OMERACT®
Outcome Measures in Rheumatology

JUDGING THE PERFORMANCE OF THE INSTRUMENT BASED ON RESULTS FOUND IN THE STUDIES

Measurement property	**Provisional standards for adequate performance**
Construct validity	Pre-specified hypotheses are met. Should be shown with similar constructs, dissimilar constructs and across known groups to show both presence and absence of a relationship as appropriate for each domain of interest.
Inter-method reliability	Intra-class correlation coefficient (ICC); weighted Kappa coefficient (Kw) Excellent > 0.90. Good >0.75 (considered the threshold for adequate performance and for a positive rating) Excellent needed for measurement if done for individual clinical decision making.
Test-retest reliability	Intra-class correlation coefficient (ICC); weighted Kappa coefficient (Kw) Excellent > 0.90. Good >0.75 (considered adequate for a positive rating) Excellent needed for measurement if done for individual clinical decision making.
Longitudinal construct validity	Consistency with a priori hypothesis of the magnitude and direction of change that should be seen in that situation of change. If a priori hypothesis suggests a large effect should be observed, one should see an effect size or standardized response mean of >0.80. If moderate is expected, look for 0.5-0.79, small effect 0.2-0.5. Consistency with a priori theory of direction and magnitude of change would be given a positive finding. Findings outside the anticipated range should be considered a negative finding.
Clinical trial discrimination (Sensitivity in clinical trials)	**Gold**: Randomized groups demonstrate change in their scores congruent with anticipated effect of the study. **Silver**: Two group comparison (not randomized) are compared and differences in their change scores are congruent with anticipated results. **Bronze**: Longitudinal data are provided for the groups that have changed and separately for groups that have remained stable or had a different amount of change compared to the first group. SRM/ES/T test is greater in change group than in stable group, or group expected to have smaller change. This relative difference is aligned with expected difference in the change experienced in each arm/group. Ratio of effect size statistics squared is also a way of articulating the relative responsiveness of one measure over another. $(ES_{group1}^2/ES_{group2}^2)$. If reporting on % exceeding a threshold of meaning (i.e., response criteria), should report proportions for each group. Results should show a logical, significant relationship to the a priori hypotheses and expectations for the relative difference in the change experienced in the two groups.
Thresholds of meaning	There are no "standards" for the value of a calculated threshold. We expect that reporting and context be as clear as possible for users [32] and matched to the intended application. • '+' = the chosen method was clearly described and results reported and is in a similar context of use (population, setting). • '±' means results were derived only from distribution-based methods (ie, 1 SEM or ½ SD) but in an otherwise similar context. **Points to consider:** • Thresholds are dependent on the anchors used and should be reported and interpreted in that context (i.e., threshold for identifying levels of disease activity), and with sensitivity and specificity of the cut point provided. • For change thresholds, describe relation of both minimal important difference (MID) and minimal detectable change (MDC) and guide interpretation accordingly. Both must be exceeded to be confident in the threshold of change. • Congruence across multiple anchors will bring confidence in the meaning of a threshold score. Difference between results from multiple anchors can be shown using empirical cumulative distribution functions.