npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-01783-z

High-dimensional item response theory analysis of patient-reported outcomes in total knee arthroplasty

Abel Díaz Berenguer[1,4] ✉, Matías Nicolás Bossa[1,4] ✉, Julien Lebleu[2], Andries Pauwels[2] & Hichem Sahli[1,3]

This study introduces a Bayesian multidimensional hierarchical item response theory (MHIRT) model to improve patient-reported outcome (PRO) assessments in total knee arthroplasty (TKA). Traditional unidimensional scoring fails to capture the multifaceted nature of recovery. Our model uncovers latent traits and inter-item relationships directly from PROMs such as the OKS and the EQ-5D-3L, without relying on predefined subscales. MHIRT flexibly decomposes PROMs into clinically meaningful traits like pain, mobility, self-care, and confidence. These traits captured more domain-specific variation, showed stronger sensitivity to temporal changes, and better reflected demographic factors than traditional total scores. The model was trained on a large NHS dataset and externally validated on PROMs from the moveUP digital platform. In predictive modeling of postoperative outcomes, MHIRT-derived features consistently outperformed unidimensional scores and conventional multidimensional IRT models. These findings suggest that MHIRT offers a potentially interpretable framework for tracking recovery and predicting health outcomes.

Knee osteoarthritis (KOA) is a degenerative condition affecting millions of individuals worldwide. Its incidence poses a major healthcare challenge, contributing to the burden of aging-related health issues[1,2]. KOA primarily affects the cartilage and bone within the knee joint, leading to symptoms such as pain, stiffness, and swelling, along with a reduction in mobility. These manifestations significantly deteriorate and impair the quality of life (QoL) for affected individuals[3]. In cases where KOA progresses to a severe stage, where conservative and non-surgical treatments, such as medication, physical therapy, and weight management, fail to alleviate the debilitating symptoms, total knee arthroplasty (TKA) often emerges as a standard surgical solution. This procedure entails the removal of the knee's damaged parts and their replacement with artificial implants[4].

With the increasing prevalence of KOA, particularly among the elderly population[1], there is a corresponding rise in the number of TKA procedures, a trend expected to continue in the coming years[5]. The outcomes of TKA are influenced by a complex interplay of the surgical procedure itself and individual patient characteristics. Unfortunately, it is reported that between 7% and 30% of patients express dissatisfaction after TKA[6,7]. The reasons for TKA failures are multifactorial, encompassing issues such as aseptic loosening, periprosthetic joint infection, and post-TKA instability, in addition to a myriad of preoperative factors[7], including patient socioeconomic status[8] and overall health conditions. This highlights the importance of comprehensive evaluations by physicians prior to TKA, wherein patient reported outcome measures (PROMs) play a crucial role. Preoperative patient characteristics, often assessed through PROMs, are crucial for evaluating the patient's overall health status and predicting the likelihood of successful TKA outcomes[9,10]. PROMs are item-based instruments to evaluate various characteristics of a patient. These item-based tools are pivotal in capturing a wide array of indicators pertinent to patients' lifestyle, health, and physical status, thereby facilitating a comprehensive assessment of various patient characteristics. They encompass a broad spectrum of factors, including daily activities, knee pain, and knee functional levels. Additionally, these instruments are critical in evaluating psychological aspects, such as levels of anxiety and depression.

PROMs are invaluable tools in preoperative patient evaluations and are essential in measuring the outcomes of TKA. Their use, pre- and post-TKA, enables healthcare providers to follow the evolution of patients' self-perceived health indicators, enabling a patient-centric assessment of the surgical intervention's effectiveness. The orthopedic literature offers a diverse array of PROM instruments and outcome measures for assessing the success of TKA, with most instruments aggregating item responses into a one-dimensional score for assessment. However, it is worth noticing that only some of these measures sufficiently fulfill their intended purposes regarding validity, reliability, and responsiveness. The Oxford Knee Score

[1]Department of Electronics and Informatics (ETRO), Vrije Universiteit Brussel (VUB), Brussel, Belgium. [2]moveUP, Brussel, Belgium. [3]Interuniversity Microelectronics Centre (IMEC), Leuven, Belgium. [4]These authors contributed equally: Abel Díaz Berenguer, Matías Nicolás Bossa. ✉e-mail: aberengu@etrovub.be; mnbossa@etrovub.be

(OKS) is widely recognized as one of the most effective condition-specific PROMs[11] due to its practicality, reliability, and clinical sensitivity for monitoring patient progress from pre- to post-TKA[12]. In conjunction with condition-specific tools like the OKS, the EuroQoL five-dimension three-level version (EQ-5D-3L) questionnaires and the visual analog scale (EQ-VAS)[13] are also self-assessed measures globally utilized for evaluating patients' overall health, say health-related quality of life, before and after TKA. These PROMs are sensitive to clinically significant changes in patient progress follow-up[14]. Several studies have evaluated the performance of widely used PROMs in TKA, particularly the OKS and variants of the EQ-5D. Lin et al.[15] compared OKS, EQ-5D-3L and EQ-VAS using classical test theory metrics, reporting that OKS demonstrated the greatest responsiveness and predictive validity while the EQ-VAS showed stronger predictive performance than the EQ-5D but lower responsiveness. Similarly[16], found that the OKS exhibited high responsiveness and the EQ-5D-3L moderate responsiveness. Both studies supported the concurrent validity of OKS and EQ-5D, reinforcing their value as complementary instruments in outcome monitoring.

However, these studies—and others using similar classical approaches—typically analyze PROMs at the instrument level, relying on pre-defined total or index scores rather than modeling item-level response patterns. While this approach is well-established and interpretable, it does not capture item-specific properties such as how items vary in difficulty, how well they differentiate between patients at different levels of recovery, or whether items reflect multiple underlying aspects of health rather than a single predefined dimension. Moreover, the reliance on fixed, predefined score structures assumes that each item contributes to a single trait, limiting the ability to explore correlations across items, especially when PROMs from different instruments are used in combination.

This motivates the use of item response theory (IRT), a family of statistical models that link item responses to one or more underlying latent traits, such as health status or physical function. Unlike traditional scoring methods that rely on aggregating item responses into a single summary score, IRT estimates item-specific properties—such as difficulty and discrimination—and provides individualized latent trait estimates for each respondent. It also accommodates missing data more effectively, which is particularly relevant in PROM collection where incomplete responses are common[17]. By modeling the latent structure directly at the item level, IRT offers a principled way to identify and interpret multidimensional patterns within and across PROM instruments.

Early IRT models, such as the Rasch logistic model[18], the 2-PL and 3-PL logistic models[19,20], and the graded response model[21], predominantly adopted a unidimensional approach in which each item contributes to a single latent trait (i.e., predefined domain or subscale). This oversimplified assumption often fails to reflect the complex interrelationships and interdependencies among various items across different instruments. Multidimensional IRT (MIRT) models attempt to address this issue by allowing for multiple latent traits. However, as highlighted in Morucci's work[22] these models still rely heavily on a predefined fixed structure for dimensions. This rigid approach may obscure latent traits that are not explicitly accounted for in the predefined model, thus limiting the discovery of novel interrelations and reducing the model's applicability in capturing the multifactorial nature of patient-reported outcomes. Furthermore, reliance on prior assumptions about dimensionality can lead to model misspecification, especially in heterogeneous clinical datasets where item relationships are complex and data-driven solutions are needed. Even flexible MIRT models require pre-specifying the number of dimensions and suffer from factor indeterminacy, a well-known issue where multiple rotated solutions yield equivalent model fit, making trait interpretation unstable and dataset-specific[23].

In this work, we aim to leverage the principles of IRT-based modeling to develop a novel method to build multidimensional assessment instruments that are reliable and very sensitive in assessing the QoL, that is, health-related indicators of patients, both pre- and post-TKA. Our work enables more comprehensive assessment tools to accurately capture the multifactorial nature of patient evolution and outcomes associated with TKA, providing a nuanced understanding of patient well-being and self-perceived effectiveness of TKA. Specifically, we propose a method to discern the correlations between questionnaire items using a novel Bayesian multidimensional hierarchical IRT (MHIRT) model. The proposed method is designed to enhance the sensitivity of pre- and post-TKA assessments of a patient's health-related QoL and foster a more nuanced understanding of the multiple traits dimensions of its health status, such as physical functioning, pain levels, and mental well-being, and how they interrelate. Hence, we seek to provide a robust analytical framework that adequately reflects the complexity and multifactorial nature of patients' experiences and outcomes. In Fig. 1, we illustrate the objectives of this work. While the proposed method is generalizable, this study restricts its scope to PROMs commonly used in TKA follow-up: the OKS and EQ-5D-3L.

To develop the MHIRT model in the scope of pre- and post-TKA assessments, we initially relied on a large dataset comprising responses and data from individuals who underwent TKA and has been made accessible by the England National Health Service (NHS) (https://digital.nhs.uk/). This dataset served as a primary source, first, to fit and confirm the efficacy of the MHIRT model in discerning patterns that reflect the true underlying correlations among questionary items, and second, to assess the usefulness of the latent traits derived from the model for predicting the likelihood of TKA success. Furthermore, we externally validate the usefulness of the proposed model to analyze data collected through the moveUP platform for extracting insights into assessing the efficacy of digital therapies following TKA. MoveUP is an FDA-approved device with registration number: 3023739055 and the product code: ISD, Device class: 2, Regulation number: 890.5360, class 2 exempt. It is also CE-certified as a medical device under the Medical Device Directive 93/42/EEC.

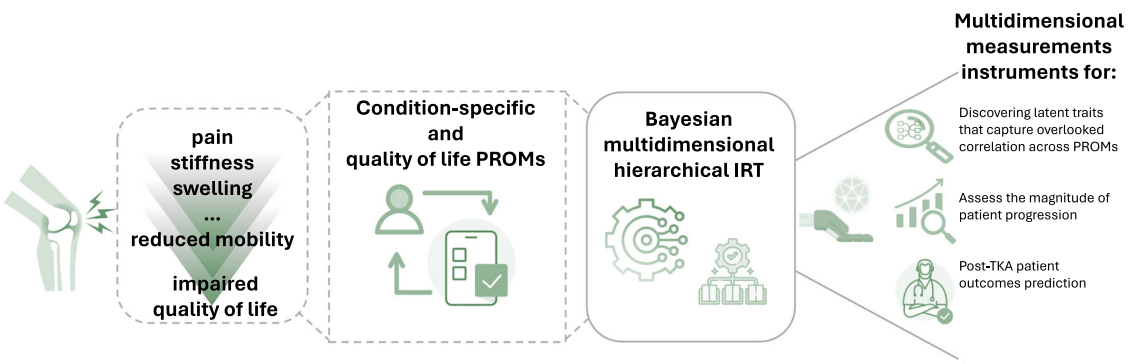

Fig. 1 | Schematic representation summarizing the objectives of this work. We employed a publicly available dataset comprising condition-specific and health-related quality of life PROMs to develop and validate a novel Bayesian multidimensional IRT (MHIRT) model to build instruments for assessing patient evolution and outcomes. We applied the model to an independent PROMs dataset collected via the moveUP platform to assess external validity.

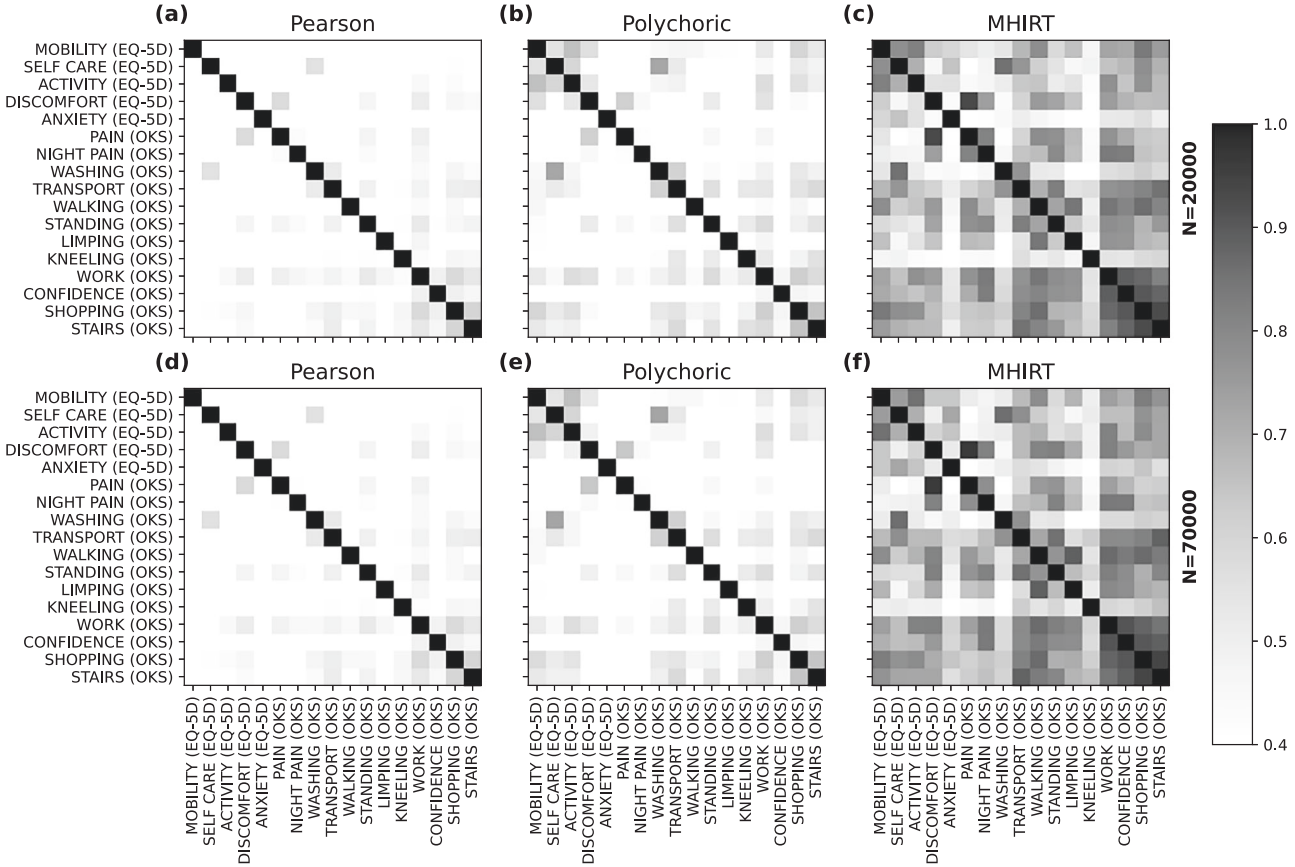

Fig. 2 | Correlation between items. Left (a, d): Pearson correlation coefficients. **Center (b, e):** Polychoric correlation. **Right (c, f):** MHIRT model covariance matrix. **Top (a, b, c):** Estimations for 20,000 random subjects. **Bottom (d, e, f):** Estimations for 70,000 random subjects. Grey levels encode correlation strength---black indicates a strong correlation, mid-grey a moderate correlation, and white weak or no correlation.

Our findings can be summarized as follows:

- **Identification of latent correlations across items and instruments**: Using the proposed MHIRT model, we identified consistent latent correlations among items from both OKS and EQ-5D-3L. These correlations, which are not directly captured by traditional pairwise correlation methods like Pearson and polychoric approaches, suggest shared constructs across instruments and support the use of multidimensional modeling in PROM-based assessment.
- **Enhanced sensitivity for measuring patient progress**: Compared to the aggregated OKS score, our model's factor analysis-derived traits capture more domain-specific variation, particularly in the dimension associated with pain, which demonstrated a larger effect size than OKS in both datasets. Additionally, effect sizes exceeded 1 across all extracted dimensions, indicating strong responsiveness to change. These findings suggest that the MHIRT model captures item-level variation that aligns with distinct aspects of recovery, such as pain and mobility.
- **Data-driven adaptability and flexibility**: Unlike traditional multidimensional IRT models that require predefined item-factor mappings, the MHIRT model infers both latent traits and their inter-item relationships directly from the observed response patterns. This adaptive structure allows the model to accommodate item groupings that may not align with assumed subscales, thereby reducing the risk of model misspecification. Moreover, the learned MHIRT model covariance matrix provides regularization that helps address factor indeterminacy, a known challenge in conventional MIRT models.
- **Improved predictive power in low-data scenarios**: Our experiments demonstrate that the latent traits derived from the MHIRT model achieved lower prediction error for post-TKA outcomes, especially in scenarios with limited training data. The linear regression model using MHIRT-derived features showed smaller prediction error across bootstrap samples than using traditional composite scores like OKS and EQ-5D-3L, or traditional IRT-based features. These results indicate that the model performs reliably in small-scale studies or clinical settings with limited data availability.
- **Potential for personalized and digital healthcare applications**: By integrating the IRT-based features derived from the proposed MHIRT model with linear regression modeling, our results suggest potential utility in developing data-driven tools that may inform personalized rehabilitation strategies. The external validation using the moveUP dataset demonstrated the model's ability to capture domain-specific recovery patterns in a digitally monitored cohort. This suggests that it could be effectively applied in digital healthcare delivery, facilitating data-driven evaluation of therapy outcomes.

While this study focuses on PROMs related to TKA, the modeling approach we apply may be transferable to other clinical areas where PROMs are commonly used, such as oncology, cardiovascular rehabilitation, chronic pain management, and mental health. By allowing latent trait structures to emerge from the data, the proposed method offers a framework that could support multidimensional assessments in a range of medical contexts for informing clinical decision-making and personalized medicine strategies.

Results

Model development and experimental setting

To evaluate the proposed MHIRT model, we designed an experimental setup utilizing the publicly available NHS Patient-Reported Outcomes Dataset, which includes data from individuals who underwent total knee arthroplasty (TKA) in England. The dataset provides pre- and post-TKA

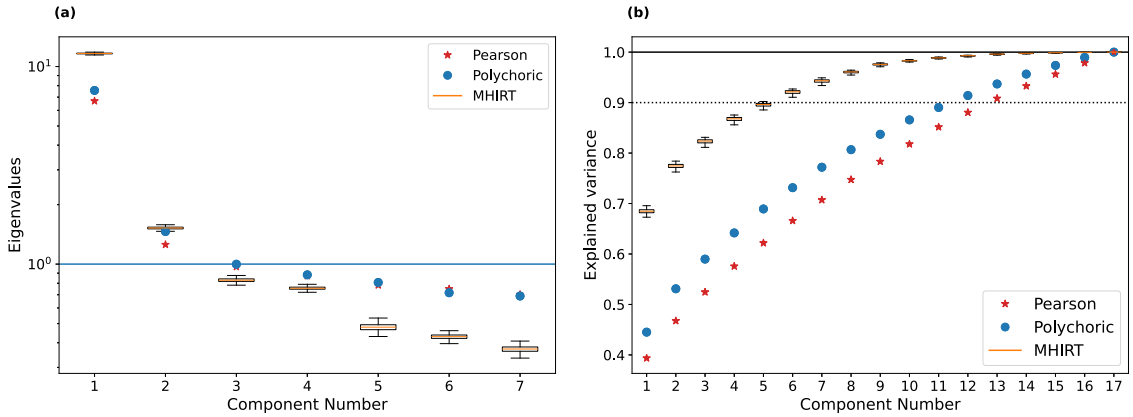

Fig. 3 | Eigendecomposition for the Pearson, polychoric and IRT correlation matrices. a Eigenvalues in logarithmic scale (Scree plot). **b** Cumulative explained variance. In both panels, separate curves correspond to the Pearson, polychoric, and MHIRT matrices, as indicated in the in-figure legend.

patient-reported outcomes using instruments such as the Oxford Knee Score (OKS), EuroQoL five-dimension three-level version (EQ-5D-3L), and EuroQol Visual Analog Scale (EQ-VAS), alongside demographic and clinical variables. To evaluate the model's robustness to sample size, we created two sub-datasets of different sizes from the NHS dataset by randomly sampling subjects: **Smaller Subset** with 20,000 subjects, and a **Larger Subset** with 70,000 subjects. To control for patient variability and reduce confounding, we incorporated covariates, such as gender, age, comorbidities (e.g., diabetes, heart disease, arthritis), and lifestyle factors (e.g., smoking status, employment, and education). These variables ensured a robust analysis and improved the interpretability of the latent traits.

To systematically evaluate the model's performance, we implemented the following steps:

- **MHIRT Model Fitting**: The proposed MHIRT model was trained on the NHS dataset sub-samples to uncover latent traits and correlations among questionnaire items. This step enabled the identification of patterns across PROMs instruments. (**Assessing IRT-based correlation structure**).
- **Factor Analysis**: Latent traits derived from the MHIRT model were subjected to factor analysis to extract interpretable dimensions (e.g., pain, mobility, self-care) and reduce data dimensionality (**Latent traits discovery through factor analysis**).
- **Effect Size**: We explored the effect sizes of latent traits derived from the MHIRT model. Specifically, we compared the sensitivity and magnitude of these latent traits to established measures like the OKS (**Effect size of latent traits**).

We also conducted an external validation study to evaluate the robustness and generalizability of the model when applied to independent data from a digital health monitoring platform following TKA (**External validation using the moveUP dataset**). We used the model developed on the Larger Subset of the NHS data to estimate the individual MHIRT latent vectors and derived traits in a cohort of 798 patients whose PROMs were collected via the moveUP platform.

Assessing IRT-based correlation structure

In order to investigate the capacity of the MHIRT model in distinguishing the underlying correlations among questionnaire items, and therefore to capture the multifactorial nature of patients' self-perceived QoL, we performed a comprehensive evaluation of the learned inter-item correlation patterns. Figure 2 illustrates the MHIRT model covariance matrix (right panels) estimated using the MHIRT-based method as defined in equation (1) of **Method**. For comparison, two classic alternatives are also presented: the Pearson correlation (left panels) and the polychoric correlation (center panels). These correlations were estimated directly from the raw responses. The top panels correspond to the smaller dataset previously mentioned, while the bottom panels correspond to the larger dataset. Despite the difference in dataset size, the correlation patterns exhibit very similar structures, which confirms the robustness of the estimations.

The Pearson correlation coefficient is not the optimal statistical tool for studying ordinal outcomes because of the underlying assumptions, such as interval measurement scales[24]. On the other hand, the polychoric correlation was explicitly defined to study categorical variables. It estimates the correlation between hypothesized normally distributed continuous latent variables associated with each outcome. It can be seen that the polychoric correlation (central panels of Fig. 2) reveals some additional associations between variables not detected by the Pearson correlation (left panels of Fig. 2), for example, the correlation between the EQ items.

The MHIRT model covariance matrix, shown on the right panels of Fig. 2, presents a much stronger correlation structure than Pearson or polychoric correlation. It is difficult to explain the reasons behind this correlation increase compared to the polychoric correlation and to foresee the effect of different hyperpriors. We used the Lewandowski-Kurowicka-Joe distribution[25], i.e., LKJ(1), which is equivalent to a uniform distribution over correlation matrices. With increasing dimensionality, the marginal distribution over the correlations concentrates around zero due to the complex constraints[26]. The LKJ prior is a weakly informative prior that works as a regularizing prior for correlation matrices. Without solid evidence from the data, the correlations will shrink toward zero and the correlation matrix toward the identity matrix. Here, the contrary is observed, suggesting that this correlation pattern reflects the true underlying correlations between items, which the polychoric correlation cannot capture effectively.

It can be noticed from Fig. 2 that, within the OKS items, Kneeling exhibited the weakest associations with all other variables in the MHIRT correlation matrix. Clinically, this is unsurprising: kneeling requires > 110° of knee flexion and many patients avoid it even when pain and basic ambulation have improved, moving only loosely related to the pain- and walking-dominated constructs captured by the remaining OKS items. The low correlations, therefore, do not necessarily indicate model misfit but highlight that the kneeling question taps a relatively independent aspect of functional demand.

A benefit of an increased correlation is that more variance is concentrated in fewer dimensions, eventually allowing the discovery of a few robust independent dimensions. Figure 3 shows the eigendecomposition of the correlation matrices for the three methods. In the case of MHIRT-based eigenvalues, the box plots summarize their posterior distribution. For all three methods, at least two independent factors can be derived according to the Kaiser criterion because the eigenvalues are higher than 1 (see Fig. 3a). However, this criterion has been criticized for being subjective and unreliable[27]. Consistent with the pain and function domains commonly reported for TKA PROMs, these first two factors account for the largest

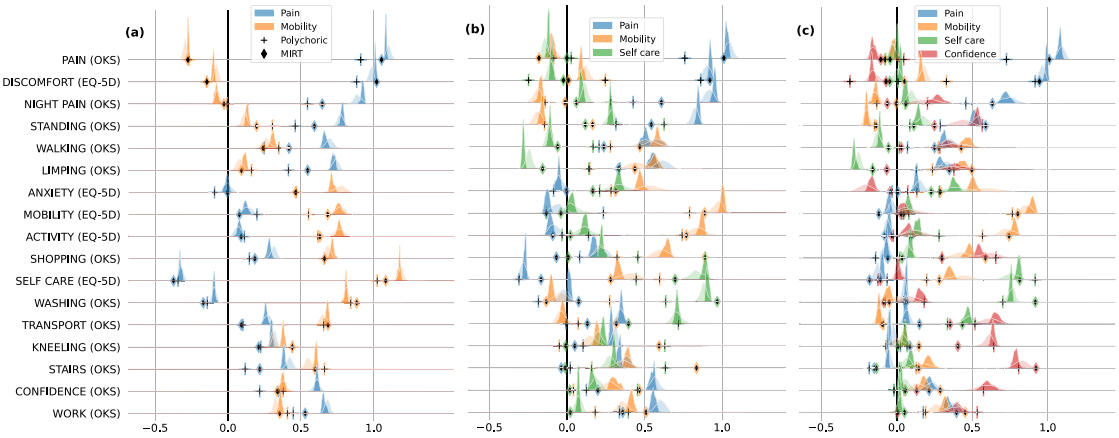

Fig. 4 | Factor Analysis (FA) of latent variables computed on each MCMC sample. The shadow areas represent the loading distributions for each factor, i.e., how much each factor contributes to each item. The FA loadings for the smaller (20,000) and larger (70,000) populations are depicted by lighter and darker areas, respectively. Panels (**a**), (**b**), and (**c**) display the FA computed for 2, 3, and 4 factors, respectively. The plus (+) and diamond marks on the horizontal axes represent the loadings for the FA estimated from the polychoric correlation and for the standard MIRT + FA approach[33], respectively. Each factor is labeled according to the common trait of the items exerting the largest influence on it, as displayed in the legend panels.

share of variance across all approaches. A substantial drop in energy after the fourth eigenvector is observed for the MHIRT model but not for Pearson or polychoric methods. This "second elbow" indicates that eigenvalues 3-5 remain clearly above the noise floor, signaling additional—albeit smaller—latent dimensions that are not captured by the classical correlations. These four eigenvectors concentrate significantly more variance in the MHIRT model than in the others, reaching almost 90% of the explained variance (see Fig. 3b). On this basis we retained up to 4 or 5 factors for subsequent analyses, as they capture clinically interpretable constructs such as self-care and confidence while still accounting for the vast majority of total variance.

Latent traits discovery through factor analysis

To investigate the ability of the approach to discover latent traits, say domains or sub-scales derived from item correlations across PROMs instruments rather than grouping them based on a defined structure, we build upon Factor Analysis (FA). Exploratory FA (EFA) was estimated with the Minimum Residual method with oblique rotation using the Promax method. An oblique rotation was used because they are more appropriate than orthogonal rotations when factors are not assumed to be independent. In particular, Promax is widely used for EFA as it is computationally efficient and provides a good balance between simplicity and interpretability[28,29]. Thereby, we estimated the FA decomposition directly from the set of latent vectors, i.e., $\{\boldsymbol{\theta}_i\}_{i=1}^{N}$ defined in **Method** to get a deeper insight into the dominant latent traits and to extract a low dimensionality representation to be used in downstream analysis or prediction tasks. As the factors were estimated based on the latent traits, we provided arbitrary names according to the common trait of the items exerting the largest influence. Interestingly, the trait relates to the International Classification of Functioning, Disability, and Health framework[30]. This framework is a common language for describing the level of function of a person. Health is divided into components, divided into a hierarchy of classification codes. Our approach brings additional insights, for example, into how standing is related to pain or how confidence, which is a mental function, is related to stairs or kneeling. Mental health is often assessed before TKA surgery through the measurement of anxiety, depression, or catastrophizing[31]. Confidence is another concept in our approach, as it is linked to several body movements, such as kneeling or stairs. This concept has been much more investigated after anterior cruciate ligament surgery, where the patients are expected to recover from sports activities and where a certain level of confidence is needed[32]. Considering confidence in further TKA research might help understand dissatisfaction after surgery, which is not related to physical components or usual mental health diseases.

Figure 4 shows the distribution of the factor weights after computing FA on each Markov Chain Monte Carlo (MCMC) sample and the two training populations. As expected, the uncertainty for the smaller population is larger. However, the factors are very similar across MCMC samples and training populations. The factors present a coherent pattern, with some items associated with a specific factor (e.g., pain with discomfort, mobility with activity, or self-care with washing) and other items belonging to many factors (e.g., limping or work). As more factors are considered, the initial factors, such as pain or mobility, maintain their primary attributes, while the additional factors describe more subtle traits.

Figure 4 also portrays the FA computed from the polychoric correlation matrix and with the classical MIRT approach using the `mirt` R package[33], both with the same parameters as previously described. Although a similar pattern is observed, some differences can be highlighted. The factors are more mixed in the polychoric correlation, as their loadings are not so extreme and are not so different between factors. This is not a limitation per se, but a deeper look into the polychoric correlation factors shows some inconsistencies, from which we argue that our model captures the latent traits better. For example, in our model, *night pain (OKS)* is more clearly assigned to the first factor, which we associate with *Pain*. *Washing (OKS)* and *self care (EQ-5D)* are highly correlated according to any of the estimated correlation matrices (see Fig. 2). However, the polychoric correlation FA with three factors fails to assign them to the same factor. MIRT factors are closer to the proposed MHIRT model than polychoric. The main differences are in the *confidence (OKS)* item.

The MHIRT model enabled fine-grained domains derivation by capturing the correlations and complementarities among items from condition-specific and more generalized health instruments. Thus, beyond the predefined subscales on traditional PROMs, MHIRT could attain an enriched representation of patient health status, allowing a practical digital solution to develop pre- and post-TKA patient assessment instruments and monitoring tools.

Effect size of latent traits

The effect size (or standardized mean difference) provides a measure of the magnitude of the difference or change. It is a critical tool in assessing the effectiveness of a treatment or intervention[34], understanding the practical significance of research findings (e.g., to quantify how much a patient's health status has changed), determining the sample size needed to detect an effect[35], and synthesizing evidence from multiple studies to inform healthcare decisions[36]. Thereby, to investigate the capacity of our model to be used as an analytical tool to assess the efficacy of therapies,

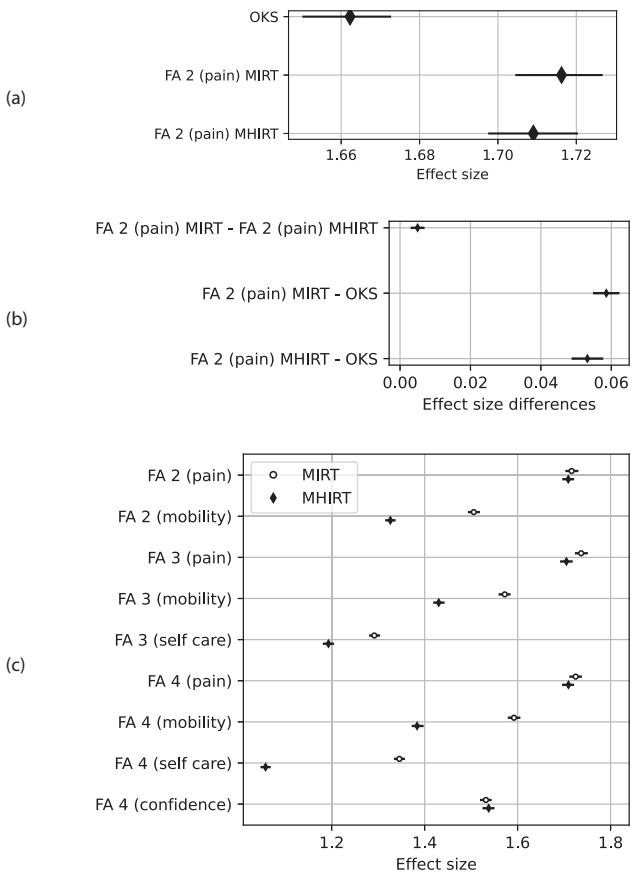

Fig. 5 | Estimated effect sizes (diamonds) and 95% confidence interval (estimated with 10,000 bootstrap sampling permutations and depicted as horizontal segments) on the NHS dataset. a Comparison of OKS and the pain trait. **b** Effect size difference between OKS and the pain trait. **c** Comparison of the traits obtained with factor analysis of different number of dimensions.

we compare the effect size of discovered latent traits to that of the OKS using the NHS dataset.

The top panel, **a**, of Fig. 5 shows the pre-post treatment effect size (ES) for the OKS and the ES of the MHIRT FA 2 (2-dimensional) latent trait associated with pain. Specifically, the ES was computed as the average score difference divided by its standard deviation. Composite scores such as OKS are designed to be highly sensitive to changes, summarizing the information in the most effective way to increase the instrument's power. On the other hand, the latent traits provided by factor analysis are multidimensional, which could limit the sensitivity of individual traits in exchange for a richer multifactorial assessment. However, it can be observed that the factor associated with pain has a larger effect size than OKS. This difference is statistically significant ($p < 0.0001$) and larger than 0.048 at a $p = 0.001$ significance level (see panel **b** of Fig. 5). EQ-5D index and EQ-VAS (not shown in the figure) had very low ESs, lower than 1.

Figure 5**c** illustrates the ES for each factor when increasing the number of factors. The trait associated with pain has the largest ES in all the cases, and its magnitude does not change with the number of factors. The second most sensitive trait is confidence, followed by mobility, while self-care is the least sensitive. Interestingly, confidence, while found to be sensitive in our approach, is a trait seldom used in TKA outcome measurement. Confidence is a trait more evident in cases where patients are expected to recover from sports activities, assuming they will need a certain level of therapy commitment to recover well[32]. Hence, the obtained results suggest that further TKA research might better understand dissatisfaction after surgery by considering confidence, which is not related to physical components or usual mental health diseases. In addition, these outcomes underscore the suitability of the MHIRT framework to design new measurement tools that allow more nuanced evaluations of traits. Overall, all the ESs portrayed in Fig. 5**c** are larger than 1 and the pain dimensions larger than OKS ($p < 0.0001$), which means that the sample size of a study to detect a 10% improvement on these scores at a 5% significance level and with a power = 0.8 would be approximately 1000 subjects for pain to 3000 subjects for self-care.

Progression prediction

In this section, we investigate the predictive utility of MHIRT latent vectors and the derived independent factors (MHIRT FA) to predict individual changes over a one-year follow-up period. In this set of experiments, we evaluated the performance of a linear regression model to predict changes in the OKS score from pre-TKA to post-TKA. Linear regression modeling was chosen for its relevance in medical studies due to its clinical interpretability and acceptance. It is the most widely used statistical tool in regression analysis and predictive modeling due to its robustness and interpretability properties.

To benchmark predictions from the linear regression model, we used the NHS regression model as a reference (baseline). The NHS provides predicted postoperative outcomes for each patient undergoing TKA, estimated from their validated, case-mix-adjusted regression equation[37]. This model was developed and refined iteratively over multiple large-scale NHS datasets, making it reliable for prediction within the NHS dataset. It incorporates preoperative OKS and EQ-5D PROMs, along with detailed patient characteristics such as age, gender, and comorbidities[38]. Thus, these NHS predictions, which are directly available in the dataset, offer a rigorous benchmark against which to compare the predictive value of IRT-based features.

In addition, to compare with previous works deriving IRT-based features from PROMs associated with TKA, we followed the protocol in ref. 39,40. Using the available implementation from[39], we employed the `mirt` R package[33] and fitted unidimensional graded response models[21] to the complete pre-TKA item sets of the OKS and the EQ-5D-3L separately to extract unidimensional IRT-based latent scores for every subject in the training set. Next, we estimated a MIRT model (also in `mirt` R package) whose structure reflected the 2-to-5 domains or sub-scales revealed by the MHIRT latent traits discovery through factor analysis. Therefore, five traditional IRT-based feature sets (derived from pre-TKA) were obtained: (i) 1D IRT OKS + 1D IRT EQ-5D[39,40], (ii) MIRT FA 2[33], (iii) MIRT FA 3[33], (iv) MIRT FA 4[33], and (v) MIRT FA 5[33]. Thus, enabling a strict like-for-like comparison of predictive performance between IRT-based feature sets extracted using conventional (unidimensional[39,40] and multidimensional[33]) IRT and the IRT-based feature extracted from the proposed MHIRT model across all sample size regimes using the baseline model from the NHS as a reference.

Finally, each IRT-based feature set was fed, in turn, as input (without covariates) into the linear regression model which was trained on multiple training set sizes (100, 160, 270, 440, 700, 1200, 2000, 3300, 5000, 9000, 15,000), chosen at approximately constant intervals on a log-scale to span small-to-large cohorts and examine how sample size affects performance. For each size, subjects were drawn with replacement from the NHS dataset, creating 100 independent bootstrap samples to quantify sampling variability. The model was refitted on each bootstrap sample, and its prediction error was re-estimated.

Figure 6 summarizes the RMSE distributions for OKS change prediction, reported as differences relative to the NHS regression model (baseline). For clarity, we present only a subset of training sizes where performance differences are most clearly illustrated, though the overall trend holds across all sizes. As expected, the linear regression model yielded more accurate results with larger training sets. Above 1200 subjects, both using the MHIRT latent vectors and independent factors (MHIRT FA 2 to MHIRT FA 5) obtained from our proposed model, approach the reference baseline. This suggests that no relevant information is lost from the OKS and

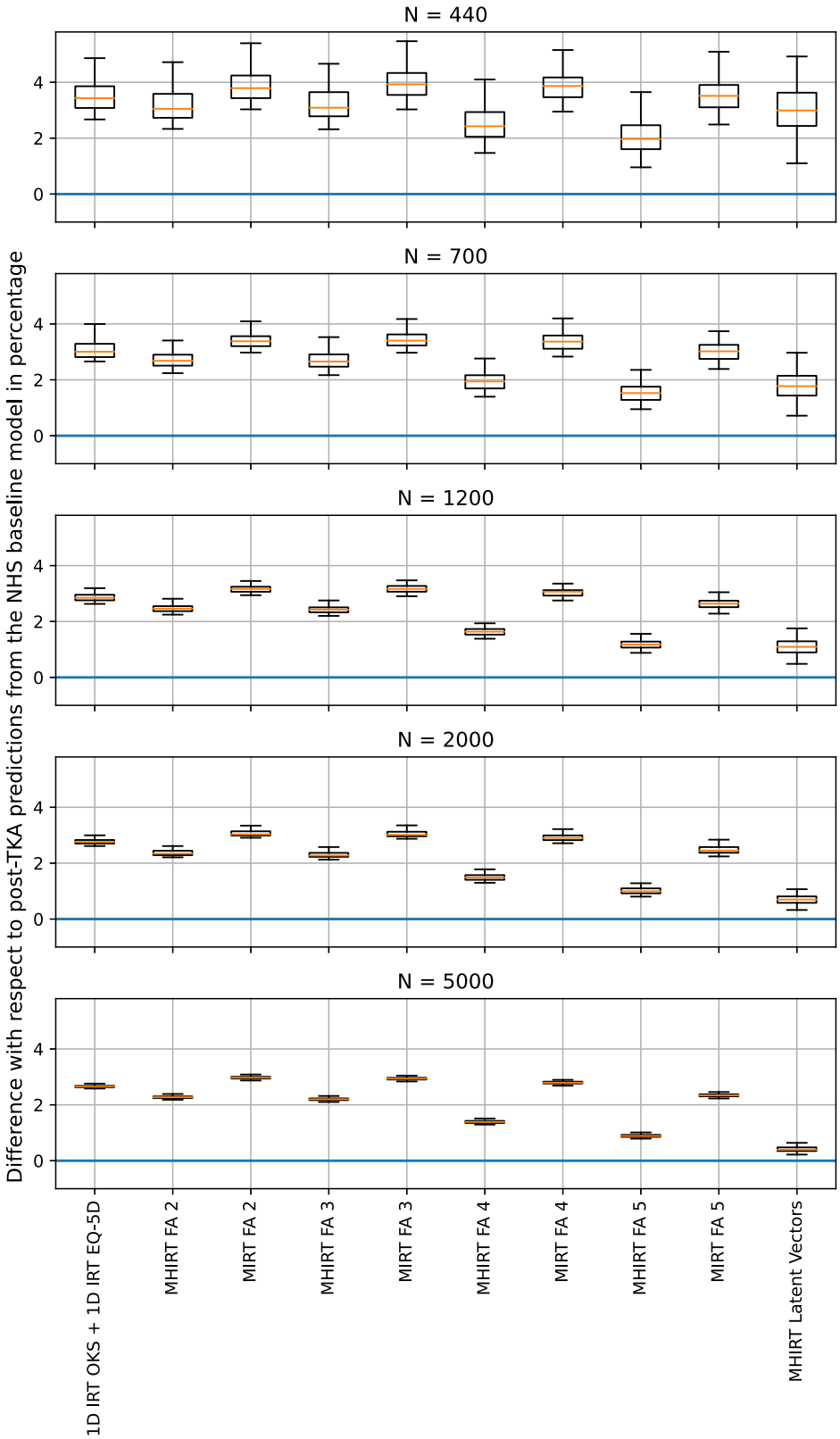

Fig. 6 | Performance of linear regression model to predict OKS change for different subsets of training set sizes across 100 bootstrap permutations. The boxes represent the interquartile range (Q3-Q1); the whiskers are the 5th and 95th percentiles. The *y* axis represents the difference with respect to post-TKA predictions from the NHS baseline model[37] in percentage: $100(RMSE_{\text{model}}/RMSE_{\text{baseline model}}-1)$. The blue line represents the post-TKA predictions from the NHS regression model used as baseline.

EQ-5D raw items to the IRT latent vectors and independent factors obtained from the MHIRT model.

One can see from Fig. 6 that for the smaller sample sizes (e.g., N=440), the linear regression model using the MHIRT FA 4 and MHIRT FA 5 independent factors as predictors outperformed traditional IRT-based features. Specifically, MHIRT FA 4 achieved a median RMSE that was ~ 1.8 percentage points lower than MIRT FA 4 and ~ 2.2 points lower than the composite 1D IRT OKS + 1D IRT EQ-5D. Similarly, MHIRT FA 5 outperformed MIRT FA 5 by ~ 1.4 percentage points, and also improved upon the 1D IRT OKS + 1D IRT EQ-5D by ~ 1.9 percentage points.

These results suggest that for studies with a small number of participants or observations, low-dimensional independent factors MHIRT FA 4

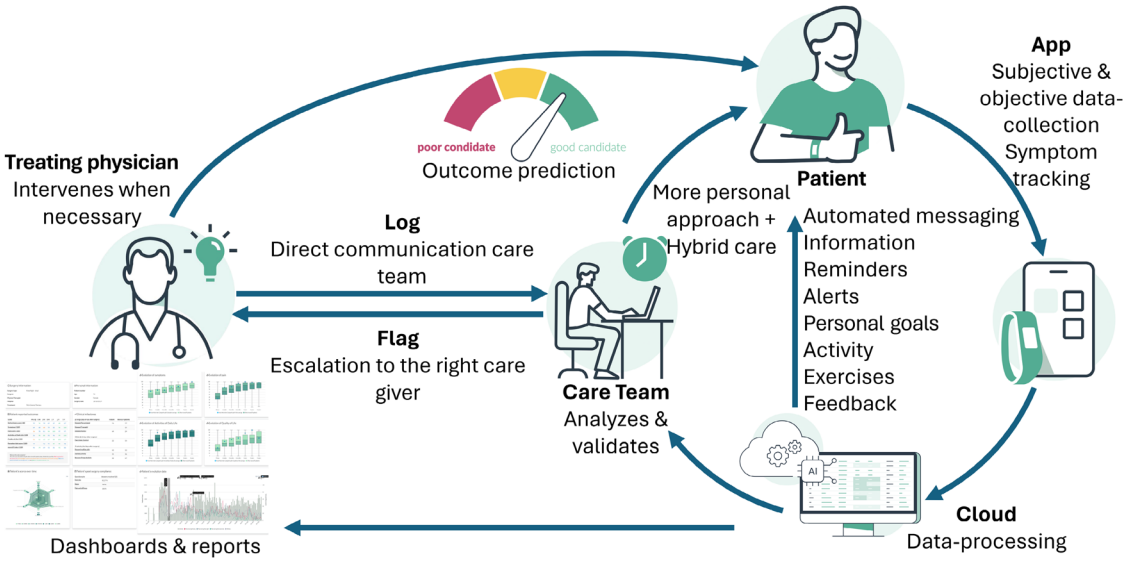

Fig. 7 | moveUP platform. This platform is an FDA approved medical device (Registration No. 3023739055; Product Code: ISD) classified as Class II (Regulation No. 890.5360, exempt). It is also CE-certified as a Class I medical device under the Medical Device Directive 93/42/EEC. The platform captures PROMs during the full patient journey.

and MHIRT FA 5 might summarize information with predictive utility and at the same time, are less prone to overfitting than the full set of MHIRT Latent Vectors (lower RMSE (up to 2.2 percentage points better than 1D IRT). Thereby, in practical digital scenarios with only small sample sizes, our IRT-derived subscales could serve as a pragmatic and suitable choice for predicting patient outcomes, and together with linear regression modeling, one can capitalize on their interpretability ability for shared decision making in digital clinical settings.

From Fig. 6 one can also observe that the linear regression model achieved significantly better performance ($p < 0.01$) using IRT-derived latent vectors and independent factors from the proposed framework, compared to traditional unidimensional (1D IRT OKS + 1D IRT EQ-5D[39]) and multidimensional (MIRT[33]) IRT-based features. These latent vectors represent normalized, regularized, and continuous transformations of PROM responses. Unlike unidimensional IRT graded response models[39], which summarize multiple items into a single latent trait, and traditional multidimensional IRT models[33], which require a substantial number of observations per factor and impose a strict upper limit on the number of factors to avoid overparameterization and numerical instability, our model flexibly associates one latent variable per individual outcome. This flexibility allows the latent vectors to automatically adapt their dimensionality based on correlations learned directly from data, rather than imposing dimensionality constraints a priori. Specifically, when inter-item correlations are low, the latent vectors resemble scaled versions of the original items. In contrast, high inter-item correlations lead to automatic dimensionality reduction, compressing the latent representation into a lower-dimensional space and potentially improving the signal-to-noise ratio. Overall, these results suggest that the independent factors derived from this framework offer a compact representation of latent traits, which may improve robustness and mitigate overfitting—particularly in studies with smaller sample sizes. In addition, for larger sample sizes (e.g., $N = 5000$), the MHIRT latent vectors closely match the predictive performance of well-validated regression models. This indicates that the proposed approach effectively retains relevant predictive information from the original PROM responses, supporting reliable outcome prediction.

Overall, these results indicate that MHIRT-derived features, particularly the MHIRT FA 4 and MHIRT FA 5 representations, yielded lower RMSE than the traditional IRT-based features and raw PROM scores. Their consistently lower RMSE values, combined with narrower bootstrap intervals (Fig. 6), suggest reduced risk of overfitting and greater stability in low-data contexts. Furthermore, the MHIRT-derived features exhibited lower prediction error compared to OKS and EQ-VAS, indicating that the model's multidimensional latent trait representations capture additional predictive information relevant to outcome forecasting.

External validation using the moveUP dataset

To evaluate the generalizability of the proposed multidimensional IRT framework beyond structured national datasets, we conducted an external validation using PROMs data collected via the moveUP platform, a certified digital medical device, that captures patient-reported outcomes along with physical activity and delivers remote rehabilitation services, including physical therapy, through digital technologies in the context of post-TKA recovery (see Fig. 7). The platform facilitates patients outcome tracking and remote monitoring[41,42].

Patients in the moveUP dataset either engaged with a fully digital follow-up protocol (cohort-1) or used the platform solely for PROM tracking while receiving standard in-person rehabilitation (cohort-2). Although group assignment was not randomized and subject to self-selection bias, the moveUP dataset offers valuable independent ground for testing model behavior under realistic variability. For instance, unlike the NHS dataset: (i) PROMs in the moveUP dataset were reported by patients through the moveUP application for digital health monitoring rather than questionnaires administered by hospital staff, (ii) patients in the moveUP dataset went into TKA under different surgery criteria as per different countries' clinical guidelines, and (iii) follow-up period after surgery covers the first 90 days rather than one year (an overview of the datasets is provided in **Method**). As such, our goals in this section are twofold: (1) to assess the robustness and interpretability of the MHIRT model on an external dataset (moveUP), independent of the NHS training data, and (2) to assess whether the model can differentiate recovery patterns between two rehabilitation strategies (cohort-1 vs. cohort-2).

We applied the MHIRT model, trained on the NHS dataset, to a retrospective cohort of 798 patients who used the moveUP platform. Importantly, this external validation allows us to assess whether the model's latent traits remain consistent when applied to PROMs collected in clinical settings that differ from those of the NHS data used for training. Figure 8 illustrates that several latent traits—particularly those related to pain, mobility, and self-care—exhibited measurable changes over a three-month

period in the moveUP dataset. Consistent with observations from the NHS dataset (see Fig. 5), the pain-associated trait demonstrated notable sensitivity to temporal progression. However, the overall effect sizes were smaller, potentially due to the shorter follow-up interval or increased variability in recovery patterns among patients in this digitally monitored setting.

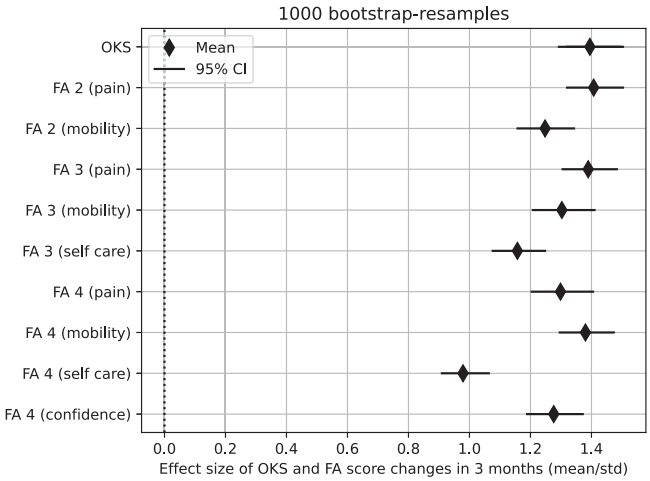

Fig. 8 | Effect sizes estimated on the moveUP dataset. Diamonds mark the estimated effect sizes and horizontal black lines denote the corresponding 95% confidence intervals. Estimates are derived from all 798 patients in the moveUP dataset.

Table 1 summarizes the results of linear regression analyses examining associations between patient characteristics and changes in outcome scores. While changes in OKS scores showed minimal correlation with demographic or biometric covariates, several FA-derived traits demonstrated significant associations. In particular, the pain and self-care traits were meaningfully linked to age and weight. For instance, improvement in self-care was positively associated with age (p-val = 0.007), and higher weight predicted greater improvement in pain-related scores (p-val = 0.03). These findings highlight the model's potential to support stratified recovery analysis and personalized rehabilitation planning based on patient-specific profiles.

To assess the model's sensitivity to detecting differences in recovery outcomes between cohort-1 (digital follow-up) and cohort-2 (standard in-person rehabilitation), we conducted two-sample t-tests on pre-to-post changes in the FA-derived traits and the OKS. This analysis evaluates whether the magnitude of improvement, differs between the two rehabilitation pathways. As shown in Fig. 9a, the OKS score revealed only a marginally significant difference between cohorts (p-val = 0.03). In contrast, Fig. 9b, shows that several FA-derived traits—most prominently pain (p-val = 0.001), but also mobility, self-care, and confidence—exhibited confidence intervals clearly further from the null (Fig. 9b). These results suggest that the FA-derived trait framework has the potential to highlight domain-specific aspects of recovery that are less clearly captured by the OKS composite score, particularly in exploratory comparisons of different care pathways.

The model captures changes in FA-derived traits such as mobility and confidence that may not be fully represented in standard PROM composite scores. For example, Fig. 9b, shows significantly greater improvement in confidence scores among patients in cohort-1 (digital follow-up), suggesting

Table 1 | Linear regression analysis

Outcome	Predictor	Coefficient	Std. Error	95% CI	p-value
OKS	Intercept	−5.3	13	[−31.4, 20.8]	0.69
	Age	0.03	0.06	[−0.08, 0.15]	0.58
	Gender	2.0	1.3	[−0.50, 4.55]	0.12
	Weight	0.075	0.04	[0.006, 0.14]	**0.03**
	Height	5.7	7.2	[−8.4, 19.8]	0.43
FA pain	Intercept	0.24	1.5	[−2.7, 3.2]	0.9
	Age	0.009	0.007	[−0.004, 0.02]	0.2
	Gender	0.04	0.14	[−0.2, 0.3]	0.8
	Weight	0.009	0.004	[0.001, 0.017]	**0.03**
	Height	−0.07	0.81	[−1.6, 1.5]	0.9
FA mobility	Intercept	0.59	1.9	[−3.1, 4.3]	0.8
	Age	0.002	0.008	[−0.01, 0.02]	0.8
	Gender	0.11	0.18	[−0.3, 0.5]	0.6
	Weight	0.008	0.005	[−0.002, 0.02]	0.1
	Height	0.31	1.0	[−1.7, 2.3]	0.8
FA self care	Intercept	−1.6	1.2	[−4.0, 0.7]	0.2
	Age	0.014	0.005	[0.004, 0.024]	**0.007**
	Gender	0.12	0.12	[−0.1, 0.3]	0.3
	Weight	0.008	0.003	[0.001, 0.014]	**0.02**
	Height	0.52	0.64	[−0.7, 1.8]	0.4
FA confidence	Intercept	−1.9	1.6	[−5.0, 1.2	0.2
	Age	0.012	0.007	[−0.001, 0.03]	0.08
	Gender	0.2	0.15	[−0.1, 0.5]	0.2
	Weight	0.005	0.004	[−0.003, 0.01]	0.2
	Height	1.24	0.85	[−0.4, 2.9]	0.1

Assessing the association of basal patients (all 798 patients from the moveUP dataset) characteristics (age, gender, weight, and height) with changes in OKS and MHIRT FA-derived traits across four factors. Bold p-values denote statistical significance ($p < 0.05$).

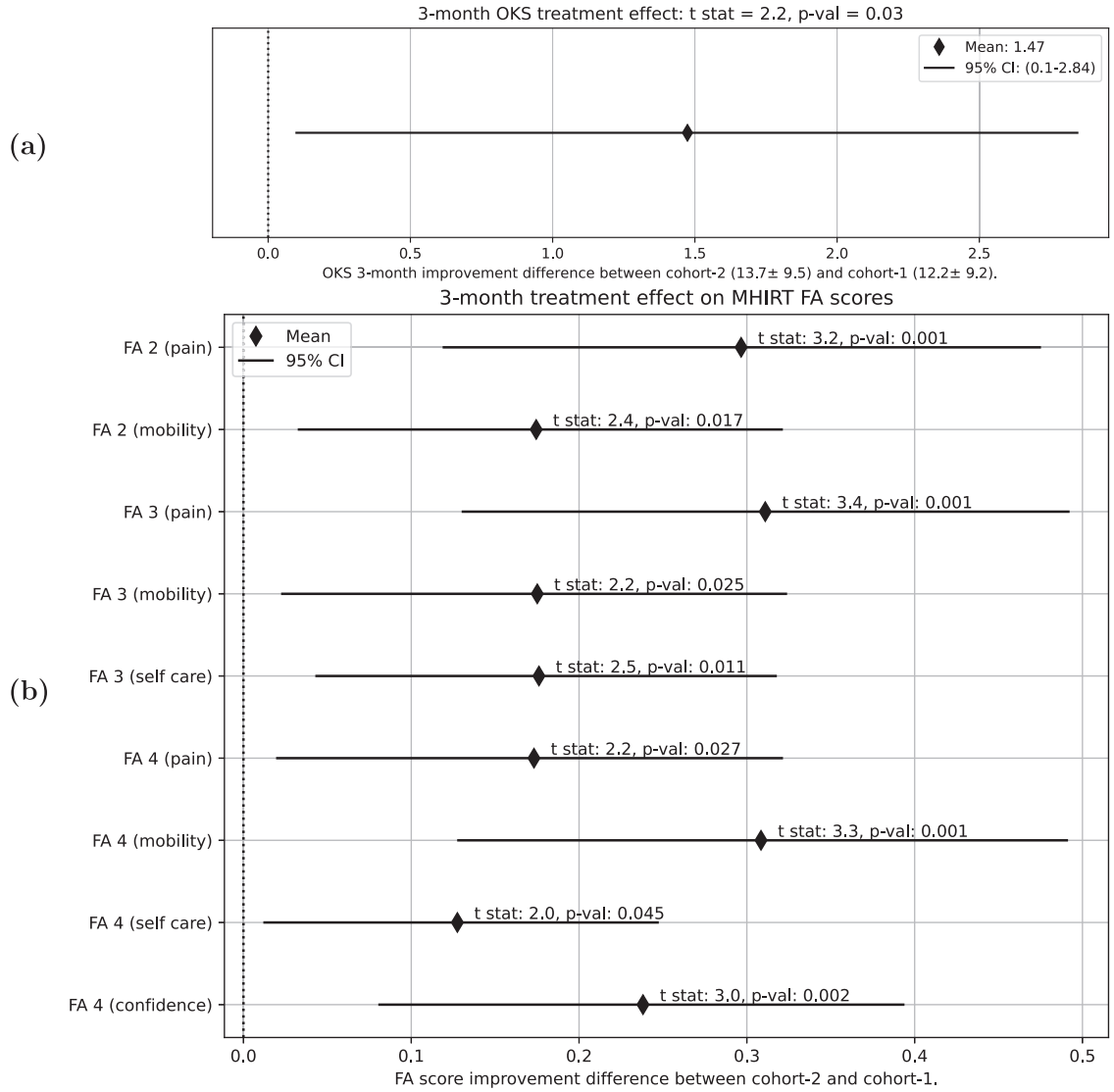

Fig. 9 | Efficacy of treatment administration on moveUP dataset. Estimated average improvement difference between the two cohorts (those who received the full digital therapy protocol, cohort-1, and those who followed standard in-person rehabilitation, cohort-2) after 3 months (diamonds) and 95% confidence interval (horizontal segments). **a** OKS. **b** Scores obtained with factor analysis of different numbers of dimensions.

that the model may uncover aspects of recovery related to self-efficacy and patient engagement, which are often overlooked in traditional assessments[43–45].

In summary, this external validation using the moveUP cohort supports the applicability of the MHIRT model across distinct populations and care delivery settings. The model, originally trained on structured registry data, performed consistently when applied to an independent, heterogeneous cohort, capturing clinically meaningful changes in FA-derived traits such as pain, mobility, self-care, and confidence (Figs. 8 and 9). Moreover, it revealed interpretable associations with patient-specific characteristics, including age and weight (Table 1), which were not evident using OKS scores.

Discussion

The interrelation between the EQ-5D and OKS instruments has been previously explored[46–48]. Yet, these studies did not examine item-level associations nor propose an integrated modeling framework that jointly analyzes both instruments. A more recent study[39] assessed the dimensionality of OKS using IRT and found it to be unidimensional, based on the Kaiser criterion applied to the polychoric correlation matrix. We replicated this result using a similar methodology on a subsample of 20000 patients, observing only one eigenvalue above one for OKS items.

However, when combining the five EQ-5D items with the twelve OKS items, we consistently identified more than two independent latent traits across all correlation estimation methods. This aligns with earlier evidence suggesting separable domains within OKS, such as pain and function[49], and highlights the added value of integrating multiple PROMs to reveal multidimensionality in recovery assessments.

As discussed in **Assessing IRT-based correlation structure**, the Kaiser criterion has notable limitations due to its subjectivity[27]. More compelling are the distinct eigenvalue patterns and improved variance concentration observed in our IRT-based correlation matrix (see Fig. 3), which better captures the underlying structure of PROM item interdependencies. The MHIRT model can support the decomposition of existing PROMs into multidimensional, clinically interpretable traits. The MHIRT framework leverages these structures to uncover latent traits—namely, pain, mobility, self-care, and confidence—which may help in understanding patient outcomes by capturing domain-specific variations often masked in composite scores[50]. These trait-level representations offer a structured means to monitor specific aspects of recovery, compare outcomes across patient

subgroups, and guide personalized care planning—without requiring changes to the original PROM instruments.

Unlike traditional MIRT models that rely on rigid subscale definitions and require substantial sample sizes, the proposed Bayesian approach adaptively learns the correlation structure among items and adjusts latent dimensionality accordingly[22]. This yields stable and interpretable factor representations, even in heterogeneous or smaller datasets.

To assess the robustness of these traits in predictive modeling, we compared MHIRT-derived features with both unidimensional (1D IRT OKS + 1D IRT EQ-5D[39]) and classical multidimensional IRT representations[33] using linear regression (see Fig. 6). Our model demonstrated lower variance and improved prediction accuracy over conventional IRT-based features, particularly in small-sample settings. These results indicate that the approach may be well-suited for outcome prediction in personalized digital health contexts where data availability is limited.

Application of the MHIRT model to the moveUP cohort provided external validation. As shown in Fig. 8, the MHIRT-derived traits—particularly those related to pain, mobility, and self-care—exhibited clear, domain-specific sensitivity to patient-reported changes over a three-month period. Compared to OKS, which showed only marginal between-group differences (see Fig. 9a), the trait-level representations derived from the MHIRT framework revealed more pronounced contrasts between cohort-1 (digital follow-up) and cohort-2 (standard care), with statistically significant differences in pain, mobility, self-care, and confidence (see Fig. 9b). Furthermore, linear regression analyses (see Table 1) revealed meaningful associations between patient characteristics and changes in FA-derived traits, such as improvements in pain correlating with weight and improvements in self-care with age, that were not evident using OKS. This suggests that MHIRT-derived traits may offer enhanced granularity for stratified analysis and personalized monitoring.

A particular finding is the emergence of a trait related to confidence, which showed significant differentiation between the two cohorts and may capture aspects of recovery tied to self-efficacy and perceived readiness to resume daily activities. Confidence, while rarely included in standard TKA PROMs, could offer valuable insight into patient engagement and satisfaction, especially for younger or more active populations[32]. Nonetheless, the analysis is limited by its reliance on the OKS and EQ-5D-3L instruments. The latter includes only a single item addressing mental health, which limits the ability to capture psychological complexity. Prior research has shown that single-item measures often lack the validity and reliability of multi-item instruments[51]. Future work could integrate PROMs like the Patient Health Questionnaire-9 (PHQ-9) for depression[52] or the Generalized Anxiety Disorder 7-item scale (GAD-7) for anxiety[53], to enhance trait specificity, although these are not commonly included in routine TKA workflows[54].

In addition, restricting the analysis to patients with complete follow-up PROMs inevitably introduces selection bias. Non-completers (i.e., patients with missing follow-up data) may differ systematically, particularly in the moveUP cohort, where 53.14% of patients were excluded compared to only 2.44% in the NHS cohort. As a result, the findings from the moveUP dataset are most representative of digitally engaged and adherent patients. Finally, because group assignment in the moveUP cohort was not randomized, selection bias cannot be ruled out. Hence, further evaluation in randomized controlled contexts is needed to confirm causal inferences.

The findings of this study have significant practical implications to inform clinical decision-making and outcome monitoring in TKA. By modeling PROMs through a data-driven, multidimensional latent trait framework, the proposed MHIRT model enables clinicians to derive more nuanced insights from established instruments such as the OKS and EQ-5D-3L, without requiring any changes to their structure. This multidimensional approach improves sensitivity by detecting domain-specific changes that may be masked in composite scores, such as improvements in pain without parallel gains in mobility, or psychological recovery (e.g., confidence) that traditional unidimensional scoring often overlooks. Specificity is enhanced by disentangling overlapping symptom domains, thereby reducing confounding between traits like self-care and mobility, which frequently co-vary in aggregated scores. For instance, Table 1 shows that self-care and mobility traits respond differently to demographic factors such as age and weight, and Fig. 9 illustrates that cohort-level improvements in these traits do not always occur in parallel, highlighting the clinical value of separating their contributions rather than aggregating them.

The model reveals heterogeneous recovery patterns by estimating individual scores across distinct traits such as pain, mobility, self-care, and confidence. This enables clinicians to tailor rehabilitation. For instance, a patient demonstrating functional gains but persistently low confidence or self-care ability may benefit from targeted balance training or psychosocial support. Such domain-specific insights go beyond conventional scoring and support personalized care. Notably, confidence, a trait often neglected in routine monitoring, emerged as clinically relevant and predictive of post-operative satisfaction, especially in younger or more active patients[55].

The model's generalizability across both structured (NHS) and independent (moveUP) datasets suggests that it may be viable for integration into digital health platforms, facilitating scalable and patient-centered outcome tracking. Aligned with ongoing efforts to embed PROMs into precision medicine[56,57], this framework provides a transferable and interpretable approach to assessing health-related quality of life. It offers a solution with potential applicability beyond orthopedics, including in oncology, cardiology, and chronic pain, domains in which multidimensional, patient-centered evaluation is increasingly emphasized, as reflected in emerging frameworks that integrate PROMs with clinical, behavioral, and biological data to inform precision care strategies[58].

Methods

The NHS patient reported outcomes dataset

To validate the efficacy of the proposed method, we first conducted experiments using publicly available PROMs data from a cohort of individuals who underwent TKA across multiple centers. This dataset, encompassing records from April 1, 2015, to March 31, 2017, was made publicly available by the England National Health Service (NHS), accessible online (https://digital.nhs.uk/). Subjects who underwent TKA were required to complete item-based questionnaires regarding their specific condition, indicated by the OKS (0 to 48 scoring system) instrument, and their overall health status, captured by the EQ-5D-3L instrument (EQ-5D-3L categories were reverse-coded to align their direction with OKS) and corresponding visual analog score (EQ-VAS), both before and following the surgical procedure. It is important to note that the EQ-5D-3L includes only a single item related to mental health ("anxiety/depression"), which inherently limits its ability to capture the complexity of mental health status[51]. Consequently, our findings related to mental well-being should be regarded as preliminary, and we recommend caution when interpreting mental health-related results. However, since this study focuses on overall well-being before and after TKA, rather than detailed mental health outcomes, the EQ-5D-3L remains an appropriate tool for assessing health-related QoL in this context. Therefore, to reflect standard TKA clinical assessments[54], our study employs OKS and EQ-5D-3L pre- and post-surgery item responses. In addition to item responses, the dataset also captures an array of variables for each patient, such as age, gender, and other comorbidities, as detailed in Table 2.

The 88,293 patients of the NHS cohort were used, as described in **Model development and experimental setting**, for MHIRT model fitting, correlation analysis, and latent traits discovery through factor analysis. However, for the effect size and progression prediction analyses, we required paired pre- and post-TKA questionnaires to ensure only data points with follow-up data. Only 86,143 patients (97.56% of the original NHS dataset), satisfied this criterion. The remaining 2150 patients were excluded in **Effect size of latent traits** (effect size) and 2.5 (progression prediction) due to lost to follow-up. While the excluded patients represent only 2.44% of the original NHS dataset, this might introduce potential selection bias and limit generalizability. Table 3 compares completers (patients with follow-up) and non-completers (patients lost to follow-up). The OKS pre-TKA showed a moderate imbalance Standardized Mean Difference (SMD = 0.28), indicating that patients lost to follow-up were more symptomatic pre-surgery.

As a result, complete-case effect size estimates may slightly overestimate recovery. We assumed that follow-up PROMs were missing at random. However, the observed imbalance in pre-TKA OKS suggests that some degree of missing not at random cannot be ruled out.

Retrospective data collection for external validation using the moveUP application

In this work, we used anonymized and depersonalized data obtained from the moveUP digital therapies database (moveUP solution, Brussels, Belgium https://www.moveup.care/ see **External validation using the moveUP dataset**), for externally validating the proposed MHIRT model in an independent post-TKA recovery setting. This dataset includes records of patients who underwent knee arthroplasty across Belgium, France, and the Netherlands. The analysis focused on a cohort of 798 patients who received elective TKA. The inclusion criteria required that patients actively use the digital application for at least 90 days following surgery and complete their PROMs three months post-TKA to guarantee follow-up data for all included participants. All patients provided written informed consent for the scientific use of their anonymized data, ensuring adherence to consent procedures. The study aligned with relevant regulatory guidelines and did not necessitate institutional review board approval due to the nature of the data used, namely anonymized patient-level data. PROMs were collected through the moveUP application, a medical device duly registered for digital health monitoring. This virtual platform integrates objective and subjective patient data and consists of a patient-oriented mobile application and a web-based dashboard for healthcare providers' use.

Table 4 details the demographic and clinical characteristics of the patient cohort, as well as pre- and post-TKA mean PROMs scores for OKS and EQ-VAS. Additionally, it summarizes comorbidities and other relevant variables. Comparing Table 2 (NHS) and Table 4 (moveUP), we observe that the digital therapy database consists of fewer patients in the age band above 80 years. This inclusion bias is known as elderly are less connected to technology[41,59]. Moreover, we also observe that the preoperative OKS in the

Table 2 | NHS dataset description

Observations	88,293
Female	50,875 (57.62%)
Age band (years)	
40 to 49	118 (0.13%)
50 to 59	8251 (9.35%)
60 to 69	31,999 (36.24%)
70 to 79	36,923 (41.82%)
80 to 89	10,996 (12.45%)
≥90	6 (0.01%)
Condition-specific patients' self-reported measures	
Mean OKS pre-TKA	19.12 (±7.72)
Mean OKS post-TKA	35.48 (±9.58)
General health status patients' self-reported measures	
Mean EQ-VAS pre-TKA	68.06 (±19.34)
Mean EQ-VAS post-TKA	70.58 (±24.35)
Covariates	
Heart disease	8629 (9.77%)
High blood pressure	40,074 (45.38%)
Stroke	1460 (1.65%)
Circulation	5523 (6.26%)
Lung disease	8331 (9.44%)
Diabetes	11,681 (13.23%)
Kidney disease	1781 (2.02%)
Nervous system	925 (1.05%)
Liver disease	466 (0.53%)
Cancer	4671 (5.29%)
Depression	7748 (8.78%)
Arthritis	68,129 (77.16%)

Details of the NHS cohort used in this study, including information on patients' demographic characteristics, PROMs, and covariates. Continuous variables are reported as mean ±SD; categorical variables as *n* (%).

Table 4 | Overview of the moveUP dataset

Observations	798
Female	482 (60.40%)
Age band (years)	
40 to 49	52 (6.52%)
50 to 59	233 (29.20%)
60 to 69	311 (38.97%)
70 to 79	178 (22.31%)
80 to 89	20 (2.51%)
≥90	-
Condition-specific patients' self-reported measures	
Mean OKS pre-TKA	24.35 (±8.33)
Mean OKS 3 months post-TKA	37.47 (±8.31)
General health status patients' self-reported measures	
Mean EQ-VAS pre-TKA	59.89 (±21.67)
Mean EQ-VAS 3 months post-TKA	68.07 (±23.05)
Covariates	
Arthritis	68 (50.75%)
Mean height	1.70 (±0.09)
Mean weight	85.42 (±16.29)

Details of the entire moveUP cohort used in this study, including information on patients' demographic characteristics, PROMs, and covariates. Continuous variables are reported as mean ±SD; categorical variables as *n* (%).

Table 3 | Comparison of completers (included patients) vs non-completers (lost to follow-up patients) in the NHS dataset

Variable	Completers (*n* = 86,143)	Non-completers (*n* = 2,150)	SMD
Female	49,565 (57.54%)	1310 (60.93%)	0.07
Age	65.67 (±8.29)	67.6 (±8.45)	0.06
Mean OKS pre-TKA	19.15 (±7.75)	16.90 (±8.35)	**0.28**
Mean EQ-VAS pre-TKA	67.87 (±20.35)	67.16 (±20.81)	0.03

Continuous variables are reported as mean ±SD; categorical variables as *n* (%). Mean OKS pre-TKA showed a Standardized Mean Difference (SMD) equal to 0.28 (highlighted in bold), suggesting a moderate imbalance.

Table 5 | Comparison of completers (included patients) vs non-completers (lost to follow-up patients) in the moveUP dataset

Variable	Completers (*n* = 798)	Non-completers (*n* = 905)	SMD
Female	482 (60.40%)	492 (54.36%)	0.11
Age	63.02 (±8.87)	63.61 (±10.78)	0.06
Mean OKS pre-TKA	24.35 (±8.33)	23.59 (±8.42)	0.09
Mean EQ-VAS pre-TKA	59.89 (±21.67)	60.32 (±22.03)	0.02

Continuous variables are reported as mean ±SD; categorical variables as *n* (%). Standardized Mean Differences (SMD) are small (SMD ≤ 0.11), suggesting that bias is likely small for moveUP.

digital therapy database is 5 points higher than the NHS database. There are differences in surgical criteria between countries and health systems. The NHS system is much more restricted than Belgium regarding arthroplasties (see e.g https://www.hweclinicalguidance.nhs.uk/clinical-policies/primary-knee-replacement. Belgium is the second country with the highest rate of arthroplasty (see e.g. https://www.oecd-ilibrary.org).

In the moveUP dataset, we included only participants who completed both pre- and post-TKA questionnaires to ensure complete data for analysis. This exclusion of patients without follow-up responses introduces a selection bias, meaning the results primarily reflect recovery patterns among patients who remained engaged in cohort-1 (digital follow-up) and cohort-2 (standard in-person rehabilitation). Specifically, 905 individuals with shorter or interrupted use of the moveUP platform (53.14% of the original 1,703 patients in the moveUP registry) were excluded due to loss to follow-up. Hence, outcomes for these excluded patients might differ systematically from those of the analyzed sample. Although we assumed that missing data in the moveUP dataset occurred at random, the high dropout rate raises the possibility that data may be missing not at random. Therefore, our findings should be interpreted as reflecting the recovery patterns of patients who remained engaged, rather than the broader moveUP population, summarized in terms of completers and non-comleters in Table 5.

Multidimensional hierarchical IRT model (MHIRT)

Multidimensional Item Response Theory (MIRT) models provide a robust framework for analyzing patient-reported outcomes by capturing relationships between questionnaire items and multiple latent traits[23]. These latent traits, representing unobserved abilities or characteristics, are linked to test items through discrimination parameters that measure the extent to which each trait influences the likelihood of a specific response. While MIRT models are powerful, they face several limitations that hinder their applicability in complex datasets like patient-reported outcomes measures (PROMs). One major challenge is factor indeterminacy, or the rotation problem, which arises when discrimination parameters are not predefined and must be estimated from data. This problem occurs because multiple parameter sets, related by a rotation, can fit the data equally well, making the interpretation of latent traits and parameter estimates challenging, particularly in high-dimensional or multifactorial datasets[60].

Additionally, current MIRT models rely on rigid, predefined dimensional structures to associate items with latent traits, limiting their ability to uncover complex interrelations in diverse datasets. They also struggle to handle redundancy or correlations among items, leading to less sensitive and precise assessments, especially for overlapping constructs. Furthermore, existing frameworks are not designed to adaptively model correlations across diverse dimensions or domains, reducing their generalizability and hindering their capacity to uncover latent traits that transcend predefined boundaries. These limitations underscore the need for more flexible, data-driven approaches to better capture the intricate relationships in PROMs data.

To address these limitations, we propose a new multidimensional hierarchical IRT model that operates as a full-rank framework to uncover latent traits and latent item-item correlation structure in PROMs data without relying on rigid predefined structures.

The key components of the framework are as follows. **Latent Variable Estimation**: The model estimates a multidimensional latent vector $\boldsymbol{\theta}_i$ for each subject, where each component represents a latent trait associated with a specific health-related dimension. The latent variables are modeled using a multivariate normal distribution with a shared covariance matrix (denoted as MHIRT model covariance matrix) that captures the interdependence among dimensions. **Covariate Integration**: Patient-specific covariates, such as age, gender, and comorbidities, are incorporated into the model to control for variability and bias, ensuring more accurate and personalized estimates of the latent traits. **Ordered Logistic Likelihoods**: Observed item responses are modeled using ordered logistic likelihoods parameterized by the latent traits and covariates. This approach ensures robust handling of ordinal data and provides interpretable thresholds for each item. **Dynamic Item Correlation Modeling**: The model includes a hierarchical structure that dynamically learns the correlation patterns between items, enabling the identification of interdependent latent traits and avoiding rigid predefined groupings. **Regularization Through Priors**: Weakly informative priors, such as the Lewandowski-Kurowicka-Joe (LKJ) prior for the covariance matrix, are used to regularize the latent variable estimates. The LKJ prior enforces constraints on the correlation matrix, encouraging sparsity or a shrinkage effect towards the identity matrix when there is insufficient data to estimate correlations robustly. This prevents overfitting and ensures that the model effectively captures meaningful patterns in the data.

The proposed framework combines Bayesian hierarchical modeling with MIRT principles to estimate latent traits and the underlying item-item correlation structure in PROMs data. This approach offers a data-driven alternative for exploring multidimensional constructs in patient-reported outcomes, which may be applicable across different medical contexts and support more interpretable clinical analysis.

To achieve this, we formalize the problem as follows: given a set of observed graded responses $\{Q_i^l\}_{i=1,l=1}^{N,L}$ from L items and N subjects, we want to estimate an L-dimensional latent vector θ_i for each subject i, in such a way that each component $\boldsymbol{\theta}_i^l$ of the latent vector is associated to the l-th item. Furthermore, we want to model the effect of a set of M covariates, $x_{i,m}$, on the response likelihood to control for these factors and reduce bias and inter-subject variability. We also want to remove the redundancy inherent in the selected subset of items to reduce the noise, increase the latent vector sensitivity to detect differences in changes, and learn the correlation patterns between items.

We propose a multidimensional hierarchical IRT model, where the item outcomes have ordered logistic priors, parameterized by a linear combination of the latent variables and the covariates. The latent variables are modeled with a multimodal distribution with a shared scale hyper-parameter, the covariance matrix.

Specifically, let $\boldsymbol{\theta}_i = [\theta_i^1; \ldots; \theta_i^L]$ be a normally distributed latent vector,

$$\boldsymbol{\theta}_i \sim \mathcal{N}(\mathbf{0}, \Sigma), \tag{1}$$

with Σ, the MHIRT model covariance matrix encodes inter-trait correlations, forming the core of the model's adaptability. It acts as a data-driven mechanism to learn inter-item and inter-trait structure. It regularizes latent variables, promotes interpretability, and avoids overfitting using an LKJ prior. This enables the model to adaptively identify multidimensional health traits from PROMs data, a capability that classical MIRT lacks without manual tuning or strong prior assumptions. Each component θ_i^l reflects the inner subject status regarding item l. We model the effect of patient characteristics and previous conditions on the likelihood of each item l outcome with fixed-effect terms as follows:

$$d_i^l = \theta_i^l \alpha^l + \sum_{m=1}^{M} x_{i,m} \beta_m^l,$$

where α^l represents the overall difficulty of item l, $x_{i,m}$ of the values of M covariates for subject i, and β_m^l is the effect of covariate m on item l.

The cumulative probability mass function of the l item is given by:

$$P(Q_i^l = 1) = 1 - \text{logit}^{-1}(d_i^l - c_1^l)$$
$$P(Q_i^l \geq k) = \text{logit}^{-1}(d_i^l - c_{k-1}^l), \quad \text{if } 1 < k \leq K,$$

where $\{c_k\}_{k=1}^{K-1}$ are thresholds for item l. This likelihood is called Ordered Logistic, and the probability for a specific value $k > 1$ is given by $P(Q_i^l = k) = P(Q_i^l \geq k) - P(Q_i^l \geq k+1)$.

The following priors were used for the model parameters and hyper-parameters:

$$\begin{aligned} \beta_m^l &\sim \mathcal{N}(0, 5) \\ \ln(\alpha^l) &\sim \mathcal{N}(0, 5) \\ \sigma_\alpha &\sim \text{half-Cauchy}(0, 5) \\ \Sigma &\sim \text{LKJ}(1), \end{aligned}$$

and the thresholds c_k^l have uniform priors and are restricted to be ordered, $c_1^l < \cdots < c_{K-1}^l$.

Inference of model parameters was made using the Stan software[61]. Stan is an implementation of a Hamiltonian Monte Carlo algorithm that efficiently explores the posterior distribution. Stan provides automated diagnostics to assess convergence and sampling reliability (e.g., R-hat, which evaluates chain mixing, or Effective Sample Size (ESS), confirming sufficient independent draws. These diagnostics confirm that the posterior is reliably characterized, and no computational issues arose. Standard MIRT and IRT analysis was performed in R (version 4.4.2) with the mirt package (version 1.44.0).

Data availability

The datasets analyzed during the current study are available in the England National Health Service (NHS) repository, https://digital.nhs.uk/. The retrospective data that support the findings of this study are available from moveUP digital therapies (moveUP solution, Brussels, Belgium. https://www.moveup.care/), but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are, however, available from the authors upon reasonable request and with permission from moveUP solution, Brussels.

Code availability

The underlying code for this study and training/validation splits is not publicly available but may be made available to qualified researchers on reasonable request from the corresponding authors.

Received: 20 January 2025; Accepted: 8 June 2025;
Published online: 01 July 2025

References

1. Briggs, A. M. et al. Musculoskeletal health conditions represent a global threat to healthy aging: a report for the 2015 world health organization world report on ageing and health. *Gerontologist* **56**, S243–S255 (2016).
2. Wong, A. Y., Samartzis, D. & Maher, C. The global burden of osteoarthritis: past and future perspectives. *Lancet Rheumatol.* **5**, e496–e497 (2023).
3. Grotle, M., Hagen, K. B., Natvig, B., Dahl, F. A. & Kvien, T. K. Prevalence and burden of osteoarthritis: results from a population survey in norway. *J. Rheumatol.* **35**, 677–684 (2008).
4. Robertsson, O. et al. Knee arthroplasty in denmark, norway and sweden: a pilot study from the nordic arthroplasty register association. *Acta Orthop.* **81**, 82–89 (2010).
5. Steinmetz, J. D. et al. Global, regional, and national burden of osteoarthritis, 1990–2020 and projections to 2050: a systematic analysis for the global burden of disease study 2021. *Lancet Rheumatol.* **5**, e508–e522 (2023).
6. Khatib, Y., Badge, H., Xuan, W., Naylor, J. M. & Harris, I. A. Patient satisfaction and perception of success after total knee arthroplasty are more strongly associated with patient factors and complications than surgical or anaesthetic factors. *Knee Surg. Sports Traumatol. Arthrosc.* **28**, 3156–3163 (2020).
7. Inui, H., Yamagami, R., Kono, K. & Kawaguchi, K. What are the causes of failure after total knee arthroplasty? *J. Jt. Surg. Res.* **1**, 32–40 (2023).
8. Demetriou, C. et al. Preoperative factors affecting the patient-reported outcome measures following total knee replacement: Socioeconomic factors and preoperative oks have a clinically meaningful effect. *J. Knee Surg.* **35**, 940–948 (2021).
9. Harris, K. et al. Systematic review of measurement properties of patient-reported outcome measures used in patients undergoing hip and knee arthroplasty.*Patient-Relat. Outcome Meas* **7**, 101–108 (2016).
10. Wang, Y. et al. Patient-reported outcome measures used in patients undergoing total knee arthroplasty. *Bone Jt. Res.* **10**, 203–217 (2021).
11. Sabah, S. A., Alvand, A., Beard, D. J. & Price, A. J. Evidence for the validity of a patient-based instrument for assessment of outcome after revision knee arthroplasty. *Bone Jt. J.* **103-B**, 627–634 (2021).
12. Lundgren-Nilsson, Å. et al. Patient-reported outcome measures in osteoarthritis: a systematic search and review of their use and psychometric properties. *RMD Open* **4** (2018).
13. Devlin, N., Parkin, D. & Janssen, B. *Methods for analysing and reporting EQ-5D data* (Springer Nature, 2020).
14. Shim, J. & Hamilton, D. F. Comparative responsiveness of the promis-10 global health and eq-5d questionnaires in patients undergoing total knee arthroplasty. *Bone Jt. J.* **101-B**, 832–837 (2019).
15. Lin, D.-Y. et al. Evaluation of the eq-5d-5l, eq-vas stand-alone component and oxford knee score in the australian knee arthroplasty population utilising minimally important difference, concurrent validity, predictive validity and responsiveness. *Health Qual. Life Outcomes* **21**, 41 (2023).
16. Kang, S. Assessing responsiveness of the eq-5d-3l, the oxford hip score, and the oxford knee score in the nhs patient-reported outcome measures. *J. Orthop. Surg. Res.* **16**, 18 (2021).
17. Mercieca-Bebber, R. et al. Design, implementation and reporting strategies to reduce the instance and impact of missing patient-reported outcome (pro) data: a systematic review. *BMJ Open* **6**, e010938 (2016).
18. Rasch, G. *Probabilistic models for some intelligence and attainment tests*. (ERIC, 1993).
19. Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. *Statistical theories of mental test scores* (1968).
20. Lord, F. M. & Novick, M. R. *Statistical theories of mental test scores* (IAP, 2008).
21. Samejima, F. Estimation of latent ability using a response pattern of graded scores 1. ETS Res. Bull. Ser. 1968, 1–169 (1968).
22. Morucci, M., Foster, M. J., Webster, K., Lee, S. J. & Siegel, D. A. Measurement That Matches Theory: Theory-Driven Identification in Item Response Theory Models. *Am. Polit. Sci. Rev.* **119**, 727–745 (2025).
23. Reckase, M. D. Multidimensional item response theory models. (Springer, New York, 2009).
24. Holgado-Tello, F. P., Chacón-Moscoso, S., Barbero-García, I. & Vila-Abad, E. Polychoric versus Pearson correlations in exploratory and confirmatory factor analysis of ordinal variables. *Qual. Quant.* **44**, 153–166 (2010).
25. Lewandowski, D., Kurowicka, D. & Joe, H. Generating random correlation matrices based on vines and extended onion method. *J. Multivar. Anal.* **100**, 1989–2001 (2009).
26. Barnard, J., McCulloch, R. & Meng, X.-L. Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Stat. Sin.* **10**, 1281–1311 (2000).

27. Fabrigar, L. R., Wegener, D. T., MacCallum, R. C. & Strahan, E. J. Evaluating the use of exploratory factor analysis in psychological research. *Psychol. Methods* **4**, 272–299 (1999).
28. Costello, A. B. & Osborne, J. Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis (2005).
29. Browne, M. W. An overview of analytic rotation in exploratory factor analysis. *Multivar. Behav. Res.* **36**, 111–150 (2001).
30. Organization, W. H. *International Classification of Functioning, Disability, and Health: Children & Youth Version: ICF-CY*. (World Health Organization, 2007).
31. Lopez-Olivo, M. A. et al. Psychosocial determinants of total knee arthroplasty outcomes two years after surgery. *ACR Open Rheumatol.* **2**, 573–581 (2020).
32. Sadeqi, M. et al. Progression of the psychological ACL-RSI score and return to sport after anterior cruciate ligament reconstruction: a prospective 2-year follow-up study from the French prospective anterior cruciate ligament reconstruction cohort study (FAST). *Orthop. J. Sports Med.* **6**, 2325967118812819 (2018).
33. Chalmers, R. P. Mirt: A multidimensional item response theory package for the R environment. *J. Stat. Softw.* **48**, 1–29 (2012).
34. Kazis, L. E., Anderson, J. J. & Meenan, R. F. Effect sizes for interpreting changes in health status. *Med. Care* **27** (1989).
35. Cohen, J. *Statistical power analysis for the behavioral sciences* (Routledge, 1988).
36. Hedges, L. V. Meta-analysis. *J. Educ. Stat.* **17**, 279–296 (1992).
37. NHS Digital. Patient reported outcome measures (PROMs) in England – data dictionary version 3.4. https://digital.nhs.uk/binaries/content/assets/website-assets/data-and-information/data-tools-and-services/data-services/proms/proms_data_dictionary.pdf (2016).
38. Huber, M., Kurz, C. & Leidl, R. Predicting patient-reported outcomes following hip and knee replacement surgery using supervised machine learning. *BMC Med. Inform. Decis. Mak.* **19**, 1–13 (2019).
39. Harrison, C. J. et al. Item response theory assumptions were adequately met by the Oxford Hip and Knee Scores. *J. Clin. Epidemiol.* **158**, 166–176 (2023).
40. Khatri, C. et al. Item response theory validation of the Oxford knee score and activity and participation questionnaire: a step toward a common metric. *J. Clin. Epidemiol.* **175**, 111515 (2024).
41. Lebleu, J. et al. Digital rehabilitation after knee arthroplasty: a multi-center prospective longitudinal cohort study. *J. Pers. Med.* **13**, 824 (2023).
42. Lebleu, J. et al. Incorporating wearable technology for enhanced rehabilitation monitoring after hip and knee replacement. *Sensors* **24**, 1163 (2024).
43. Zeni, J. A. et al. Rehabilitation exercises post-TKA: Impact on functional outcomes. *J. Arthroplast.* **25**, 220–226 (2010).
44. Kutzner, I. et al. Real-world activities and joint loading after TKA. *J. Orthop. Res.* **31**, 345–351 (2013).
45. Ghomrawi, H. M. K. et al. Patient confidence as a predictor of surgical success in TKA. *J. Arthroplast.* **35**, 961–968 (2020).
46. Dakin, H., Gray, A. & Murray, D. Mapping analyses to estimate EQ-5D utilities and responses based on Oxford knee score. *Qual. Life Res. : Int. J. Qual. Life Asp. Treat. Care Rehab.* **22**, 683–694 (2013).
47. Clement, N. D. et al. Mapping analysis to predict the associated EuroQol five-dimension three-level utility values from the Oxford Knee Score. *Bone Jt. Open* **3**, 573–581 (2022).
48. Fawaz, H., Yassine, O., Hammad, A., Bedwani, R. & Abu-Sheasha, G. Mapping of disease-specific Oxford knee score onto EQ-5D-5L utility index in knee osteoarthritis. *J. Orthop. Surg. Res.* **18**, 84 (2023).
49. Harris, K. 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.* **22**, 2561–2568 (2013).
50. Nelson, E. C., Eton, D. T., Revicki, D. A., Valderas, J. M. & Rumsfeld, J. S. Developing and interpreting patient-reported outcomes for clinical decision-making: a review of recent methods and applications. *Health Serv. Res.* **57**, 1067–1083 (2022).
51. Allen, M. S., Iliescu, D. & Greiff, S. Single item measures in psychological science: A call to action. *Eur. J. Psychol Assess.* **38**, 1–5 (2022).
52. Manea, L., Gilbody, S. & McMillan, D. A diagnostic meta-analysis of the patient health questionnaire-9 (PHQ-9) algorithm scoring method as a screen for depression. *Gen. Hosp. Psychiatry* **37**, 67–75 (2015).
53. Johnson, S. U., Ulvenes, P. G., Øktedalen, T. & Hoffart, A. Psychometric properties of the general anxiety disorder 7-item (gad-7) scale in a heterogeneous psychiatric sample. *Front. Psychol.* **10**, 1713 (2019).
54. Ramkumar, P., Harris, J. D. & Noble, P. Patient-reported outcome measures after total knee arthroplasty: a systematic review. *Bone Jt. Res.* **4**, 120–127 (2015).
55. Scott, C. E., MacDonald, D. J., Howie, C. R. & Biant, L. C. Predictors of satisfaction after total knee arthroplasty: A prospective cohort study. *J. Arthroplast.* **35**, 1764–1771 (2020).
56. Black, N. et al. Patient-reported outcomes: pathways to better health, better services, and better societies. *Qual. Life Res.* **25**, 1103–1112 (2016).
57. Faust, J. S., Mathew, J., Gottlieb, M. & Gottlieb, M. Humanising ai in healthcare: a matter of patient-reported outcomes. *Lancet Digit. Health* **4**, e766–e767 (2022).
58. Amoei, M. & Poenaru, D. Patient-centered data science: an integrative framework for evaluating and predicting clinical outcomes in the digital health era. *arXiv preprint* https://arxiv.org/abs/2408.02677 (2024).
59. Van der Vaart, R. et al. E-healthmonitor 2022. stand van zaken digitale zorg. *RIVM rapport 2022-0153* (2023).
60. Liu, T., Wang, C. & Xu, G. Estimating three- and four-parameter MIRT models with importance-weighted sampling enhanced variational auto-encoder. *Front. Psychol* **13**, 935419 (2022).
61. Stan Development Team. Stan Modeling Language Users Guide and Reference Manual, Version 2.29 https://mc-stan.org (2019).

Acknowledgements

This work was partially funded by INNOVIRIS (Brussels Capital Region, Belgium) under the projects: Augmented Intelligence in Orthopedics Treatments “ANTICIPATE” (BHG/2020-RDIR-6a) and Towards Data Driven Precision Medicine in Chronic Obstructive Pulmonary Disease “COPD-PROMPT” (BHG/2024-JRDIC-3b).

Author contributions

A.D.B. and M.B.: Conceptualization; methodology; software; formal analysis; investigation; validation; visualization; writing—original draft; writing—review & editing. J.L. and A.P.: Data curation; provision of independent data for external validation; interpretation of results; writing—review & editing. H.S.: Conceptualization; methodology; supervision; writing—original draft; writing—review & editing. All authors have read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Abel Díaz Berenguer or Matías Nicolás Bossa.