Can Short but Reliable Measures of Knee-specific Function be Constructed Using Item Response Theory?

Barbara Gandek, PhD, John E. Ware, PhD, Milena Anatchkova, PhD, Courtland Lewis, MD, Patricia Franklin, MD, MBA, MPH

Introduction: Patient-reported outcomes (PRO) data are important in evaluating total knee replacement (TKR), but knee- specific PRO measures need to be shorter for routine use in clinical care. To examine how knee-specific function might be measured using fewer questions, this study used item response theory (IRT) and computerized adaptive testing (CAT) methods to evaluate the 22-item Function in Activities of Daily Living (ADL) and Sport/Recreation measures from the Knee injury and Osteoarthritis Outcome Score (KOOS).

Methods: 1,179 randomly-selected TKR patients (mean age=66.1, 61% female) from five medical centers completed the KOOS before and 6 months after TKR. To represent a full range of functional states, one survey per patient (pre-TKR or post-TKR) was randomly selected for psychometric evaluation. IRT model assumptions of unidimensionality, item local independence and monotonicity were evaluated. All 22 function items were calibrated using the graded response IRT model. Item usage was evaluated with real-data CAT simulations.

Results: Analyses supported calibration of all function items. Full 22-item bank and simulated CAT scores correlated highly (r=0.96). Reliability of the full item bank was ≥0.95 (as recommended for individual patient monitoring) across a range from -2.5 standard deviations (SD) below to 1.7 SD above the combined pre-TKR/post-TKR mean. CAT scores with reliability ≥0.95 were achieved in 5-10 items for 96% of patients before TKR but only 66% post-TKR; patients for whom highly reliable CAT scores could not be estimated post-TKR generally had higher function. Eight ADL items accounted for the majority of CAT administrations; most Sport/Recreation items had low IRT information and were rarely selected by.

Conclusion: IRT and CAT methods allow knee-specific function to be measured efficiently but reliably at the individual patient and aggregate practice level, using only 5-10 questions. However, higher levels of function at six months in some TKR patients may not be reliably captured.