## Ratings of All Proposed Outcome Measures for TJAOM Toolkit

This table provides the consensus rating, across raters, for criteria relevant to potential inclusion of each option in the TJAOM Toolkit.

| **Performance-based tests** | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Face/ content for OA/TJA or both** | **Criterion (AUC ≥0.70) and/ or construct (>0.5)** | **Test-retest OR intra-rater** | **Inter-rater >0.8** | **Effect size or SRM (e.g., for ES, 0.1 = low, 0.3=medium, >0.5, large)** | **Floor or ceiling effects** | **Set up/scoring burden + Space, equipment, training** | **Safety, patient burden/ effort** | **SEM or MDC or MCID available** | **Norms/ref values available** | **Threshold/ cut points or PASS** | **In-person/ virtual** | **Instructions in other languages** | **OVERALL RATING** |
| **4mWT** | | | | | | | | | | | | | | |
| **6MWT** | | | | | | | | | | | | | | |
| **10mWT** | | | | | | | | | | | | | | |
| **30sCST** | | | | | | | | | | | | | | |
| **40mFPWT** | | | | | | | | | | | | | | |
| **BBS** | | | | | | | | | | | | | | |
| **BESTest** | | | | | | | | | | | | | | |
| **Brief-BESTest** | | | | | | | | | | | | | | |
| **Mini- BESTest** | | | | | | | | | | | | | | |
| **Floor Sitting-Rising Test** | | | | | | | | | | | | | | |
| **FRT** | | | | | | | | | | | | | | |
| **FSST or 4SST** | | | | | | | | | | | | | | |
| **L-test** | | | | | | | | | | | | | | |
| **SLS** | | | | | | | | | | | | | | |
| **SCT** | | | | | | | | | | | | | | |
| **ST** | | | | | | | | | | | | | | |
| **TUG or TUAG** | | | | | | | | | | | | | | |

| Patient-reported outcome measures (PROMs) | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Face/ content for OA/TJA or both | Criterion (AUC ≥0.70) and/ or construct (>0.5) | Internal Consistency (e.g., ICC (>0.9 for scores and >0.7 for subscales) | Test-retest (>0.9 for scores and >0.7 for subscales) | Effect size or SRM (e.g., for ES, 0.1 = low, 0.3=medium, >0.5 = large) | Floor or ceiling effects (< 15%) | Clinician burden | Respondent burden | SEM/ MDC/MCID available | Norms/ref values available | Threshold/ cut points Or PASS available | Pen/paper online, App | Translated, validated versions available | OVERALL RATING |
| EQ-5D-5L | | | | | | | | | | | | | | |
| FJS-12 | | | | | | | | | | | | | | |
| HOOS | | | | | | | | | | | | | | |
| HOOS- JR | | | | | | | | | | | | | | |
| KOOS | | | | | | | | | | | | | | |
| KOOS-JR | | | | | | | | | | | | | | |
| LEFS | | | | | | | | | | | | | | |
| NPRS | | | | | | | | | | | | | | |
| OKS | | | | | | | | | | | | | | |
| OHS | | | | | | | | | | | | | | . |
| PSFS | | | | | | NA | | | | NA | | | | |
| Pain VAS | | | | | | | | | | | | | | |

Abbreviations: 4mWT, 4m walk test; 10mWT, 10m walk test; 6MWT, 6 Minute Walk Test; 30 sec CST, 30 Second Chair Stand Test; 40mFPWT, 40m Fast-paced Walk Test; BESTest, Balance Evaluation Systems Test; BBS, Berg Balance Scale; Brief Balance Evaluation Systems Test; Brief-BESTest; EQ-5D-5L, EuroQoL-5D-5L; FJS-12, Forgotten Joint Score-12; FSST or 4SST, Four Square Step Test; FRT, Functional Reach Test; HOOS, Hip Disability and Osteoarthritis Outcome Score; HOOS-JR, Hip Disability and Osteoarthritis Outcome Score-Joint Replacement; KOOS: Knee Injury and Osteoarthritis Outcome Score; KOOS- JR, Knee injury and Osteoarthritis Outcome Score for Joint Replacement; LEFS, Lower Extremity Functional Scale; Mini-BESTest, Mini Balance Evaluation Systems Test; NPRS, Numeric Pain Rating Scale; OHS: Oxford Hip Score; OKS: Oxford Knee Scale; PSFS, Patient Specific Functional Scale; SCT, Stair Climb Test; ST, Step Test; TUG or TUAG, Timed Up and Go; Pain VAS, Pain Visual Analogue

Response options: Yes, green; Partially, yellow; No, red; No information, grey

Overall rating**:** Green, strongly recommend; Yellow, conditionally recommend; Red, don't recommend; Grey, no information available to ascertain whether the criteria is met

## Rating Guide

**Response options:**

For each outcome measure, enter in each column your rating according to the following scale:

- Yes – meets all the criteria (green)
- Partially – meets some of the criteria (yellow)
- No – does not meet any of the criteria (red)
- No information – no information available to ascertain whether the criteria is met (grey)

## Patient-reported outcome measures

- Overall Consideration: For each category there must be psychometric data for either pre-op TKA/THA, TKA or THA populations for a Green rating to be assigned (i.e. if there is only data on hip or knee OA, the best available rating would be a yellow).

**Validity:** The degree to which the instrument measures what it is supposed to measure.

- Face/ content for OA/TJA or both – write "TJA", "OA" or "both" that it measures what it is supposed to measure i.e. it measures hip function if it is a hip related OM
- Criterion - it has been compared to another OM which is considered to be a 'gold standard' and the number should be ≥0.70.
- Construct –it reasonably reflects the intended construct – i.e. the OM behaves in a way that makes theoretical sense and the number should be ≥ 0.5.

**Reliability**: The degree to which an instrument is free from random error

- Internal consistency e.g. ICC (>0.9 for scores and >0.7 for subscales): An intraclass correlation coefficient (ICC) is used to measure the reliability of ratings where there are two or more raters
- Test-retest (>0.9 for scores and >0.7 for subscales) a measure of reliability obtained by administering the same test twice over a period of time to the same person

**Responsiveness:** An instrument's ability to detect change over time

- Effect size or SRM: The amount of change over time necessary to indicate significant change (e.g., for ES, 0.1 = low, 0.3=medium, >0.5 = large; SRM >0.8 = large responsiveness, 0.5–0.8 = moderate, and 0.2–<0.5 = low)
- Floor / ceiling effects: a considerable proportion of subjects score the best/maximum or worst/minimum score, rendering the measure unable to discriminate between subjects at either extreme of the scale. An acceptable amount is < 15% i.e. > than 15% is unacceptable

**Feasibility**

- Clinician burden - time & effort to administer and/or score, equipment
- Respondent burden – time, effort to undertake

**Interpretability** The degree to which one can assign easily understood meaning to an instrument's quantitative scores

- SEM or MDC or MCID available
- Norms/ref values available – note, these values should ideally represent a Canadian or North American context. (i.e., Green rating if Canadian/ NA values available, otherwise Yellow would be the top rating).
- Threshold/ cut points or PASS available

**Alternative modes** (Pen/paper, online, App) Different ways that the OM can be administered

**Cultural / Language adaptations:** Adaptations have been made to address different cultures and/or languages. If translated to at least one other language AND validated in that language for patients with TKA/THA or Hip/knee OA, it would be a green; if translated to at least one other language and not validated in that language for a TKA/THA/knee or hip OA population it would be a yellow and if not available in any other language than English it would be a red.

## Performance measures

- Overall Consideration: For each category there must be psychometric data for either pre-op TKA/THA, TKA or THA populations for a Green rating to be assigned (i.e. if there is only data on hip or knee OA, the best available rating would be a yellow).

**Validity:** The degree to which the instrument measures what it is supposed to measure.

- Face/ content for OA/TJA or both – write "TJA", "OA" or both that it measures what it is supposed to measure – e.g. hip or knee
- Criterion (AUC ≥0.70) / compared to gold standard – it has been compared to another OM which is considered to be a 'gold standard'

**Reliability:** The degree to which an instrument is free from random error

- Test-retest / intra-rater: consistency of measures when undertaken on the same patient more than once by the same clinician
- Inter-rater: consistency of measures when undertaken on the same patient more than once by different clinicians

**Responsiveness** An instrument's ability to detect change over time

- Estimates of magnitude of change available (e.g.ES effect size, SRM standardized response mean). Effect size is low if the value of r varies around 0.1, medium if r varies around 0.3, and large if r varies > 0.5. An SRM >0.8 indicates large responsiveness, 0.5–0.8 moderate, and 0.2–<0.5 low.
- Floor/ceiling effects- a considerable proportion of subjects score the best/maximum or worst/minimum score, rendering the measure unable to discriminate between subjects at either extreme of the scale. An acceptable amount is < 15% i.e. > than 15% is unacceptable.

**Feasibility**

- Space, equipment, training
- Safety

**Interpretability** The degree to which one can assign easily understood meaning to an instrument's quantitative scores

- SEM or MDC or MCID available
- Norms/ref values available (i.e., Green rating if Canadian/ NA values available, otherwise Yellow would be the top rating).
- Threshold/ cut points Or PASS

**Alternative modes:** (Virtual, in-person) Has been evaluated both in-person and virtually

**Cultural / Language adaptations:** Adaptations have been made to address different cultures and/or languages

**Overall rating**: Please provide your opinion as to whether you think this OM should be considered by PT clinicians for this patient population. Please provide the reasons for your opinion.

- Green: Strongly recommended
- Yellow: Provisionally recommended
- Red: Not recommended
- Grey: Not enough info at this time to recommend