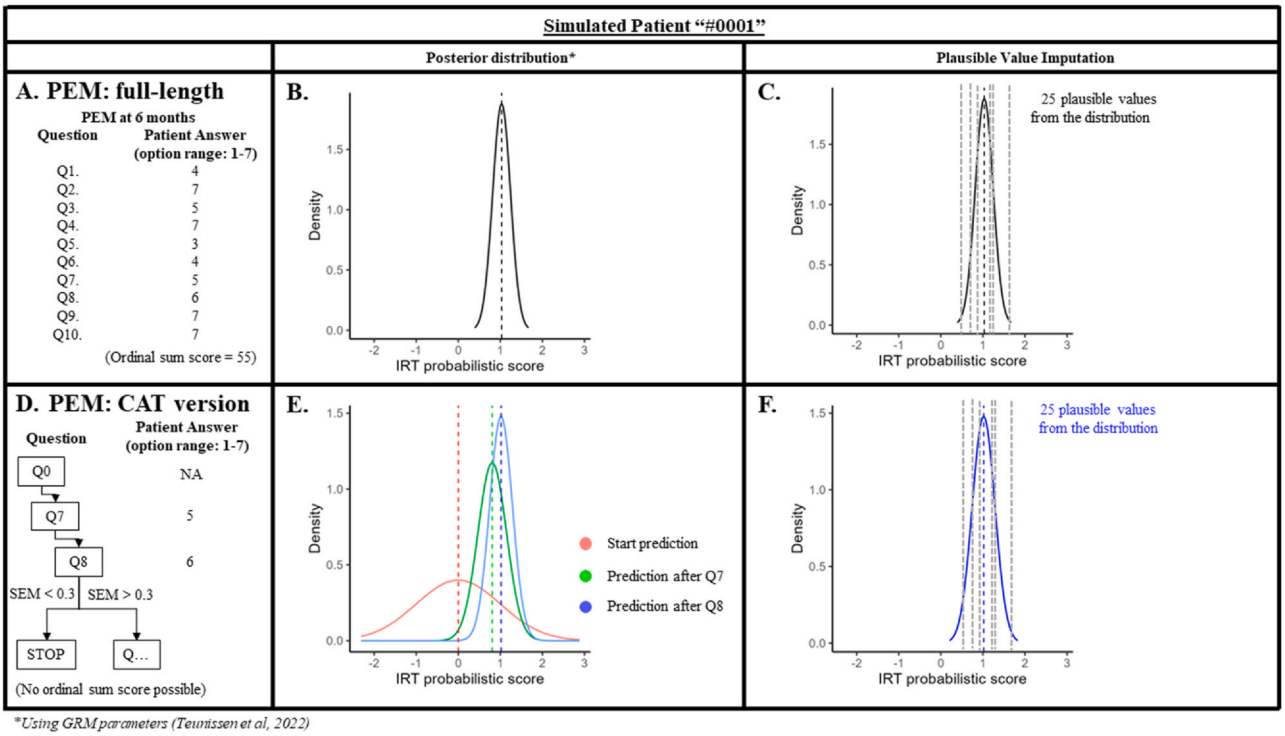

Figure 1 Schematic visualisation of the different statistical approaches. Panels A-C indicate the full-length PEM IRT analysis pathway, panels D-E indicate the pathway if CAT had been used instead of the full-length PEM. Abbreviations: *PEM* Patient Evaluation Measure; *GRM* Graded Response Model. **A.** Traditionally, a patient fills in all 10 questions of the PEM. Subsequently, the ordinal sum score (in this case 55) can be calculated. **B.** IRT constructs a posterior distribution (black parabola) based on the patient's response pattern. This patient's response pattern corresponds with an IRT score between 0.5 and 1.8, with the most likely score being approximately 1.0 (vertical dotted black line). **C.** Instead of taking only the most likely score, 25 values are randomly drawn from the posterior distribution (vertical dotted grey lines) to account for its breadth. **D.** Using CAT, only relevant questions are administered to the patient based on their previous answer until a prespecified precision threshold is met. In this example, the patient only needs to fill in Q7 and Q8 to meet the prespecified threshold (Sem < 0.3), instead of all 10 questions. **E.** Before the patient completes the first question, the IRT score is very imprecise (red parabola) indicated by its large breadth. As more questions are completed, the IRT scores become increasingly precise (green parabola and subsequent blue parabola). Notably, the final prediction (blue parabola) is very similar to the prediction from panel B (black parabola), but continues to be moderately broad (imprecise). **F.** Again, 25 values are randomly drawn (vertical dotted grey lines) from the posterior distribution.

In IRT, a patient's response pattern to the PEM is modelled into a probability distribution (the posterior distribution) of their hand function (Figure 1B, 1E). The graph (Figure 1B, 1E) shows that the patient may have a certain level of hand function, based on their answers. These IRT scores are on a different scale than the original PEM (Appendix B) and range from −2.31 to 2.88 in this patient population.

The breadth of the posterior distribution represents the imprecision of the measurement (broader distribution indicates more imprecision) and may differ between response patterns and number of responses.

As CAT administers fewer questions than the full-length questionnaire, the breadth of the posterior distribution is potentially larger and could introduce imprecision. To account for this imprecision, we randomly drew 25 plausible values from the posterior distribution per patient, for each time point (Figure 1C,1F). Subsequently, we performed an LMM for every set of draws (25 LMMs in total) and pooled their results using Rubin's rules.[10]

First, we used the posterior distribution of the full-length PEM responses of all patients (Figure 1A-C). Then, we repeated the analysis a second time through simulation, assuming that the patients had only responded to the questions selected by CAT instead of responding to the full-length PEM (Figure 1D-F). By comparing the mean IRT scores and confidence intervals (CIs) of the full-length PVI analysis and CAT PVI analysis, we aimed to understand the impact of any CAT-related imprecision on the analysis.

All analyses were performed using R statistical software. Missing data were not imputed as they did not appear to increase the precision of the LMM.[11] A p-value < 0.05 was considered statistically significant.

Results

The original study consisted of 268 adult patients treated between February 2012 and April 2019 and found a significant improvement in hand function after surgery using traditional PEM sum-scoring analysis.[5] Characteristics of the 268 patients are shown in Table 1; the availability of data points is presented in Table S1.

Table 1 Summary characteristics of the patients included in the analysis.

Characteristics	Value
N	268
Age (years, median (IQR))	55 (45–65)
Sex = Females (N (%))	134 (50)
Operation (N (%))	
In-situ decompression	222 (83)
+ subcutaneous transposition	46 (17)

IQR Interquartile range.

PVI scores in full-length PEM responses vs. CAT PEM

Full-length and CAT assessments showed a significant improvement in IRT scores for the overall group and subgroups in the first 6 months after surgery (all $p < 0.01$; Figure 2). The estimated PVI scores at intake, 2 months and 6 months were similar (Figure 2), with mostly overlapping point estimates and CIs. The exact estimated marginalised means including CI are presented in Table S2. When considering all patients, the largest absolute mean difference in θ scores was 0.036 points between the full-length questionnaire and

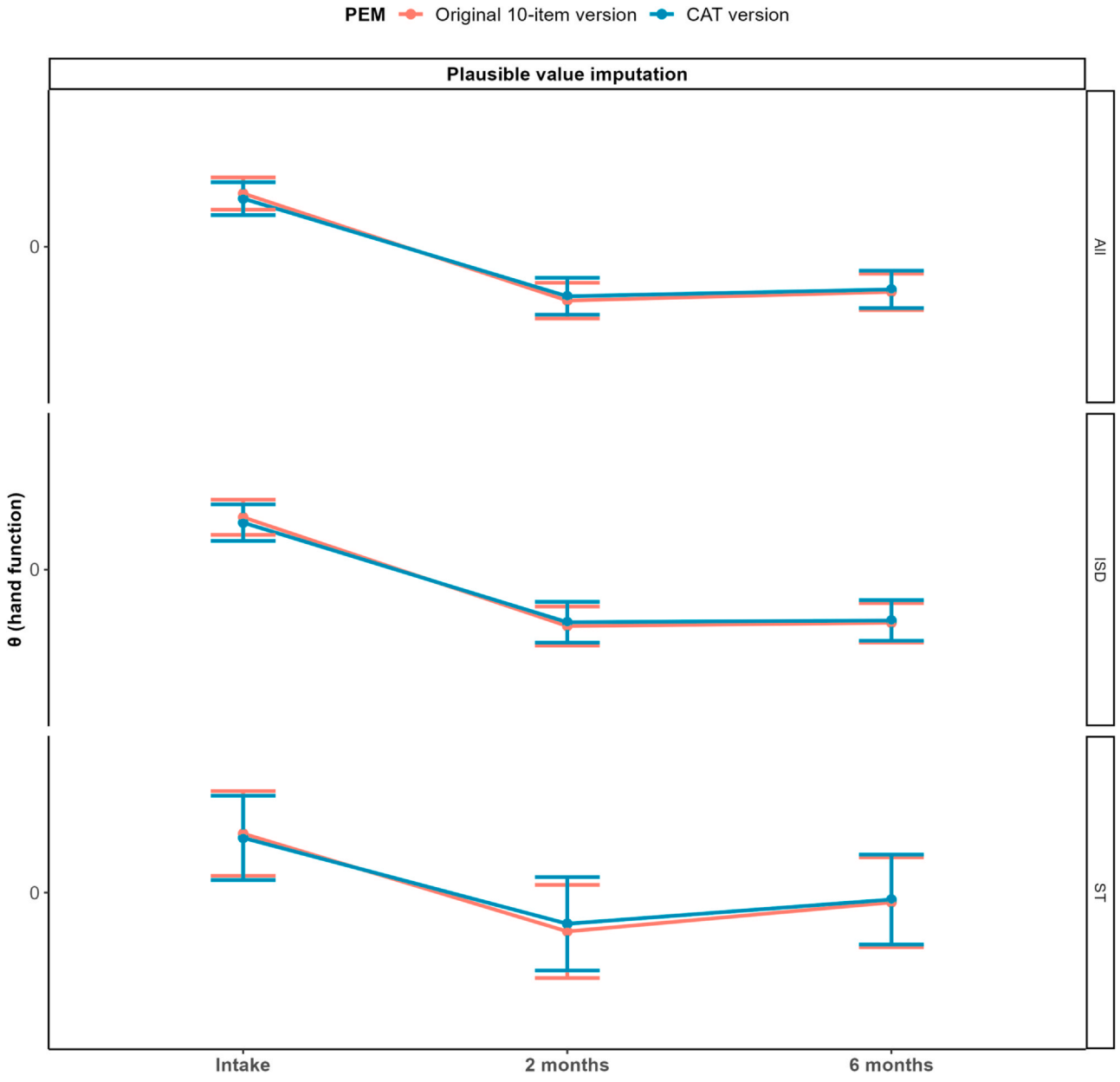

Figure 2 Estimated marginalised means of the PEM including 95% CI derived from the full-length (FL) questionnaire (red lines) or computerized adaptive test (CAT) (blue lines) at intake, 2 months and 6 months for all patients, and stratified based on the type of surgery (either in-situ decompression (ISD) or decompression + subcutaneous transposition (ST)). The estimates from the CAT are remarkably similar to the estimates derived from the full-length questionnaire as is shown by the mostly overlapping lines and CIs.

CAT. The CIs of the CAT version were slightly broader as expected; however, this did not change any conclusions.

Discussion

CAT is a psychometric technique that can efficiently reduce the length of a hand surgery PROM, thereby decreasing the patient burden of filling in the complete questionnaires. However, hand surgeons and researchers may be concerned that CAT may introduce imprecise measurements, which might bias the longitudinal results of an observational study or RCT. In this simulation study, we repeated our analysis of UKHR patients undergoing cubital tunnel release, assuming that they had completed CAT instead of the full-length PEM questionnaire. We found that using CAT would have reduced the number of questions answered from 10 to a median of 2 without introducing any meaningful imprecision or changing the conclusion of the primary analysis. This should reassure hand surgeons and researchers who are considering the use of the growing number of CAT algorithms that are available for capturing patient-reported outcome data in hand surgery.

The findings of this study complement the existing evidence demonstrating the similarity between CAT and full-length PROM scores. Previous studies have shown that CAT versions of PEM for conditions such as thumb base osteoarthritis and CuTS can accurately predict outcomes with significantly fewer questions.[2,3] However, these studies focused on the cross-sectional equivalence of CAT and full-length assessments. The current study was needed to evaluate the impact of CAT on longitudinal analyses, specifically to determine if the potential imprecision introduced by fewer questions in CAT would affect the robustness of longitudinal outcome data in hand surgery. Our results indicate that even when fewer questions are asked, the longitudinal conclusions remain consistent. These results reinforce the reliability of CAT in tracking patient outcomes over time and support its use as an efficient and dependable method for outcome measurement in clinical practice.

The use of PVI to account for measurement imprecision is a strength of this study. By incorporating the breadth of the posterior distribution of IRT scores, PVI provides a method to assess the potential impact of imprecision on PROM scores. All measurements have potential for error. Typically, PROM scores are presented as a single value, without accounting for this potential measurement error. However, with modern electronic scoring systems, we can present PROM scores together with credible intervals, indicating the 'margin of error' in each score.[12] This margin of error can be accounted for at the group level using PVI.[4]

This study has some limitations. First, the PEM is particularly suited for CAT because each question provides a large amount of discriminatory information (each question has 7 response options). However, CAT may not be suitable for every other hand surgery PROM. Second, this analysis was performed using a CAT that was already developed with high precision (SEM < 0.3). CATs that are developed using less strict precision criteria (e.g. paediatric PROMIS) could generate larger CIs once the imprecision is accounted for. In such cases, including PVI as a sensitivity analysis may be beneficial. Third, although PVI is a powerful tool for addressing imprecision, it currently requires relatively advanced and niche statistical expertise and is not readily accessible to those who might benefit from it. Lastly, the follow-up was limited to 6 months, by which time the patients may not have reached their final functional outcomes. Future research should assess the stability of CAT precision over longer follow-up periods.

This research adds to the validity of routine use of CAT in research and clinical settings. Using stopping rules with an SEM < 0.3 ensures precise measurements for groups, and in some instances, lenient stopping rules might be sufficient. PVI can be applied when there are concerns regarding measurement imprecision. Creating user-friendly PVI software would help make this approach more accessible in clinical practice.

Statement of ethics

Ethical approval for this study was waived by The University of Oxford Clinical Trials and Research Governance team because it comprises secondary use of anonymous data controlled by a registered charity for research purposes (the data were originally collected for quality assurance); this does not require ethical approval in the United Kingdom.

Funding

Conrad J. Harrison was funded by a National Institute for Health Research (NIHR) Doctoral Research Fellowship (NIHR300684) for this research project. Jeremy N. Rodrigues is funded by a NIHR postdoctoral fellowship (PDF-2017-10-075). This document presents independent research funded by the NIHR. The views expressed are those of the author(s) and not necessarily those of the United Kingdom's National Health Service, NIHR or Department of Health.

Author contributions

All authors were involved in protocol development. J.S. Teunissen and C.J. Harrison performed the statistical analysis. All authors interpreted the results. J.S. Teunissen and C.J. Harrison wrote the first draft of the manuscript. All authors reviewed and edited the manuscript and approved the final version of the manuscript.

Declaration of Competing Interest

The author(s) declare no potential conflicts of interest concerning the research, authorship, and/or publication of this article.

Acknowledgments

We would like to thank all those who have contributed to the UK Hand Registry, without whom this work would not be

possible. In particular, we would like to acknowledge and thank Sue Fullilove, who established the UK Hand Registry and continues to champion this valuable resource for UK hand surgery research and innovation.

Appendix A. Non-author contributors

Alastair Graham, Andrew Logan, Andy McKee, Anthony Heywood, Charles Gozzard, Claire Edwards, Donald Sammut, Gillian Eastwood, Graham Cheung, Harry Lyall, Ian Grant, Ian McNab, Ibrahim Roushdi, Indranil Chakrabarti, James Nicholl, Jeremy Field, Jill Arrowsmith.

John Harrison, Jonathan Jones, Kate Owers, Lindsay Muir, Lisa Leonard, Malcolm Jones, Margaret Birks, Mark Hobson, Martin Wood, Michael Eames, Michael Shanahan, Nicholas Downing, Oliver Harley, Paul Critchley, Paul Stuart, Ramesh Chennagiri, Rashpal Bassi, Rebecca Dunlop, Rob Savage, Rosalyn Harper, Sean Walsh, Simon Knight, Soham Gangopadhyay, Stephen Bostock, Stephen Royle, Steve Hodgson, Sue Fullilove, Tim Davis, Timothy Hems, Vivien Lees, Will Mason.

Appendix B. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.bjps.2025.02.013.

References

1. Wouters RM, Jobi-Odeneye AO, de la Torre A, Joseph A, Hovius SER. A standard set for outcome measurement in patients with hand and wrist conditions: consensus by the International Consortium for Health Outcomes Measurement hand and wrist working group. *J Hand Surg Am* 2021 Jul;**46**(10):841–55.
2. 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 trapezio-metacarpal arthritis. *J Hand Surg Eur Vol* 2022 Mar;**47**(9):893–8.
3. Teunissen JS, Hovius SER, Ulrich DJO, Issa F, Rodrigues JN, Harrison CJ. Computerized adaptive testing for the patient evaluation measure (PEM) in patients undergoing cubital tunnel syndrome surgery. *J Hand Surg Eur Vol* 2023 Apr;**48**(10):1042–7.
4. Harrison CJ, Plessen CY, Liegl G, et al. 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. *J Clin Epidemiol* 2023 Jun;**158**:62–9.
5. Teunissen JS, Griffiths TT, van der Heijden BEPA, et al. Changes in hand function and health state utility after cubital tunnel release using the United Kingdom Hand Registry. *J Hand Surg Eur Vol* 2024 Sep;**50**(3):344–51.
6. Macey AC, Burke FD, Abbott K, et al. Outcomes of hand surgery. British Society for Surgery of the Hand. *J Hand Surg Br* 1995 Dec;**20**(6):841–55.
7. Dias JJ, Bhowal B, Wildin CJ, Thompson JR. Assessing the outcome of disorders of the hand. Is the patient evaluation measure reliable, valid, responsive and without bias? *J Bone Joint Surg Br* 2001 Mar;**83**(2):235–40.
8. Gibbons LE, Feldman BJ, Crane HM, et al. Migrating from a legacy fixed-format measure to CAT administration: calibrating the PHQ-9 to the PROMIS depression measures. *Qual Life Res* 2011 Nov;**20**(9):1349–57.
9. Reeve BB, Hays RD, Bjorner JB, et al. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Med Care* 2007 May;**45**(5 1):S22–31.
10. Rubin DB. *Multiple imputation for nonresponse in surveys*. John Wiley & Sons; 2004.
11. Peters SAE, Bots ML, den Ruijter HM, et al. Multiple imputation of missing repeated outcome measurements did not add to linear mixed-effects models. *J Clin Epidemiol* 2012 Jun;**65**(6):686–95.
12. Harrison CJ, Trickett R, Wormald J, et al. Remote symptom monitoring with ecological momentary computerized adaptive testing: Pilot cohort study of a platform for frequent, low-burden, and personalized patient-reported outcome measures. *J Med Internet Res* 2023 Sep;**25**:e47179.