

Appendix

Online materials

Leung et al. **OMERACT Filter 2.1 Instrument Selection for Physical Function Domain in Psoriatic Arthritis: Provisional Endorsement for HAQ-DI and SF-36 PF.**

Table of Contents

Provisional standards for adequate performance of instruments, OMERACT Filter 2.1 2

Evidence of measurement properties for HAQ-DI 3

 Table A.1. Reporting of Evidence of Construct Validity for HAQ-DI in PsA with OMERACT Filter 2.1. 3

 Table A.2. Report of Studies of Test-Retest Reliability for HAQ-DI in PsA with OMERACT Filter 2.1..... 9

 Table A.3. Reporting of Evidence of Responsiveness (Longitudinal Construct validity) for HAQ-DI in PsA with OMERACT Filter 2.1 11

 Table A.4. Reporting of Evidence of Clinical Trial Responsiveness for HAQ-DI in PsA with OMERACT Filter 2.1 15

 Table A.5. Reporting of Evidence of Threshold of meaning for HAQ-DI in PsA with OMERACT Filter 2.1. 19

Evidence of measurement properties for SF-36 Physical functioning domain 24

 Table B.1. Reporting of Evidence of Construct Validity for SF-36 PF in PsA with OMERACT Filter 2.1. 24

 Table B.2. Report of Studies of Test-Retest Reliability for SF-36 PF in PsA with OMERACT Filter 2.1 27

 Table B.3. Reporting of Evidence of Responsiveness (Longitudinal Construct validity) for SF-36 PF in PsA with OMERACT Filter 2.1..... 28

 Table B.4. Reporting of Evidence of Responsiveness (Clinical Trial Discrimination) for SF-36 PF in PsA with OMERACT Filter 2.1 30

 Table B.5. Reporting of Evidence of Threshold of meaning for SF-36 PF in PsA with OMERACT Filter 2.1. 31

Provisional standards for adequate performance of instruments, OMERACT Filter 2.1

Pillar (and Question)	Measurement property	OMERACT Filter 2.1 Provisional standards for adequate performance
Truth. (Question 3. Do the numeric scores make sense?)	Internal consistency	Alpha >0.75, higher if target application is individual clinical decision making (0.90).
	Construct validity	Pre-specified hypotheses are replicated. Should be shown with similar constructs, dissimilar constructs and known groups in order to show both presence and absence of a relationship as appropriate.
Discrimination (Question 4: Can it discriminate between groups of interest?)	Test retest reliability	Excellent > 0.90. Good >0.75 (considered adequate for a green). ICC, Kw Excellent needed for measurement if done for individual clinical decision making. Please also report on SEdiff and MDC-95, Bland-Altman graph is helpful.
	Longitudinal construct validity	Consistency with a priori theory in studies that look at situation similar to the intended application. Anticipated large effect expect SRM >0.80, medium/moderate effect, SRM 0.5-0.79, small effect 0.2-0.5. Findings outside the anticipated range should be considered a negative finding.
	Sensitivity in clinical trials	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 is greater in change group than in stable or different change group. This difference is also reported in a relative effectiveness statistic ($ES_{\text{group1}}^2/ES_{\text{group2}}^2$) = hypothesized magnitude and direction.
	Thresholds of meaning	There are not “standards” for a calculated threshold. We ask only that reporting and context be as clear as possible for users. Report threshold value and how it was calculated, error boundaries if possible. Thresholds should be related to the anchors used (i.e., threshold for predicting disease activity), sensitivity and specificity of the cut point. For change thresholds, describe relation of both MID and MDC and guide interpretation accordingly.

Extracted from Boers M, Kirwan JR, Tugwell P, et al. The OMERACT Handbook. Online, available at: <https://omeracthandbook.org/handbook>

Evidence of measurement properties for HAQ-DI

Table A.1. Reporting of Evidence of Construct Validity for HAQ-DI in PsA with OMERACT Filter 2.1.

Truth: 8 articles (Blackmore 1995, Taylor 2007, Taccari 1998, Leung 2008, Brodsky 2010, Katchamart 2014, Leung 2020, Wan 2020)

Author Year	Study characteristics		Methods and results used to assess construct validity. Two categories provided, correlational testing and known group comparisons. Both need brief description of a priori hypotheses (including expected results and how they were defined) and the results found and what the authors synthesis (was that good?) Look for strong associations/differences and the absence of the same (no correlation, no differences expected and none seen)				
	Sample description (mean age, % gender, disease type/severity/duration)	Study design/ methods	Correlation with other measures		Differences in scores across groups known to be different		
			A priori hypothesis (described expected results, and comparison instrument)	Results	A priori hypotheses of differences between the groups	Results	Judgement adequacy (+, +/-, -)
Blackmore 1995	<ul style="list-style-type: none">• 114 PsA patients who fulfilled the Moll and Wright criteria• 70 men, 44 women• Seen at the University of Toronto• Mean (SD) age 49.3 (13.2) years• Mean (SD) duration of illness 15.1 (9.6) years• Mean (SD) HAQ 0.50 (0.58)	<p>The HAQ-DI/ HAQ-S measure disability</p> <p>Comparison:</p> <ul style="list-style-type: none">• Measures of disease activity: morning stiffness, tender/ swollen joint number, ESR, PASI• Clinical measure of function: grip strength, ACR functional class, finger to floor	<p>HAQ-DI/ HAQ-S would correlate highly with clinical measures of function and deformities</p> <p>Definition of correlations: poor: <0.25 moderate: ≤0.25 to <0.40 high: ≥0.40</p>	<ul style="list-style-type: none">• HAQ-DI correlated highly with clinical measures of function including grip strength (r=-0.63), ACR functional class (r=0.59).• HAQ-DI correlated poorly with chest expansion and finger to floor distance (r=-0.16 to 0.25)• number of fibrositic TP (r=0.54) – not hypothesized• HAQ-DI correlated poorly with measures of disease activity (morning stiffness, total number of joint effusions, ESR and PASI (r ranges 0.04 to 0.36), but correlated highly with total	<p>HAQ-DI and HAQ-S scores would be higher in patients with spondyloarthropath y or fibromyalgia than patients without</p>	<ul style="list-style-type: none">• HAQ-DI was higher in patients with spondyloarthropathy than in patients without spinal disease as expected, but not significantly different. The mean (SD) HAQ scores were 0.61 (0.64) vs. 0.49 (0.56) (t=-1.12, df=112, p=0.26) in patients with and without spondyloarthropathy• Patients with fibromyalgia had significantly higher HAQ-DI, [mean (SD)=1.32 (0.49) vs. 0.43 (0.52) (t=-6.44, df=112, p=0.0001) than those without	<p>(+/-)</p> <p>Working group remarks: HAQ-DI performed ~50% of what authors hypothesized. However, the direction and magnitude of correlations, and distinguishing between groups aligned with conceptual framework.</p> <p>The working group acknowledge the authors proposed interpretations for high correlation (>0.4) was not adequately high with current standards. However, the final results of correlation between grip strength and ACR functional class fall in the</p>

		distance, Chest expansion • Pain: fibrositic tender points (TP)		number of active joints (r=0.49)			acceptable ranges (0.59- 0.63)
Taylor 2007	<ul style="list-style-type: none"> • 134 patients with clinic diagnosed PsA in New Zealand • 43% women • Mean age 52.3 (14.1) years • Median (IQR) disease duration 12.3 (6.3-21.1) years • Mean (SD) HAQ-DI 0.50 (0.59) • Mean (SD) SF-36 PF 60.4 (27.1) 	<p>HAQ-DI measures functional ability</p> <p>Comparison: SF-36 PF; the physical function domain measures functional ability</p>	Fit of HAQ-DI and SF-36 PF to the Rasch model, including unidimensionality, linear model, free of misfit items, no differential item functioning (DIF)	<ul style="list-style-type: none"> • Both HAQ-DI and SF-36 PF fit into Rasch model • Showed HAQ-DI and SF-36 were unidimensional, and measured the same construct of physical disability • SF-36 PF had fewer misfitting items, better distribution, better scale length, less DIF than HAQ-DI. • Floor effect of HAQ-DI: 30.4% vs. SF-36 PF: 3.1% 	<ul style="list-style-type: none"> • Both HAQ-DI and SF-36 PF fit the Rasch model analysis, indicative of measuring similar construct. • SF-36 PF have slightly better measurement span, separation between items, and less floor effects. 	<ul style="list-style-type: none"> • Both met criteria of fit to Rasch model analysis • Bland and Altman plot indicated significant relationship between the mean of both scores 	<p>(+)</p> <p>Remark: Overall solid support for structure validity of HAQ-DI and SF-36 PF</p> <p>Floor effect of HAQ-DI may affect interpretability of the score, but not overall construct validity.</p>
Taccari 1998	<ul style="list-style-type: none"> • 72 patients with PsA (both arthritis and PsO, and neg rheumatoid factor) in Italy • 30.6% women • Mean (SD) age 55 (12.6) years • Mean (SD) arthritis duration: 	<p>HAQ-DI</p> <p>Comparison:</p> <ul style="list-style-type: none"> • AIMS physical function • Disease activity: morning stiffness, number of painful joints, Ritchie's articular 	<ul style="list-style-type: none"> • HAQ-DI and AIMS physical function scale closely correlated • HAQ-DI moderately correlated with disease activity • HAQ-DI moderately correlated with disease severity (by erosion, sacroiliitis, 	<ul style="list-style-type: none"> • HAQ-DI and AIMS physical function highly correlated with each other (r=0.747, p<0.0001) • HAQ-DI had low to moderate correlation with disease activity (range of r=0.162 to 0.496), highly correlated with axial stiffness (r=0.724) • HAQ-DI had low correlation with radiographic severity (r=0.089 to 0.286) 	None hypothesized	• NA	<p>(+/-)</p> <p>Working group remark:</p> <ul style="list-style-type: none"> • Authors in this paper hypothesized moderate correlations between HAQ-DI with disease activity and radiographic severity, which was too optimistic • The final results actually showed correlations with disease activity and severity were lower than hypothesized

	11.1 (8.1) years	<p>index, ESR, pain VAS</p> <ul style="list-style-type: none"> • Radiographic severity: number of erosions, severity of wrist involvement 0-3, sacroiliitis • Skin severity: PASI, skin VAS 	<p>damage score, vertebral score</p> <ul style="list-style-type: none"> • Definition of correlations. Poor: <0.25. Moderate: 0.2 to <0.40. Strong: ≥ 0.40 source? 	<ul style="list-style-type: none"> • HAQ-DI had low correlation with skin severity (range of $r=0.095$ to 0.148) – not hypothesized 			<ul style="list-style-type: none"> • In general, the magnitude and directions of these correlations were aligned to what is expected from all other literature. Therefore, the working group think these data can be used, but recommend some caution in interpretation • The working group acknowledge the interpretation for correlations were lower than current standards. However, the results for measurements that should have high correlations with HAQ-DI falls in the acceptable ranges (ACR functional class: 0.75)
Leung 2008	<ul style="list-style-type: none"> • 108 patients with PsA (by CASPAR) in Hong Kong • 51.9% women • Mean (SD) age 49.3 (12.6) years • Mean (SD) duration of PsA 9 (6.8) years • Mean (SD) HAQ-DI 0.69 (0.67) • Mean (SD) SF-36 PF 63.3 (25.5) 	<p>HAQ-DI (Chinese) measures physical function</p> <p>Comparison:</p> <ul style="list-style-type: none"> • SF-36 PF • BASFI • Dougados Functional Index 	<ul style="list-style-type: none"> • HAQ-DI (and other instruments) evaluated with Rasch model analysis for: <ul style="list-style-type: none"> ○ Unidimensionality ○ Item fit ○ DIF ○ Item/person reliability and span 	<ul style="list-style-type: none"> • HAQ-DI and SF-36 PF fit the Rasch model, demonstrating unidimensionality, and measuring the same construct of physical disability • HAQ-DI, SF-36 PF and BASFI highly correlated with each other (range of $r=0.76$ to 0.80) [<i>not a priori</i>] • HAQ-DI correlated moderately with pain and patient global ($r=0.56$ and 0.53) [<i>not a priori</i>] • SF-36 PF had longer measurement span, 	<ul style="list-style-type: none"> • BASFI and Dougados functional index may behave differently in patients with/without sacroiliitis (DIF) • Nothing hypothesized for HAQ-DI 	<ul style="list-style-type: none"> • NA 	<p>(+)</p> <p>Remark: Solid support for structure validity of HAQ-DI and SF-36 PF</p> <p>Floor effect of HAQ-DI may affect interpretability of the score, but overall construct validity is good.</p>

				<p>item separation, and less floor effect.</p> <ul style="list-style-type: none"> Floor effect of HAQ-DI: 24.5% vs. SF-36 PF: 7.4% 			
<p>Brodzky 2010</p>	<ul style="list-style-type: none"> 183 patients with PsA (either Moll & Wright or CASPAR) in Hungary 57% women Mean (SD) age 50.1 (12.9) years Mean (SD) duration of PsA 9.2 (9.2) years Mean (SD) HAQ-DI 1.0 (0.7) 	<p>HAQ-DI (Hungarian)</p> <p>Comparison: PsAQoL EQ-5D</p> <p>Disease duration</p> <p>Disease activity (DAS28)</p> <p>BASDAI</p> <p>Patient global VAS</p> <p>Pain VAS</p> <p>Physician global VAS</p> <p>PASI</p>	<p>Higher PsAQoL/ HAQ/ lower EQ-5D scores were hypothesized for groups with severe PsA. (Implied expecting correlations not in low ranges).</p> <p>Correlations defined as strong: > 0.5, Moderate: 0.30–0.49, Weak: 0.1–0.29</p>	<p>Magnitude and direction of correlations generally aligned with findings from literature:</p> <ul style="list-style-type: none"> Strong correlations between HAQ-DI with DAS28, BASDAI, patient global, and pain VAS ($r=0.52$, 0.59, 0.50 and 0.54). HAQ-DI had high correlation with PsAQoL ($r=0.64$) and EQ-5D ($r=-0.71$) HAQ-DI had weak correlation with disease duration and PASI ($r=0.18$ and 0.21) and moderate correlation with physician global ($r=0.39$) 	<ul style="list-style-type: none"> Severe group classified by <ol style="list-style-type: none"> Hospital admission in past year Receiving work pension Use devices Use home care <p>Effect size (Cohen's d) was calculated, larger effect sizes represent better discriminative ability, defined as small (0.2–0.5), medium (0.5–0.8), or large (> 0.8).</p>	<ul style="list-style-type: none"> HAQ-DI were significantly different between severity groups. (all $p<0.05$) <ol style="list-style-type: none"> Hospital admission in past year (standardized mean difference, SMD 0.41) Receiving work pension (SMD 0.67) Use devices (SMD 0.46) Use home care (SMD 1.54) 	<p>(+/-)</p> <p>Working group remarks:</p> <p>HAQ-DI differentiate groups with severe disease as hypothesized.</p> <p>HAQ-DI correlated strongly with PsAQoL and EQ-5D with the hypothesized direction, and strongly with the implied magnitude. Although the magnitude of correlations were not explicitly spelled out.</p> <p>The correlations in results were generally aligned with conceptual expectation in literature. In view of the magnitude of correlations not explicitly spelled out, working group recommended (+/-) for adequacy for performance.</p>
<p>Katchamart 2014</p>	<ul style="list-style-type: none"> 47 patients with PsA (by CASPAR) 55.3% women Mean (SD) age 49 (10.1) years 	<p>HAQ-DI (Thai)</p> <p>Comparison:</p> <ul style="list-style-type: none"> Disease activity: ASDAS, BASDAI, Disease severity: 	<p>Indirect evidence of a priori hypothesis available: Sample size estimation was based on a significant correlation</p>	<p>Magnitude and direction of correlations generally aligned with findings from literature:</p> <ul style="list-style-type: none"> HAQ highly correlated with BASDAI ($r=0.81$), ASDAS ($r=0.76$), physician global 	<ul style="list-style-type: none"> None hypothesized 	<ul style="list-style-type: none"> NA 	<p>(+)</p> <p>Working group remarks:</p> <p>The correlation between HAQ-DI and BASDAI was 0.81, higher than requirement for sample size estimation. Other correlations found</p>

	<ul style="list-style-type: none"> • Mean (SD) duration of illness 6.97 (6.17) years • Mean (SD) HAQ 0.47 (0.47) • 38% spondylitis, 34% oligoarthritis, 17% polyarthritis, and 4% DIP joint arthritis 	<p>damaged peripheral joint count, spinal mobility</p> <ul style="list-style-type: none"> • Functional status: Grip strength <p>Others: Patient/physician global VAS, pain VAS, ESR, morning stiffness, joint counts (TJC, SJC)</p>	<p>between HAQ and BASDAI with a correlation coefficient of 0.59, implied significant and high correlation hypothesized.</p> <p>No definition of interpretation of correlation given.</p>	<p>($r=0.78$), pain ($r=0.71$), and patient global ($r=0.65$). All $p=0.01$</p> <ul style="list-style-type: none"> • HAQ-DI correlated moderately with grip strength ($r=-0.39$, $p=0.01$), morning stiffness ($r=0.44$, $p<0.05$), and ESR (0.52, $p<0.05$) • HAQ-DI weakly correlated with TJC ($r=0.32$, $p<0.05$) <p>HAQ-DI had no association with spinal mobility measures, damaged joints, grip strength, or SJC</p>			generally aligned with what is expected from literature.
Leung 2020	<ul style="list-style-type: none"> • 414 PsA consecutive patients with at least 2 years duration of PsA • 48% women • From 14 countries, 2 follow up • Mean (SD) age 52.4 (12.5) years • Mean (SD): duration of PsA 10.9 (8.1) years • Mean (SD) HAQ-DI 0.64 (0.68) 	<p>HAQ-DI</p> <p>Comparison:</p> <ul style="list-style-type: none"> • SF-12 PCS • PsAID-12 functional capacity (FC) • Patient global arthritis NRS and skin NRS • 68/66 TJC/SJC • Pain NRS <p>Disease activity (DAPSA, MDA)</p>	<p>HAQ-DI correlates highly with SF-12 PCS, PsAID FC</p> <p>HAQ-DI correlates weakly to moderately with patient global, disease activity, and pain</p> <p>HAQ-DI correlates very weakly with skin</p> <p>HAQ-DI differentiated between disease severity states (remission, LDA vs. higher disease activity)</p>	<p>HAQ-DI correlated strongly with SF-12 PCS, PsAID FC: $r=0.71$ to 0.79</p> <p>HAQ-DI correlated moderately with patient global arthritis ($r=0.61$ to 0.78), TJC ($r=0.45$ to 0.48), DAPSA ($r=0.59$ to 0.60)</p> <p>HAQ-DI correlated weakly with TJC/SJC ($r=0.26-0.32$),</p> <p>HAQ-DI correlated very weakly with patient global skin (0.23 to 0.29)</p>	<ul style="list-style-type: none"> • HAQ-DI distinguished groups of patients with defined remission (REM)/low disease activity (LDA) 	<ul style="list-style-type: none"> • Patients in REM had lower HAQ-DI than non-REM (0.27 vs. 0.73, $p<0.001$) • Patients in LDA had lower HAQ-DI than non-LDA (0.44 vs. 0.93, $p<0.001$) 	<p>(+)</p> <p>Overall, 78% of hypothesized correlations achieved.</p> <p>100% of hypothesized differences between known groups achieved</p>

			Definition of correlations: Very weak <0.3, Weak 0.3-0.5, Moderate 0.5-0.7, Strong >0.7				
Wan 2020	<ul style="list-style-type: none"> • 274 PsA patients from 4 USA centers • Mean (SD) age 49.3 (14.2) years • 49% women • Mean (SD) HAQ-DI 0.6 (0.6) 	LOS, F/U between week 12 to 52 HAQ-DI Comparators: <ul style="list-style-type: none"> • MDHAQ • PROMIS GH 10a, global physical health subscore (GPH) 	Strong correlation between HAQ-DI and MDHAQ ($r > 0.8$) Correlation HAQ-DI and GPH ($r = 0.6-1.0$) Moderate correlation with PsAID functional item Weak to moderate ($r < 0.6$) with pain, patient global, PsAID skin. 13 hypotheses given in table.	High correlation with MDHAQ ($r = 0.85$) Correlation with GPH $r = -0.75$ PsAID functional item ($r = 0.64$) Pain ($r = 0.61$) Patient global ($r = 0.52$) PsAID skin ($r = 0.30$) Overall, 80% hypotheses achieved	NA	NA	(+) Working group remark: Good quality paper 80% hypothesis achieved.

(+) indicates findings of the study had adequate performance of the instrument; (+/-) indicates equivocal performance; (-) indicates poor performance (less than adequate).
Abbreviations.

ACR: American College of Rheumatology; ASDAS: Ankylosing Spondylitis Disease Activity Score ;BASDAI: Bath Ankylosing Spondylitis Disease activity index; BASFI: Bath Ankylosing Spondylitis Functional Index; ESR: erythrocytes sedimentation rate; EQ-5D: EuroQol-5 items; HAQ: Health assessment Questionnaire; -DI: disability index; DAPSA: Disease Activity index for Psoriatic Arthritis; fibrositic TP: fibrositic tender point based on ACR criteria for fibromyalgia (Wolfe et al. Arthritis Rheum 1990;33:160-72); F/U: follow up; MDHAQ: multidimensional HAQ; IQR: interquartile range; MDA: minimal disease activity; NA: not applicable; NRS: numeric rating scale; PsA: psoriatic arthritis; PsAID: Psoriatic Arthritis Impact of Disease Questionnaire; -FC: functional capacity; PsAQoL: PsA quality of life index; LOS: longitudinal observation study; LDA: low disease activity; LOS: longitudinal observational study; MDA: minimal disease activity; PROMIS: Patient Reported outcome Measure Information System; -GPH: global physical health; -GH: global health; VAS: visual analogue scale; SF-36 PF: Medical Outcome Survey Short form -36 Physical Functioning domain; PCS: physical function summary score; SD: standard deviation; SMD: standardized mean difference; SJC: swollen joint count; TJC: tender joint count.

Table A.2. Report of Studies of Test-Retest Reliability for HAQ-DI in PsA with OMERACT Filter 2.1

2 data sets: Leung 2016 (unpublished data), Tillett 2019 (unpublished data). Final detail of both studies published in Leung et al., *The Journal of rheumatology*: jrheum-210175.

Author, year	Study description Characteristics of sample	Characteristics of testing situation	Results Sample recruited and sample considered stable for analysis	Scores at baseline and retest	Statistic used	Results	Minimal detectable change (95%CI) $SEM=SD_{baseline} \times \sqrt{(1-ICC)}$ $MDC=1.96 \times SEM \times \sqrt{2}$ $SEM=0.55 \times \sqrt{(1-0.94)}$ =0.13 $MDC=1.96 \times 0.13 \times \sqrt{2}$ = 0.36	Judgement Interpretation of authors of adequacy (+, +/-, -) (+)
Leung 2016	<ul style="list-style-type: none">•Consecutive patients with PsA fulfilled CASPAR recruited for validation of PsAQoL study•48% women•Mean (SD) age 51.5 (13.8) years•Mean (SD) duration of PsA 5.5 (8.4) years	<ul style="list-style-type: none">•2 weeks apart•Stability between timepoints assumed given PsA is a chronic illness and patients not required medication change	<ul style="list-style-type: none">•98 patients recruited to the main study, planned to recruit 40 patients who required no medication change for test-retest reliability•38 patients provided complete dataset•21 (55.3%) men•Mean (SD) age 53.9 (11.5) years	Mean (SD) T1: 0.38 (0.55) T2: 0.35 (0.56) Mean difference=0 (SD=0.19), p=1.00 95% CI: -0.373 to 0.373	ICC Spearman's rho (r)	ICC=0.94 (95% CI: 0.89-0.97) r=0.83 (p<0.001)		Good ICC and correlation between scores that changes were not expected. Bland/ Altman plot provided supportive evidence.
Tillett 2019	<ul style="list-style-type: none">•Consecutive patients with PsA fulfilled CASPAR, recruited for validation of composite measures	<ul style="list-style-type: none">•1 week apart•Assumed no change in condition in stable patients without medication change	<ul style="list-style-type: none">•140 patients recruited to the main study, planned to recruit 30 patients who were stable and did not require medication change for test-retest reliability•31 patients provided complete dataset•Mean (SD) age 54 (11) years•Mean (SD) duration of PsA 5.7 (4.7) years	Mean (SD) T1: 0.49 (0.59) T2: 0.50 (0.65) Mean difference=-0.004 (SD=0.28), p=0.96 95% CI: -0.114 to 0.105	ICC Spearman's rho (r)	ICC=0.94 (95% CI: 0.88-0.97) r=0.93 (p<0.01)	SEM=0.59 x $\sqrt{(1-0.94)}$ =0.14 $MDC=1.96 \times 0.14 \times \sqrt{2}$ = 0.39	(+) Good ICC and correlation between scores that changes were not expected. Bland/ Altman plot provided supportive evidence.

(+) indicates findings of the study had adequate performance of the instrument; (+/-) indicates equivocal performance; (-) indicates poor performance (less than adequate).

Abbreviations.

CASPAR: Classification of Psoriatic Arthritis Study criteria; HAQ: Health assessment Questionnaire; -DI: disability index; DAPSA: Disease Activity index for PSoriatic Arthritis; F/U: follow up; ICC: intraclass correlation coefficients; MDC: Minimal detectable change; MDA: minimal disease activity; NA: not applicable; NRS: numeric rating scale; SEM: standard error of measurement; SF-36 PF: Medical Outcome Survey Short form -36 Physical Functioning domain; PCS: physical function summary score; SD: standard deviation; SMD: standardized mean difference.

Table A.3. Reporting of Evidence of Responsiveness (Longitudinal Construct validity) for HAQ-DI in PsA with OMERACT Filter 2.1

3 articles: 1. Husted 1998; 2. Leung 2011; 3. Leung 2020.

Author Year	Study characteristics		Methods and results					
	Sample description (mean age, % gender, disease type/severity/dura- tion)	Study structure	Groups being contrasted**	Hypothesis of change (described expectation or anchor)	Anticipated results (NR=not reported in article)	Statistic used	Results observed	Adequacy (+, +/-, -)
Husted 1998	<ul style="list-style-type: none"> • 80 patients with PsA attending Toronto center • 70 patients had completed data at 2 occasions • 38.6% women • Mean (SD) age: 46.3 (11.4) years • Mean (SD) duration of PsA: 12.98 (7.68) years • NSAIDs only 27.1%, DMARDs 24.3%, MTX 31.4%, oral steroid 2.9% 	Two visits, F/U=6 months	Within person change baseline and F/U	<p>Three anchors used as following</p> <p>For the change in HAQ-DI, significant association with:</p> <ol style="list-style-type: none"> 1. Change in active joint number 2. Change in damaged joint number 3. Change in global health <p>HAQ-DI can distinguish patients 'and clinicians' impression of a clinically meaningful change:</p> <ol style="list-style-type: none"> 1. Improvement in global health 2. 30% reduction in active joints 3. Deterioration in global health 4. 30% increase in active joints 5. (>=1) increase in damage joint 	<ul style="list-style-type: none"> • HAQ-DI should have significant associations with the 3 anchors • SRM magnitude not hypothesized; but writing imply expecting SRM not low with the clinically meaningful change chosen. • SRM responsiveness defined as: 0.2 low; 0.5 moderate; 0.8 high 	SRM	<p>Change in HAQ-DI has significant association with 3 anchors:</p> <ol style="list-style-type: none"> 1. Change in # active joints (B=0.35, p=0.028, R²=0.12), 2. Change in # damaged joints (B=0.02, p=0.895, R²=0.00) -- NS 3. Change in perceived general health (B=0.37, p=0.002, R²=0.14) <p>SRM for HAQ-DI</p> <ol style="list-style-type: none"> 1. Improvement in global health (-0.482) 2. 30% reduction in active joints (-0.458) 3. Deterioration in global health (0.337) 4. 30% increase in active joints (0.043) 5. >=1 in damage joint (-0.070) 	<p>(+)</p> <p>Working group remarks:</p> <p>Correlation with 2 out of 3 (67%) anchors achieved statistical significance.</p> <p>Although SRM magnitude not explicitly stated, authors did imply anticipation of a result not in the low range. The results aligned in 3 out of 5 (60%) clinically meaningful changes proposed showing a reasonable SRM (improvement in global health, 30% reduction in active joints, deterioration in global health)</p>
Leung 2011	<ul style="list-style-type: none"> • 20 PsA patients with active 	<ul style="list-style-type: none"> • After 12 weeks, 11 	Within person change baseline and	Changes in HAQ-DI associated with change in anchor	Correlation between change in HAQ-DI and change	• r for change score	Change in HAQ-DI and change in anchor	<p>(+/-)</p> <p>Working group remark:</p>

	<p>disease who started TNFi</p> <ul style="list-style-type: none"> • Mean (SD) age 49.8 years • Mean (SD) duration of PsA: 7.8 years • 40% women • 68/66 TJC/SJC: 10.4/3.7 	<p>patients discontinued TNFi due to cost</p> <ul style="list-style-type: none"> • Patients seen at baseline and weeks 12, 24, 36, and 52 	<p>F/U timepoints</p> <p>With 4 F/U timepoints, 78 sets of PRO data (from 20 patients) were analyzed</p>	<p>(current health status compared to last visit: much better, slightly better, similar, slightly worse, much worse).</p>	<p>in anchor would be significant, with correlation (r) at least >0.3</p> <p>No hypothesis for magnitude of effect size</p> <p>Effect defined as small if < 0.2, medium if 0.3–0.5, and large if > 0.5.</p>	<ul style="list-style-type: none"> • Cohen's d effect size (ES) • SRM 	<p>correlation, $r=0.30$, $p<0.01$</p> <p>ES=-0.22</p> <p>SRM=-0.22</p>	<p>Correlation with anchor achieved statistical significance</p> <p>The ES/ SRM were lower than previous study, because data were derived from patient group reporting only "slight" change on F/U; therefore, this result is expected</p> <p>Overall, the working group thinks the paper can be considered. Although no formal ES magnitude was explicitly stated, an ES not in the low range was implied. However, as there was no formal hypothesis for ES magnitude, working group recommends caution in interpretation</p>
Leung 2020	<ul style="list-style-type: none"> • 414 PsA consecutive patients with at least 2-year duration of PsA • From 14 countries, 2 F/U • 48% women • Mean (SD) age: 52.4 (12.5) years • Mean (SD) duration of PsA 10.9 (8.1) years • Mean (SD) HAQ-DI 0.64 (0.68) 	<ul style="list-style-type: none"> • Patients seen at baseline, then F/U at 1-6 months • F/U data available for 350 patients 	<p>Patients endorsing improved/ same/ worsened compared to last visit</p>	<ul style="list-style-type: none"> • Change scores of HAQ-DI strongly correlated with change scores of other function scores (SF-12 PCS, PsAID FC) • Change scores of HAQ-DI with disease activity (DAPSA) moderately correlated 	<p>Correlations between change scores, direction and magnitudes given in table</p>	<ul style="list-style-type: none"> • Correlation (r) for change scores 	<ul style="list-style-type: none"> • High correlation with change scores of function: SF-12 PCS $r=-0.71$ -PsAID FC $r=0.68$ • Moderate correlation with change in PGA, pain, DAPSA ($r=0.54-0.57$) • weak correlation with change in tender/ swollen joint ($r=0.34 - 0.37$) • very weak correlation with 	<p>(+)</p> <p>Working group remarks:</p> <ul style="list-style-type: none"> • 78% <i>a priori</i> hypothesis on correlation of change scores achieved. • Higher effect sizes for change scores in change groups compared to no change group • Statistical significantly difference in change scores only seen in

				<ul style="list-style-type: none"> • Change scores statistically different from 3 groups (improved/same/worsened) • Higher effect sizes (Cohen's d/ SRM) for groups who were improved/worsen as compared to group with no change 	<ul style="list-style-type: none"> • Change scores between patients with change vs. no change statistically significant, • Higher effect size in groups with change than in group with no change • Magnitude of ES not explicitly mentioned. ES interpreted as 'trivial' (<0.20), 'small' (≥0.20 to <0.50), 'moderate' (ES ≥0.50 to <0.80), or 'large' (≥0.80) 	<ul style="list-style-type: none"> • Cohen's d • SRM 	<p>change in skin global r=0.29</p> <p>Significant difference in change scores in worsened group compared to no change group (p<0.05) and improved group (NS)</p> <p>SRM of HAQ-DI change: -improved: -0.19 -no change: -0.07 -worsen: 0.37</p> <p>Similar results for Cohen's d: -improved: -0.24 -no change: -0.08 -worsen: 0.46</p>	<p>worsen group compared to no change group (this is expected)</p> <ul style="list-style-type: none"> • Data derive from stable patients in outpatient clinics, further improvement is less likely. • Although no formal hypothesis for magnitude of effect size, results aligned general believe.
Wan 2020	<ul style="list-style-type: none"> • 274 PsA patients from 4 USA centers • 95 patients gave F/U data • Mean (SD) age 49.3 (14.2) years • 49% women • Mean (SD) HAQ-DI 0.6 (0.6) 	LOS, F/U between week 12 to 52	<ol style="list-style-type: none"> 1. All patients 2. Therapy initiator 3. Patient reported improvement 4. 20% improvement in SJC/TJC 	<p>Change of HAQ-DI has high correlation with change of MDHAQ, slightly lower for that with GPH.</p> <p>Higher SRM in patients expected to change (group 2,3,4) compared to all patients (group 1).</p> <p>SRM similar to SRM of MDHAQ and GPH</p> <p>SRM similar to SRM of pain and patient global</p>	SRM 0.3-0.5	SRM	<p>Change scores of HAQ-DI and MDHAQ (r=0.71-0.84)</p> <p>Change scores of HAQ-DI and GPH (r=-0.42-0.62)</p> <p>SRM for HAQ-DI</p> <ol style="list-style-type: none"> 1. All patients (n=95): 0.13 2. Rx initiator (n=65): 0.18 3. Patient improved (n=19): 1.06 4. >20% TJC/SJC (n=21): 0.39 <p>Similar SRM ranges for MDHAQ, slightly higher SRM for GPH.</p> <p>SRM pain: 0.38</p>	<p>(+)</p> <p>Working group remark:</p> <ul style="list-style-type: none"> • Good quality paper • Multiple clinically meaningful anchors used, patient reported improvement and 20% improvement in TJC/SJC likely constitute true changes • SRM with magnitude hypothesized <i>a priori</i> and achieved in 60% of cases. • Overall >75% of all hypothesis achieved

							SRM patient global: 0.16	
--	--	--	--	--	--	--	-----------------------------	--

***Groups being contrasted can include: Within person change, between person differences (if used for the study of responsiveness) or Between group differences of within person change (contrasting relative change between two groups).*

(+) indicates findings of the study had adequate performance of the instrument; (+/-) indicates equivocal performance; (-) indicates poor performance (less than adequate).

Abbreviations.

CASPAR: Classification of Psoriatic Arthritis Study criteria; DAPSA: Disease Activity index for Psoriatic Arthritis; DMARDs: Disease modifying anti-rheumatic drugs; ES: effect size; ESR: erythrocytes sedimentation rate; EQ-5D: EuroQol-5 items; HAQ: Health assessment Questionnaire; -DI: disability index; F/U: follow-up; NA: not applicable; NRS: numeric rating scale; NS: not statistically significant; NSAID: non-steroidal anti-inflammatory drug; PsA: psoriatic arthritis; PsAID: Psoriatic Arthritis Impact of Disease Questionnaire; -FC: functional capacity; PsAQoL: PsA quality of life index; LOS: longitudinal observation study; LDA: low disease activity; MDA: minimal disease activity; MTX: methotrexate; PROMIS: Patient Reported outcome Measure Information System; -GPH: global physical health; -GH: global health; SEM: standard error of measurement; SRM: standardized response mean; SF-36 PF: Medical Outcome Survey Short form -36 Physical Functioning domain; PCS: physical function summary score; Rx: treatment; SD: standard deviation; SMD: standardized mean difference; SJC: swollen joint count; TJC: tender joint count; TNFi: tumor necrosis factor inhibitors; VAS: visual analogue scale.

Table A.4. Reporting of Evidence of Clinical Trial Responsiveness for HAQ-DI in PsA with OMERACT Filter 2.1

29 articles in:

Leung et al. Clinical trial discrimination of physical function instruments for psoriatic arthritis: a systematic review. Semin Arthritis Rheum 2020;50(5):1158-

81. <https://doi.org/10.1016/j.semarthrit.2020.05.022>

Author/ year/ (study acronyms)	Intervention/ comparator (sample size, N)	Primary outcome/ time point	P value of change scores of treatment arm compared to placebo arm	Effect sizes at primary endpoint (unless specified)	Fulfills <i>a priori</i> hypothesis [‡]	Adequacy of instrument performance
Antoni, et al. 2005 (IMPACT)	IFX 5mg/kg vs. PCB (N=104)	ACR20/ Week 16	<0.001	SRM [¶] (for improvement) [§] at Week 16: IFX = 6.07 PCB = -0.19	1, 2, 3	(+)
Antoni, et al. 2005 (IMPACT2)	IFX 5mg/kg vs. PCB (N=200)	ACR20/ Week 14	<0.001	SRM [¶] (for improvement) [§] at Week 14: IFX = 1.08 PCB = -0.19	1, 2, 3	(+)
Mease, et al. 2005 (ADEPT)	ADA 40mg Q2W vs. PCB (N=313)	ACR20/ Week 12; Δ in modified Total Sharp Score/ Week 24	<0.001	SRM at Week 12: ADA = -0.8 PCB = 0.2	1, 2, 3	(+)
Genovese, et al. 2007	ADA 40mg Q2W vs. PCB (N=100)	ACR20/ Week 12	0.010	SRM at Week 12: ADA = -0.6 PCB = -0.33	1, 2, 3	(+)
Mease, et al. 2000	ETN 25mg BIW vs. PCB (N=60)	PsARC/ Week 12	<0.0001	ES ₂ [*] at Week 12: ETN = -0.547 PCB = -0.237	1, 2, 3	(+)
Mease, et al. 2010	ETN 25mg BIW vs. PCB (N=205)	ACR20/ Week 12	<0.001	Effect size for Week 12: NA ES ₂ at Week 24 (end of double-blind phase): ETN = -0.597 PCB = -0.098	1, 2, 3	(+)
Gniadecki, et al. 2012 (PRESTA)	ETN 50mg BIW/QW vs. ETN 50mg QW/QW (N=752)	Psoriasis clear or almost clear/ Week 12	0.792	ES ₂ at Week 12: ETN 50mg BIW/QW: -0.74 ETN 50mg QW/QW: -0.69	1, 3	(+)
Mease, et al. 2019 (SEAM-PsA)	ETN 50mg QW vs. ETN 50mg QW plus MTX vs. MTX alone (N=851)	ACR20/ Week 24	Etanercept = 0.67 Combination =0.34	SRM at Week 24: ETN = -0.733 ETN plus MTX = -0.685 MTX alone = -0.646	1, 3	(+/-)

Kavanaugh, et al. 2009 (GO-REVEAL)	GOL (2 doses) vs. PCB (N=405)	ACR20/ Week 14	<0.001	SRM at Week 14: GOL 100mg = -0.75 GOL 50mg = -0.62 PCB = -0.09	1, 2, 3	(+)
Kavanaugh, et al. 2017 (GO-VIBRANT)	GOL IV 2mg/kg vs. PCB (N=480)	ACR20/ Week 14	<0.001	SRM at Week 14: GOL IV = -1.13 PCB = -0.26	1, 2, 3	(+)
Gladman, et al. 2014 (RAPID-PsA)	CZP (2 doses) vs. PCB (N=409)	ACR20, EULAR response/ Week 12	<0.001	SRM at Week 12: CZP 400mg Q4W = -0.83 CZP 200mg Q2W = -0.80 PCB = -0.44	1, 2, 3	(+)
McInnes, 2014 (Phase II)	SEC (10mg/kg) vs. PCB (N=42)	ACR20/ Week 6	0.002	SRM at Week 6: SEC = -0.680 PCB = 0.018	1, 2, 3	(+)
Mease, et al. 2015 (FUTURE I)	SEC (2 doses with loading) vs. PCB (N=606)	ACR20/ Week 24*	0.0001	SRM at Week 24: SEC 150mg = -0.703 SEC 75mg = -0.721 PCB = 0.239	1, 2, 3	(+)
McInnes, et al. 2015 FUTURE II)	SEC (3 doses) vs. PCB (N=397)	ACR20/ Week 24*	(300mg): 0.004 (150mg): 0.056 (75mg): 0.92	SRM at Week 24: SEC 300mg = -1.12 SEC 150mg = -0.96 SEC 75mg = -0.644 PCB = -0.522	1, 2, 3	(+)
Kavanaugh, et al. 2016 (FUTURE II) -subgroup analysis	TNFi-naïve vs. TNFi-exposed	ACR20/ Week 24*	TNFi naïve: p<0.05 for SEC 300mg and 150mg TNFi exposed: P<0.05 for SEC 300mg	SRM at Week 24 (TNFi-naïve vs. -exposed): SEC 300mg: -1.20 vs. -1.02 SEC 150mg: -1.15 vs. -0.71 SEC 75mg: -0.77 vs. -0.44 PCB: -0.63 vs. -0.35	4	(+)
Nash, et al. 2018 (FUTURE III)	SEC (2 doses) vs. PCB (N=414)	ACR20/ Week 24*	<0.01	SRM at Week 24: SEC 300mg = -0.81 SEC 150mg = -0.57 PCB = -0.24	1, 2, 3	(+)
Mease, et al. 2017 (SPIRIT-P1)	IXE (2 doses) vs. PCB vs. ADA (N=417)	ACR20/ Week 24*	Ixe (Q4W/Q2W): <0.001 Ada: <0.01	SRM at Week 24: IXE Q2W = -0.98 IXE Q4W = -0.85 PCB = -0.35 ADA = -0.74	1, 2, 3	(+)
Nash, et al. 2017 (SPIRIT-P2)	IXE (2 doses) vs. PCB (N=363)	ACR20/ Week 24*	<0.0001	SRM at Week 24: IXE Q2W = -0.36 IXE Q4W = -0.55 PCB = -0.18	1, 2, 3	(+)

Mease, et al. 2014 Phase II	BRO (2 doses) vs. PCB (N=168)	ACR20/ Week 12	NS	SRM [‡] at Week 12: BRO 280mg = -0.60 BRO 140mg = -0.38 PCB = -0.21	1, 2, 3	(+)
Gottlieb, et al. 2009	UST (2 doses) vs. PCB (N=146)	ACR20/ Week 12	0.0005	SRM [‡] at Week 12: UST = -0.66 PCB = -0.14	1, 2, 3	(+)
McInnes, et al. 2013 (PSUMMIT I)	UST (2 doses) vs. PCB (N=615)	ACR20, EULAR response PASI75/ Week 24*	<0.0001	SRM [‡] at Week 24: UST 90mg = -0.59 UST 45mg = -0.62 PCB = -0.21	1, 2, 3	(+)
Ritchlin, et al. 2014 (PSUMMIT II)	UST (2 doses) vs. PCB (N=312)	ACR20, EULAR response PASI75/ Week 24*	<0.001	SRM [‡] at Week 24: UST 90mg = -0.66 UST 45mg = -0.59 PCB = 0.00 SRM [‡] at Week 24 Subgroups analysis (TNFi-naïve vs. -exposed): UST 90mg = -0.66 vs. -0.66 UST 45mg = -0.66 vs. -0.59 PCB = 0 vs. 0	1, 2, 3	(+)
Araugo, et al. 2019 (ECLIPSA)	UST (45mg or 100mg if body weight >100kg) vs. TNFi (N=47)	Enthesitis resolution (SPARCC = 0) at Week 24	<0.001 compared to baseline.	ES ₂ [‡] at Week 24: UST = -1.81 TNFi = -1.74	1, 2, 3	(+)
Deodhar, et al. 2018	GUS 100mg vs. PCB (N=149)	ACR20/ Week 24*	0.00025	SRM at Week 24: GUS = -0.82 PCB = -0.11	1, 2, 3	(+)
Mease, et al. 2017 (ASTRAEA)	ABT vs. PCB (N=424)	ACR20/ Week 24*	0.097	SRM [‡] at Week 24: ABT = -0.69 PCB = -0.40 SRM [‡] at Week 24: Subgroup analysis: (TNFi-naïve vs. -exposed) ABT = -0.62 vs. -0.64 PCB = -0.39 vs. -0.34	1, 3	(+)
Mease, et al. 2017 (OPAL Broaden)	TOF (2 doses) vs. PCB vs. ADA (N= 422)	ACR20/ 3 months	<0.05 compared to placebo; NS compared to Adalimumab	SRM at 3 months: TOF 10mg = -0.78 TOF 5mg = -0.69 PCB = -0.36 ADA = -0.76	1, 2, 3	(+)

Gladman, et al. 2017 (OPAL Beyond)	TOF (2 doses) vs. PCB (N=395)	ACR20, Δ in HAQ-DI/ 3 months	<0.05	SRM at 3 months: TOF 10mg = -0.64 TOF = -0.70 PCB = -0.26	1, 2, 3	(+)
Mease, et al. 2018 (EQUATOR) Phase II	FIL 200mg vs. PCB (N=131)	ACR20/ Week 16	0.0009	SRM at Week 16: FIL = -1.14 PCB = -0.56 ES ₂ at Week 16: FIL = -1.04 PCB = -0.47	1, 2, 3	(+)
Mease, et al. 2016 Phase II	CLZ (3 doses) vs. PCB (N=165)	ACR20/ Week 16	Not significant	SRM [‡] at Week 16: CLZ 200mg = -0.51 CLZ 100mg = -0.77 CLZ 50mg = -0.83 PCB = -0.52	1, 2, 3	(+)

[‡] SRM calculated using percentage change score and SD of percentage change; [§] SRM for improvement, a negative value indicate deterioration; [¥] Effect sizes estimated based on mean and SD of change scores calculated from median and IQR from original publication; * early escape for patients with inadequate response in the PCB group to active treatment group at Week 16; ** option to switch TNFi at Week 24; (+) indicates findings of the study had adequate performance of the instrument; (+/-) indicates equivocal performance; (-) indicates poor performance (less than adequate); NA: not applicable.

Abbreviations: Δ: change; ACR: American College of Rheumatology Response criteria; ABT: abatacept; ADA: adalimumab; ALC: alefacept; BIW: twice a week; BRO: brodalumab; CI: confidence interval; CLZ: clazakizumab; CZP: certolizumab pegol; ES₂: Effect size 2 (the mean difference divided by the pooled standard deviation, i.e Cohen's *d*); ETN: etanercept; EULAR: European League Against Rheumatism; FIL: filgotinib; GOL: golimumab; GUS: guselkumab; HAQ-DI: Health Assessment Questionnaire – Disability Index; IFX: infliximab; IXE: ixekizumab; IV: intravenous; IQR: interquartile range; MTX: methotrexate; NA: not available; NS: not significant; PASI: Psoriasis Area and Severity Index; PCB: placebo; PsARC: Psoriatic arthritis Response Criteria; QW: once a week; Q2W: once every 2-week; SD: standard deviation; SE: standard error; SEC: secukinumab; SPARCC: *Spondyloarthritis Research Consortium of Canada* enthesitis index; SRM: Standardized response mean (mean difference divided by the standard deviation of the differences between baseline and assessment end point); TOF: tofacitinib; TNFi: tumor necrosis factor inhibitor; UST: ustekinumab; vs.: versus.

[¥]A priori hypothesis:

1. At the primary endpoint/end of double blinded phase, patients given bDMARDs have significant change in HAQ-DI, whereas patients on placebo arm do not (except for alefacept and clazakizumab where no significant difference is expected)
2. The change scores of HAQ-DI among patients given bDMARDs are significantly higher than those of the placebo arm
3. Within individual trial, the effect sizes of change scores of HAQ-DI are higher in the bDMARD arms compared to the MTX or csDMARD arms, but do not differ significantly with different bDMARD doses (or with TNFi as comparison).
4. If data for subgroup analysis is available, the effect sizes of change scores of HAQ-DI are higher in TNF naïve versus TNF exposed subgroup

Table A.5. Reporting of Evidence of Threshold of meaning for HAQ-DI in PsA with OMERACT Filter 2.1.

4 articles: Kwok 2010, Mease 2011, Leung 2011, Leung 2020

Kwok 2010 was excluded due to quality

Author Year	Study characteristics		Methods and results						Interpretation of authors of adequacy (+, +/-, -)
	Sample description (mean age, % gender, disease type/severity/ duration)	Study structure (type of study, timing of F/U, interventions)	Threshold assessed (MID, MCID, PAS, LDA)	Method (anchor or distribu- tional)	Threshold method: anchor used and categories in that anchor	Definition of threshold of meaning using this approach	Threshold of meaning (specify value), AUC if available.	% of sample meeting/ exceeding this threshold	
Kwok 2010	<ul style="list-style-type: none"> • 200 patients with PsA (by Moll and Wright) in Ontario, Canada • 58.5% women • Mean (SD) age: 51.1 (14.1) years • Mean (SD) duration of PsA: 11.2 (7.9) years • DMARD use: 64% • Biologic use: 18.5% • Steroid use: 4% 	<ul style="list-style-type: none"> • Completed PROs in 2 consecutive visits • Mean time between F/U visits: 8.28 (5.94) months 	Minimally important difference (MID)	Anchor	Overall status compared to last visit: 1. much better 2. better 3. the same 4. worse 5. much worse	<p>Mean change of scores in those rated “better” or “worse” groups</p> <p>Correlation between change in HAQ-DI and anchor >0.37</p>	<p>MID for HAQ-DI: Mean (SD) [95% CI]</p> <ul style="list-style-type: none"> • “Better” (n=35): -0.131 (0.411) [95% CI: -0.272 to 0.011] • “Worse” (n=50) 0.131 (0.309) [95% CI: 0.044 to 0.219] • No overlap of CI • Change in HAQ-DI correspond to much better (n=22): -0.36 (0.432) [95% CI: -0.554 to -0.171] • Change in HAQ-DI correspond to much worse (n=4): 0.44 (0.315) [95% CI: -0.063 to 0.938] • Correlation of change in HAQ-DI & anchor =0.374, p<0.01 	<p>Better/much better: 28.5%</p> <p>Worse/much worse: 27%</p> <p>To establish the MID for improvement/ worsening for Patient-reported outcomes</p>	<p>(+/-)</p> <p>Working group remark:</p> <p>This paper only used one statistical method to estimate MID with no sensitivity/ specificity analysis. Thus, the quality of paper is borderline. However, working group consider this as the first paper attempting to establish MID for numerous MID, it still have reference value to include.</p> <p>The MID estimated was very small. This is expected as data were collected from known clinic population</p>

Mease 2011	<ul style="list-style-type: none"> • 205 PsA patients in an RCT receiving etanercept vs. PCB • 161 patients (92 etanercept, 69 PCB) who had HAQ-DI showing improvement • 48% men • Mean (SD) age 46.8 (11.1) years • Mean (SD) duration of PsA 9.1 (range 0-41.4) years • Mean (SD) HAQ-DI 1.16 (range: 0.13-2.88) 	Patients seen at baseline and weeks 4, 12, and 24	Minimally clinically important improvement (MCII)	<ul style="list-style-type: none"> • Anchor • Distributional 	<ul style="list-style-type: none"> • Anchor 1: Patients rating importance of improvement • Anchor 2: patient rating of satisfaction • Both anchors (1=not at all to 7=maximum) <ul style="list-style-type: none"> ○ ratings 2-3 deemed minimally important/satisfied ○ ratings 6-7 deemed very important improvement /satisfied <p>Distributional method: 1.96 x standard error of measurement</p>	Mean change of HAQ-DI corresponding to anchor groups	<p>Anchor 1:</p> <ul style="list-style-type: none"> ○ 11 (2.8%) responded to ratings 2-3, corresponding change in HAQ-DI 0.335 to 0.360. Rounding up to give MCII as 0.35 ○ 81 (75%) responded to ratings 6-7, corresponding HAQ-DI 0.435 to 0.460. Rounding up to very important improvement of HAQ-DI: 0.45 <p>Anchor 2:</p> <ul style="list-style-type: none"> ○ 83 (21.4%) responded to ratings 2-3, corresponding to change in HAQ-DI 0.293 to 0.360. ○ 157 (40.5%) responded to ratings 6-7, corresponding to change in HAQ-DI 0.559 to 0.625 <p>Distributional methods: Half SD=0.293 1.96 x SEM=0.266</p>	<p>From anchor 1</p> <ul style="list-style-type: none"> ○ 2.8% of response, MCII 0.35 ○ 75% of response, very important change 0.45 	<p>(+)</p> <p>Working group remarks:</p> <ul style="list-style-type: none"> • 2 different anchors used, giving sensible results for minimal change for improvement. Close ranges for ratings 2-3 for important/satisfaction, which is higher than the error bar by distributional methods. Thus, the statistical interpretation was reasonable • No data for MCII for deterioration, which is not the purpose of this paper • In patients from an RCT, improvement is expected. Only MCII was derived • This MCII is important and relevant for interpretation of results in RCTs
Leung 2011	• 20 PsA patients with active	• Patients seen at baseline	MID	Anchor	Current health status	Mean change of scores in those	Mean (SD)	Much better: 34.6%	(+/-)

Author Year	Study characteristics		Methods and results						Interpretation of authors of adequacy (+, +/-, -)
	Sample description (mean age, % gender, disease type/severity/ duration)	Study structure (type of study, timing of F/U, interventions)	Threshold assessed (MID, MCID, PAS, LDA)	Method (anchor or distribu- tional)	Threshold method: anchor used and categories in that anchor	Definition of threshold of meaning using this approach	Threshold of meaning (specify value), AUC if available.	% of sample meeting/ exceeding this threshold	
	<p>disease who started TNFi</p> <ul style="list-style-type: none"> • Mean (SD) age 49.8 years • Mean (SD) duration of PsA: 7.8 years • 60% men • 68/66 TJC/SJC: 10.4/3.7 	<p>and weeks 12, 24, 36, and 52</p> <ul style="list-style-type: none"> • With 4 F/U timepoints, 78 sets of PRO data (from 20 patients) were analysed 			<p>compared to last visit:</p> <ol style="list-style-type: none"> 1. much better, 2. slightly better, 3. similar, 4. slightly worse, 5. much worse. 	<p>rated “slightly better” or “slightly worse” groups</p> <p>Correlation between change in HAQ-DI and anchor</p>	<p>MID for improvement (n=17): -0.27 (0.42)</p> <p>Mean (SD) MID for deterioration (n=21): 0.095 (0.18)</p> <p>No significant differences in MID when stratified by sex or higher baseline HAQ>1.0.</p> <p>Correlation of change in HAQ-DI and Anchor=0.30, p<0.01</p>	<p>Slightly better: 21.3%</p> <p>Similar: 10.3%</p> <p>Slightly worse: 26.9%</p> <p>Much worse: 6.4%</p> <p>To establish the MCID for improvement/worsening for PROs</p>	<p>Working group remarks:</p> <p>This paper only used one statistical method to estimate MID with no sensitivity/specificity analysis. The sample size was very small. Thus, quality of paper is borderline</p> <p>However, the MID for improvement - 0.27 align with results found in another paper in TNFi treatment population, and think this paper also give support to the measurement property of HAQ-DI in this setting.</p> <p>Working group give (+/-) as performance adequacy, while recommends interpreting evidence with caution</p>

Leung 2020	<ul style="list-style-type: none"> • 414 PsA consecutive patients (52% men) with at least 2 year duration of PsA • From 14 countries, 2 follow up • Mean (SD) age: 52.4 (12.5) years • Mean (SD) duration of PsA 10.9 (8.1) years • Mean (SD) HAQ-DI 0.64 (0.68) 	<ul style="list-style-type: none"> • Patients seen at baseline, then F/U at 1-6 months • 350 patients gave F/U data 	<ul style="list-style-type: none"> • Minimally clinically important difference (MCID) • Patient defined remission (REM)/ low disease activity state (LDA). Wordings derived with patient input • Patient acceptable symptom state (PASS) 	Anchor	<p>MCID – compared to last visit – improved, same, or worse</p> <p>LDA: Patient defined LDA/MDA/ DAPSA LDA (Yes/No)</p> <p>REM: Patient defined REM/VLDA/ DAPSA REM (Yes/No)</p> <p>PASS (Yes/No)</p>	<p>MCID: mean change of HAQ-DI corresponds to patient endorsing improvement/ worsening</p> <p>REM/LDA/PASS via 75th percentile of scores, corresponding to Youden's Index</p> <p>Sensitivity analysis/ AUC</p>	<p>MCID improvement: -0.16 (0.87)</p> <p>MCID worsening: 0.30 (0.81)</p> <p>Patient defined LDA (n=245) 75th percentile: 0.75</p> <p>ROC cut-off: 0.75</p> <p>Sensitivity/ specificity/AUC (0.79/0.55/0.69)</p> <p>MDA (n=165) 75th percentile: 0.25</p> <p>ROC cut-off: 0.50</p> <p>Sensitivity/ specificity/AUC (0.94/0.72/0.88)</p> <p>DAPSA LDA (n=230) 75th percentile: 0.63</p> <p>ROC cut-off: 0.50</p> <p>Sensitivity/ specificity/AUC (0.74/0.71/0.78)</p> <p>Patient defined REM (n=86) 75th percentile: 0.50</p> <p>ROC cut-off: 0.63</p> <p>Sensitivity/ specificity/AUC (0.88/0.47/0.71)</p> <p>VLDL (n=54)</p>	<p>Improved: 27.3%</p> <p>Worsened: 18.4%</p> <p>Patient defined REM: 24.6%</p> <p>Patient defined LDA: 70%</p> <p>PASS: 80%</p>	<p>(+)</p> <p>Working group remarks:</p> <p>Several anchors used, showing sensible results. Used multiple statistical methods, and sensitivity analysis done.</p> <p>Data come from cohort of clinic patients not expecting improvement. MCID worsening derived from this study would be more relevant</p>
------------	---	---	---	--------	---	---	--	--	--

Author Year	Study characteristics		Methods and results						Interpretation of authors of adequacy (+, +/-, -)
	Sample description (mean age, % gender, disease type/severity/ duration)	Study structure (type of study, timing of F/U, interventions)	Threshold assessed (MID, MCID, PAS, LDA)	Method (anchor or distribu- tional)	Threshold method: anchor used and categories in that anchor	Definition of threshold of meaning using this approach	Threshold of meaning (specify value), AUC if available.	% of sample meeting/ exceeding this threshold	
							75 th percentile: 0.0 ROC cut-off: 0.13 Sensitivity/ specificity/AUC (0.91/0.69/0.85) DAPSA REM (n=83) 75 th percentile: 0.13 ROC cut-off: 0.13 Sensitivity/ specificity/AUC (0.77/0.72/0.78) PASS (n=280) 75 th percentile: 0.63 ROC cut-off: 0.63 Sensitivity/ specificity/AUC (0.76/0.72/0.81)		

(+) indicates findings of the study had adequate performance of the instrument; (+/-) indicates equivocal performance; (-) indicates poor performance (less than adequate).
Abbreviations.

ACR: American College of Rheumatology; AUC: area under curve; CASPAR: Classification of Psoriatic Arthritis Study criteria; CI: confidence interval; ESR: erythrocytes sedimentation rate; EQ-5D: EuroQoL-5 items; HAQ: Health assessment Questionnaire; -DI: disability index; DAPSA: Disease Activity index for Psoriatic Arthritis; F/U: follow up; ICC: intraclass correlation coefficients; MDHAQ: multidimensional HAQ; IQR: interquartile range; MDC: Minimal detectable change; MDA: minimal disease activity; MID: minimally important difference; NA: not applicable; NRS: numeric rating scale; PsA: psoriatic arthritis; PsAID: Psoriatic Arthritis Impact of Disease Questionnaire; -FC: functional capacity; PsAQoL: PsA quality of life index; LOS: longitudinal observation study; LDA: low disease activity; MDA: minimal disease activity; PROMIS: Patient Reported outcome Measure Information System; -GPH: global physical health; -GH: global health; VAS: visual analogue scale; PASS: patient acceptable symptom status; REM: remission; ROC: receiver operating characteristic; SEM: standard error of measurement; SF-36 PF: Medical Outcome Survey Short form -36 Physical Functioning domain; PCS: physical function summary score; SD: standard deviation; SMD: standardized mean difference; SJC: swollen joint count; TJC: tender joint count; TNFi: tumor necrosis factor inhibitors.

Evidence of measurement properties for SF-36 Physical functioning domain

Table B.1. Reporting of Evidence of Construct Validity for SF-36 PF in PsA with OMERACT Filter 2.1.

Truth: 4 articles (Husted 1997, Taylor 2007, Leung 2008, Leung 2010)

Author Year	Study characteristics		Methods and results used to assess construct validity. Two categories provided, correlational testing and known group comparisons. Both need brief description of a priori hypotheses (including expected results and how they were defined) and the results found and what the authors synthesis (was that good?) Look for strong associations/differences and the absence of the same (no correlation, no differences expected and none seen)				
	Sample description (mean age, % gender, disease type/severity /duration)	Study design/ methods	Correlation with other measures		Differences in scores across groups known to be different		Judgement adequacy (+, +/-, -)
			A priori hypothesis (described expected results, and comparison instrument)	Results	A priori hypotheses of different between the groups	Results	
Husted 1997	<ul style="list-style-type: none"> • 133 patients with PsA, F/U in Toronto center • Women/men: 43/70 • Mean (SD) age 50.5 (12.6) years • Mean (SD) duration of PsA 21.1 (8.04) years • DMARDs 21.2%, MTX 25.7%, oral steroid 4.4% 	<ul style="list-style-type: none"> • Cross sectional study • Completed SF-36 <p>Comparison: SF-36 PF with total score and other domains</p> <ul style="list-style-type: none"> • Physical function: Grip strength, ACR functional class • Pain: # fibromyalgia points • Disease activity: # active joints, morning stiffness, PASI • Disease severity: # damage joints 	<ul style="list-style-type: none"> • Item-total correlation >0.4 • Items closely related to own scales than other scales • SF-36 PF should correlate highly with measures of function • SF-36 PF should correlate moderately to highly with disease activity and severity • Study authors defined correlations a priori as: Correlations -poor $r < 0.25$ -moderate ≤ 0.25- 0.40 -high $r > 0.40$ 	<ul style="list-style-type: none"> • Item-total correlations range 0.52 to 0.85 • Correlations between items in own scale higher than other scales • Correlation between SF-36 PF and -Grip strength: 0.46 -ACR fx class: -0.62 Pain: -# fibromyalgia pt: -0.51 Disease activity: -# active joints: -0.46 -morning stiffness: -0.44 -PASI: -0.05 (not hypothesized) Disease severity: -# damage joint: -0.33 	All SF-36 domains of PsA patients should be significantly lower than the general population (both US and UK population data used)	<ul style="list-style-type: none"> • Significant lower SF-36 PF scores for PsA patients (68.8 ± 2.65) compared to both US (85.2 ± 0.68) and UK (86.2 ± 0.26) populations ($p < 0.0001$) 	<p>(+)</p> <p>Working group remarks:</p> <p>All hypothesis satisfied</p>

Taylor 2007	<ul style="list-style-type: none"> • 134 patients with clinic diagnosed PsA in New Zealand • 43% women • Mean age 52.3 (14.1) years • Median disease duration 12.3 (IQR 6.3-21.1) years • Mean HAQ-DI 0.50 (0.59) • Mean SF-36 PF 60.4 (27.1) 	<p>SF-36 PF measures functional ability</p> <p>Comparison: HAQ-DI measures physical function</p>	<p>Fit of HAQ-DI and SF-36 PF to the Rasch model, including unidimensionality, linear model, free of mis-fit items, no differential item functioning (DIF)</p>	<ul style="list-style-type: none"> • Both HAQ-DI and SF-36 PF fit into Rasch model. • Showing unidimension, measure same construct of physical disability • SF-36 PF has fewer misfit items, better distribution, scale length, less DIF. • Floor effect of SF-36 PF: 3.1% vs. HAQ-DI: 30.4% 	<ul style="list-style-type: none"> • Both SF-36 PF and HAQ-DI fit the Rasch model analysis, indicative of measuring similar construct. • SF-36 PF have slightly better measurement span, separation between items, and less floor effects. 	<ul style="list-style-type: none"> • Both met criteria of fit to Rasch model analysis • Bland and Altman plot indicated significant relationship between the mean of both scores. 	<p>(+)</p> <p>Remark: Overall solid support for structure validity of HAQ-DI and SF36 PF</p>
Leung 2008	<ul style="list-style-type: none"> • 108 patients with PsA (by CASPAR) in Hong Kong • Women/men: 56/52 • Mean (SD) age 49.3 (12.6) years • Mean (SD) duration of PsA 9 (6.8) years • Mean (SD) HAQ-DI 0.69 (0.67) • Mean (SD) SF-36 PF 63.3 (25.5) 	<p>SF-36 PF (Chinese)</p> <p>Comparison – other functional scales:</p> <ul style="list-style-type: none"> • SF-36 PF • BASFI • Dougados Functional Index 	<ul style="list-style-type: none"> • SF-36 PF (and other instruments) evaluated with Rasch model analysis for: <ul style="list-style-type: none"> ○ Unidimensionality ○ Item fit ○ DIF ○ Item/person reliability and span 	<ul style="list-style-type: none"> • SF-36 PF and HAQ-DI fit the Rasch model, unidimensionality, measure same construct of physical disability • SF-36 PF, HAQ-DI and BASFI highly correlated with each other (r 0.76 to 0.80) [<i>not a priori</i>] • SF-36 PF correlated moderately with pain and patient global (r 0.44 and 0.449), [<i>not a priori</i>] • Floor effect of SF-36 PF: 7.4% vs. HAQ-DI: 24.5% 	<ul style="list-style-type: none"> • BASFI and Dougados functional index may behave differently in patients with/without sacroiliitis (DIF) • Nothing hypothesized for HAQ-DI 	<ul style="list-style-type: none"> • NA 	<p>(+)</p> <p>Remark: Solid support for structure validity of HAQ-DI and SF-36 PF</p>
Leung 2010	<ul style="list-style-type: none"> • 168 patients with PsA (by 	<p>SF-36 (Chinese) with 8 domains</p>	<ul style="list-style-type: none"> • Item-total correlations > 0.4 	<ul style="list-style-type: none"> • Item-total correlations > 0.4 	<ul style="list-style-type: none"> • Item-own vs. item-other correlations should be > 2 SD 	<ul style="list-style-type: none"> • SF-36 PF has lower correlations with RE, MH (0.35, 0.33), 	<p>(+)</p>

	CASPAR) in Hong Kong •46.4% women •Mean (SD) age 47.7 (11.9) years •Mean (SD) duration of PsA 8.4 (7.3) years		•High item-own correlations than item-other correlations	•Item-own correlations exceed item-other correlations by 2 SD. •SF-36 PF correlates better with physical domains: RP, BP (r= 0.52, 0.53); than psychological domains RE, MH (r=0.35, 0.33)	•SF-36 domains worse than that of normal population	than RP, BP (r 0.52, 0.53); •Significant lower SF-36 PF (65.5 ± 25.4) than normal population (91.8 ± 12.9); p<0.0001	Working group remarks: All hypothesis satisfied
--	--	--	--	---	---	---	--

(+) indicates findings of the study had adequate performance of the instrument; (+/-) indicates equivocal performance; (-) indicates poor performance (less than adequate).
 Abbreviations.

ACR: American College of Rheumatology; ASDAS: Ankylosing Spondylitis Disease Activity Score ;BASDAI: Bath Ankylosing Spondylitis Disease activity index; BASFI: Bath Ankylosing Spondylitis Functional Index; ESR: erythrocytes sedimentation rate; EQ-5D: EuroQol-5 items; HAQ: Health assessment Questionnaire; -DI: disability index; DAPSA: Disease Activity index for Psoriatic Arthritis; F/U: follow up; MDHAQ: multidimensional HAQ; IQR: interquartile range; MDA: minimal disease activity; NA: not applicable; NRS: numeric rating scale; PsA: psoriatic arthritis; PsAID: Psoriatic Arthritis Impact of Disease Questionnaire; -FC: functional capacity; PsAQoL: PsA quality of life index; LOS: longitudinal observation study; LDA: low disease activity; MDA: minimal disease activity; PROMIS: Patient Reported outcome Measure Information System; -GPH: global physical health; -GH: global health; VAS: visual analogue scale; SF-36 PF: Medical Outcome Survey Short form -36 Physical Functioning domain; RP: role physical domain; BP: bodily pain domain; RE: role emotion domain; MH: mental health domain; PCS: physical function summary score; SD: standard deviation; SMD: standardized mean difference; SJC: swollen joint count; TJC: tender joint count.

Table B.2. Report of Studies of Test-Retest Reliability for SF-36 PF in PsA with OMERACT Filter 2.1

One dataset: Tillett 2019 (unpublished data)

Author, year	Study description		Results					Judgement
	Characteristics of sample	Characteristics of testing situation	Sample recruited and sample considered stable for analysis	Scores at baseline and retest	Statistic used	Results	Minimal detectable change (95%CI) $SEM = SD_{baseline} \times \sqrt{(1-ICC)}$ $MDC = 1.96 \times SEM \times \sqrt{2}$	Interpretation of adequacy (+, +/-, -)
Tillett 2019	<ul style="list-style-type: none"> Consecutive patients with PsA fulfilled CASPAR, recruited for validation of composite measures 	<ul style="list-style-type: none"> 1 week apart; Assumed no change in condition in stable patients without medication change 	<ul style="list-style-type: none"> 140 patients recruited for the main study, planned to recruit 30 patients who were stable and did not require medication change for test-retest reliability 31 patients provided complete dataset Mean (SD) age 54 (11) years Duration of PsA 5.7 (4.7) years 	Mean (SD) T1: 62.5 (36.3) T2: 61.6 (41.6) Mean difference = -0.89 (SD=27.6), p = 0.89 95% CI: -8.92 to 11.6	ICC Spearman's rho	ICC = 0.86 (95% CI: 0.69-0.93) Rho = 0.74 (p<0.01)	SEM = $36.3 \times \sqrt{(1-0.86)}$ = 13.6 MDC = $1.96 \times 13.6 \times \sqrt{2}$ = 37.6	(+)

(+) indicates findings of the study had adequate performance of the instrument; (+/-) indicates equivocal performance; (-) indicates poor performance (less than adequate).
Abbreviations.

CASPAR: Classification of Psoriatic Arthritis Study criteria; HAQ: Health assessment Questionnaire; -DI: disability index; DAPSA: Disease Activity index for Psoriatic Arthritis; F/U: follow up; ICC: intraclass correlation coefficients; MDC: Minimal detectable change; MDA: minimal disease activity; NA: not applicable; NRS: numeric rating scale; SEM: standard error of measurement; SF-36 PF: Medical Outcome Survey Short form -36 Physical Functioning domain; PCS: physical function summary score; SD: standard deviation; SMD: standardized mean difference.

Table B.3. Reporting of Evidence of Responsiveness (Longitudinal Construct validity) for SF-36 PF in PsA with OMERACT Filter 2.1

2 articles (1. Husted 1998; 2. Leung 2011).

Author Year	Study characteristics		Methods and results					
	Sample description (mean age, % gender, disease type/severity/dura- tion)	Study structure (type of study, timing of F/U, interventions)	Groups being contrasted**	Hypothesis of change (described expectation or anchor)	Anticipated results (NR = not reported in article)	Statistic used	Results observed	Interpretation of adequacy (+, +/-, -)
Husted 1998	<ul style="list-style-type: none"> • 80 patients with PsA attending Toronto center • 70 patients had completed data at 2 occasions • F/M = 27/43 • Mean (SD) age: 46.3 (11.4) years • Duration of PsA: 12.98 (7.68) years • NSAIDs only 27.1%, DMARDs 24.3%, MTX 31.4%, oral steroid 2.9% 	Two visits, F/U = 6 months	Within person change baseline and F/U	<p>Three anchors used as following</p> <p>The change of SF-36 domains, significant association with:</p> <ol style="list-style-type: none"> 4. Change in active joints number, 5. Change in damage joints number, 6. Change in global health <p>SF-36 domains can distinguish patients 'and clinicians' impression of a clinically meaningful change:</p> <ol style="list-style-type: none"> 6. Improvement in global health 7. 30% reduction in active joints 8. Deterioration in global health 9. 30% increase in active joints 10. (≥ 1) increase in damage joint 	<ul style="list-style-type: none"> • SF-36 domains should have significant associations with other change scores • SRM magnitude not hypothesized; but writing imply expecting SRM not low with the clinically meaningful change chosen. • SRM responsiveness defined as low: 0.2 low; 0.5 moderate; 0.8 high. 	SRM	<p>Change in SF-36 PF has significant association with:</p> <ol style="list-style-type: none"> 4. Change in # active joints ($B=-0.37$, $p=0.002$, $R^2=0.13$), 5. Change in # damage joints ($B=0.11$, $p=0.317$, $R^2=0.01$), 6. Change in perceived general health ($B=-0.55$, $p=0.0001$, $R^2=0.31$) <p>SRM for SF-36 PF</p> <ol style="list-style-type: none"> 6. Improvement in global health (0.642) 7. 30% reduction in active joints (0.670) 8. Deterioration in global health (-0.345) 9. 30% increase in active joints (-0.624) 10. (≥ 1) increase in damage joint (0.199) 	<p>(+)</p> <p>Working group remarks:</p> <p>2 out of 3 (67%) of the correlation hypothesis achieved.</p> <p>4 out of 5 (80%) of the SRM hypothesis achieved.</p> <p>Overall achieved 73.5% of the hypothesis</p>

Leung 2011	<ul style="list-style-type: none"> • 20 PsA patients with active disease who started TNFi • Mean (SD) age 49.8 years • Mean (SD) duration of PsA: 7.8 years • 40% women • 68/66 TJC/SJC: 10.4/3.7 	<ul style="list-style-type: none"> • After 12 weeks, 11 patients discontinued TNFi due to cost • Patients F/U at baseline, week 12, 24, 36 and 52 	<p>Within person change baseline and fu time points</p> <p>With 4 F/U time points, 78 sets of PRO (from 20 patients) were analysed</p>	<p>Changes in SF-36 domains and change in anchor (current health status compared to last visit: much better, slightly better, similar, slightly worse, much worse.</p>	<p>Correlation between change in SF-36 domains and change in anchor would be significantly, with Rho at least >0.3</p> <p>No hypothesis for effect size</p> <p>Effect defined as small if < 0.2, medium if 0.3–0.5, and large if > 0.5.</p>	<ul style="list-style-type: none"> • Rho for change score • Cohen's d Effect size (ES) • SRM 	<p>Change in SF-36 PF and change in anchor correlation, Rho= -0.34, p <0.01</p> <p>ES = 0.35 SRM = 0.37</p>	<p>(+/-)</p> <p>Working group remark:</p> <ul style="list-style-type: none"> • Correlation with anchor achieved statistical significance • The ES/ SRM were in the medium range which is reasonable. Considering that data were derived from patient group reporting only "slight" change on F/U. • Overall, the working group thinks the article can be considered. Although no formal ES magnitude was explicitly stated, an ES not in the low range was implied. However, as there was no formal hypothesis for ES magnitude, working group recommends caution in interpretation
------------	--	---	--	--	--	---	--	--

***Groups being contrasted can include: Within person change, Between person differences (if used for the study of responsiveness) or Between group differences of within person change (contrasting relative change between two groups).*

(+) indicates findings of the study had adequate performance of the instrument; (+/-) indicates equivocal performance; (-) indicates poor performance (less than adequate).

Abbreviations.

CASPAR: Classification of Psoriatic Arthritis Study criteria; DAPSA: Disease Activity index for Psoriatic Arthritis; DMARDs: Disease modifying anti-rheumatic drugs; ES: effect size; ESR: erythrocytes sedimentation rate; EQ-5D: EuroQoL-5 items; HAQ: Health assessment Questionnaire; -DI: disability index; F/U: follow-up; NA: not applicable; NRS: numeric rating scale; NS: not statistically significant; NSAID: non-steroidal anti-inflammatory drug; PsA: psoriatic arthritis; PsAID: Psoriatic Arthritis Impact of Disease Questionnaire; -FC: functional capacity; PsAQoL: PsA quality of life index; LOS: longitudinal observation study; LDA: low disease activity; MDA: minimal disease activity; MTX: methotrexate; PROMIS: Patient Reported outcome Measure Information System; -GPH: global physical health; -GH: global health; SEM: standard error of measurement; SRM: standardized response mean; SF-36 PF: Medical Outcome Survey Short form -36 Physical Functioning domain; PCS: physical function summary score; PRO: patient-reported outcome; SD: standard deviation; SMD: standardized mean difference; SJC: swollen joint count; TJC: tender joint count; TNFi: tumor necrosis factor inhibitors; VAS: visual analogue scale.

Table B.4. Reporting of Evidence of Responsiveness (Clinical Trial Discrimination) for SF-36 PF in PsA with OMERACT Filter 2.1

2 papers included: Mease 2011, Gladman 2017

2 papers excluded for evidence synthesis as not passed quality assessment

Author/ year/ (study acronyms)	Intervention/ comparator (sample size, N)	Sample size (% women)	PsA duration (years)	TNFi IR	MTX use	Primary outcome/ time point	P value of Δ scores of intervention vs. PCB	Effect sizes at primary endpoint (unless specified)	Fulfills <i>a priori</i> hypothesis [‡]	Adequacy of instrument performance (+, +/-, -)
Kavanaugh 2006 (IMPACT2)	IFX 5mg/kg vs. PCB (N=200)	N=200 (39.0)	8	0%	46%	ACR20/ Week 14	<0.001	Insufficient data for effect size calculation	2	NA
Mease 2017 (OPAL Broaden)	TOF (2 doses) vs. PCB vs. ADA (N= 422)	N=422 (53.0)	6.1	0%	83.9%	ACR20/ 3 months	All NS	SRM at 3 months: TOF 10mg = 0.64 TOF 5mg = 0.64 PCB = 0.23 ADA = 0.58	1, 3	(+/-)
Gladman 2017 (OPAL Beyond)	TOF (2 doses) vs. PCB (N=395)	N=395 (55.0)	9.4	100%	73.6%	ACR20, Δ in HAQ-DI/ 3 months	<0.05 for both doses	SRM at 3 months: TOF 10mg = 0.53 TOF 5mg = 0.64 PCB = 0.22	1, 3	(+)
Mease 2018 (EQUATOR) Phase II	FIL 200mg vs. PCB (N=131)	N=131 (50.4)	7.0	0%	54.2% (any csDMARD 74%)	ACR20/ Week 16	0.0009	Insufficient data for effect size calculation	2	NA

[‡] SRM calculated using percentage change score and SD of percentage change; [§] SRM for improvement, a negative value indicate deterioration; [‡] Effect sizes estimated based on mean and SD of change scores calculated from median and IQR from original publication; * early escape for patients with inadequate response in the control group to active treatment group at week 16; ** option to switch TNFi at Week 24; (+) indicates findings of the study had adequate performance of the instrument; (+/-) indicates equivocal performance; (-) indicates poor performance (less than adequate); NA: not applicable.

Abbreviations: Δ : change; ACR: American College of Rheumatology Response criteria; ADA: adalimumab; bDMARDs: biological disease modifying anti-rheumatic drugs; csDMARDs: conventional synthetic disease modifying anti-rheumatic drugs; ES₂: Effect size 2 (the mean difference divided by the pooled standard deviation, i.e. Cohen's *d*); FIL: filgotinib; HAQ-DI: Health Assessment Questionnaire – Disability Index; IFX: infliximab; IQR: interquartile range; MTX: methotrexate; NA: not applicable; PCB: placebo; PF: physical functioning domain of SF-36; SD: standard deviation; SF-36: Medical Outcomes Study 36-item Short Form Survey; SRM: Standardized response mean (mean difference divided by the standard deviation of the differences between baseline and assessment end point); TOF: tofacitinib; vs.: versus.

[‡]A *priori* hypothesis:

1. At the primary endpoint/end of double blinded phase, patients given bDMARDs have significant change in SF-36 PF, whereas patients on placebo arm do not (except for alefacept and clazakizumab where no significant difference is expected)
2. The change scores of SF-36 PF among patients given bDMARDs are significantly higher than those of the placebo arm
3. Within individual trial, the effect sizes of change scores of SF-36 PF are higher in the bDMARD arms compared to the MTX or csDMARD arms, but do not differ significantly with different bDMARD doses (or with TNFi as comparison).
4. If data for subgroup analysis is available, the effect sizes of change scores of SF-36 PF are higher in TNF naïve versus TNF exposed subgroup

Table B.5. Reporting of Evidence of Threshold of meaning for SF-36 PF in PsA with OMERACT Filter 2.1.

Leung 2011

Author Year	Study characteristics		Methods and results						Interpretation of adequacy (+, +/-, -)
	Sample description (mean age, % gender, disease type/severity/ duration)	Study structure (type of study, timing of fu, interventions)	Threshold assessed (MID, MCID, PAS, LDA)	Method (anchor or distributi onal)	Threshold method: anchor used and categories in that anchor	Definition of threshold of meaning using this approach	Threshold of meaning (specify value), AUC if available.	% of sample meeting/excee ding this threshold	
Leung 2011	<ul style="list-style-type: none"> • 20 PsA patients with active disease who started TNFi • Mean (SD) age 49.8 years • Mean (SD) Duration of PsA: 7.8 years • 40% women • 68/66 TJC/SJC: 10.4/3.7 	<ul style="list-style-type: none"> • Patients F/U at baseline, week 12, 24, 36 and 52 • With 4 F/U time points, 78 sets of PRO (from 20 patients) were analysed 	MID	Anchor	Current health status compared to last visit: <ul style="list-style-type: none"> • much better, • slightly better, • similar, • slightly worse, • much worse. 	Mean change of scores in those rated “slightly better” or “slightly worse” groups Correlation between change in sf-36 domains and anchor	Mean (SD) MID for improvement (n=17): 4.41 (14.99) MID for deterioration (n=21): -6.25 (18.77) No significant differences in MID when stratified by sex or higher baseline HAQ>1.0. Correlation of change in SF-36 PF and Anchor = -0.34, p<0.01	Much better: 34.6% Slightly better: 21.3% Similar: 10.3% Slightly worse: 26.9% Much worse: 6.4% To establish the MCID for improvement/worsening for PROs	(+/-) Working group remarks: This paper only used one statistical method to estimate MID with no sensitivity/specificity analysis. The sample size was very small. Thus, quality of paper is borderline The MID for HAQ-DI improvement - 0.27 align with results found in another paper in TNFi treatment population, given support to the results. This is the only study that report estimate of MID

									<p>for SF-36 PF, which would serve as reference values for further study.</p> <p>Due to some concerns in quality assessment, the working group give (+/-) as performance adequacy, and recommends interpreting evidence with caution</p>
--	--	--	--	--	--	--	--	--	--

(+) indicates findings of the study had adequate performance of the instrument; (+/-) indicates equivocal performance; (-) indicates poor performance (less than adequate).
Abbreviations.

ACR: American College of Rheumatology; AUC: area under curve; CASPAR: Classification of Psoriatic Arthritis Study criteria; CI: confidence interval; ESR: erythrocytes sedimentation rate; EQ-5D: EuroQol-5 items; HAQ: Health assessment Questionnaire; -DI: disability index; DAPSA: Disease Activity index for PSoriatic Arthritis; F/U: follow up; ICC: intraclass correlation coefficients; MDHAQ: multidimensional HAQ; IQR: interquartile range; MDC: Minimal detectable change; MDA: minimal disease activity; MID: minimally important difference; NA: not applicable; NRS: numeric rating scale; PsA: psoriatic arthritis; PsAID: Psoriatic Arthritis Impact of Disease Questionnaire; -FC: functional capacity; PsAQoL: PsA quality of life index; LOS: longitudinal observation study; LDA: low disease activity; MDA: minimal disease activity; MCID: minimally clinically important difference; MID: minimally important difference; PROMIS: Patient Reported outcome Measure Information System; -GPH: global physical health; -GH: global health; VAS: visual analogue scale; PASS: patient acceptable symptom status; REM: remission; ROC: receiver operating characteristic; SEM: standard error of measurement; SF-36 PF: Medical Outcome Survey Short form -36 Physical Functioning domain; PCS: physical function summary score; PRO: patient-reported outcome; SD: standard deviation; SMD: standardized mean difference; SJC: swollen joint count; TJC: tender joint count; TNFi: tumor necrosis factor inhibitors.