SUPPLEMENTARY APPENDIX A

2019 American College of Rheumatology Recommended Patient Reported Functional Status Assessment Measures in Rheumatoid Arthritis

Members of the ACR Functional Assessment Status Measure Workgroup

Claire Barber MD, PhD, FRCPC - **Lead author**
Laura Cappelli MD
Aileen M. Davis, PhD
Linda Ehrlich-Jones, PhD, RN
Donna Everix, MPA, BS, PT
Kaleb Michaud, PhD - **Chair**
Carter Thorne, MD, FRCPC
Jinoos Yazdany, MD, MPH
JoAnn Zell, MD

Alex Limanni, MD – Quality Measures Subcommittee representative and voting member
Lisa Suter, MD – Quality Measures Subcommittee representative

Regina Parker – ACR administrative staff assigned to group
Amy Turner – ACR administrative staff assigned to group

Medline Search Strategy

The Medline search strategy is described below. This strategy uses MeSH terms and keywords across three themes: #1 construct search (for assessment of functional status), #2 population search (rheumatoid arthritis) and #3 instrument search (including terms for instruments of interest e.g., questionnaires, etc.). The Boolean search operator "AND" was used to combine the 3 search themes.

1. exp Health status/
2. 'Health level*'.tw,kw.
3. 'Health Status*'.tw,kw.
4. 'Level* of health'.tw,kw.
5. exp Disability evaluation/
6. (Disability adj2 assessment*).tw,kw.
7. (functional adj2 assessment*).tw,kw.
8. (Disability adj2 evaluation*).tw,kw.
9. exp Health status indicator/
10. 'Health status index*'.tw,kw.
11. 'Health status indic*'.tw,kw.
12. exp Severity of illness index/
13. 'Severity of illness ind*'.tw,kw.
14. exp Activities of daily living/
15. daily life activit*.tw,kw.
16. ADL*.tw,kw.
17. (Activit* adj2 living).tw,kw.
18. exp patient outcome assessment/
19. 'Patient-centered outcome* research'.tw,kw.
20. 'Patient reported outcome*'.tw,kw.
21. 'Patient perspective*'.tw,kw.
22. 'outcome* research'.tw,kw.
23. (outcome* adj2 assessment*).tw,kw.
24. 'functional status'.tw,kw.
25. 'function* impair*'.tw,kw.
26. 'Health assessment questionnaire'.tw,kw.
27. HAQ*.tw,kw.
28. MHAQ.tw,kw.
29. MDHAQ.tw,kw.
30. PROMIS.tw,kw.
31. 'Short Form 36'.tw,kw.
32. SF-36.tw,kw.
33. or/1-32

34. exp "Surveys and Questionnaires"/
35. Survey*.tw,kw.
36. Questionnaire*.tw,kw.
37. Index*.tw,kw.
38. Scale*.tw,kw.
39. Instrument*.tw,kw.
40. tool*.tw,kw.
41. diar*.tw,kw.
42. assessment*.tw,kw.
43. 'self-report*'.tw,kw.
44. measure*.tw,kw.
45. prom.tw,kw.
46. checklist*.tw,kw.
47. rating.tw,kw.
48. or/34-47
49. instrumentation.fs.
50. methods.fs.
51. validation studies.pt.
52. comparative study.pt.
53. exp Validation studies/
54. exp "Outcome Assessment (Health Care)"/
55. outcome measure*.tw,kw.
56. validation Stud*.tw,kw.
57. Validate.tw,kw.
58. Validity.tw,kw.
59. valid*.tw,kw.
60. (homogeneity or homogeneous).tw,kw.
61. ((minimal* or clinic*) and (important or significant or detectable) and (change or difference)).tw,kw.
62. 'minimal* real difference*'.tw,kw.
63. 'ceiling effect'.tw,kw.
64. 'floor effect'.tw,kw.
65. detect* change*.tw,kw.
66. exp "reproducibility of results"/
67. reproducib*.tw,kw.
68. (reliab* or unreliab*).tw,kw.
69. (reliab* and (test or retest)).tw,kw.
70. responsiveness*.tw,kw.
71. 'test-retest'.tw,kw.

72. (test adj1 retest).tw,kw.
73. discriminant analysis.tw,kw.
74. exp observer variation/
75. 'observer variation'.tw,kw.
76. exp Psychometrics/
77. Psychometr*.tw,kw.
78. clinometr*.tw,kw.
79. clinimetr*.tw,kw.
80. coefficient.tw,kw.
81. 'internal consistency'.tw,kw.
82. (cronbach* and alpha*).tw,kw.
83. 'item correlation*'.tw,kw.
84. 'item selection*'.tw,kw.
85. 'item reduction*'.tw,kw.
86. agreement.tw,kw.
87. precision.tw,kw.
88. imprecision.tw,kw.
89. 'precise values'.tw,kw.
90. stability.tw,kw.
91. interrater.tw,kw.
92. 'inter rater'.tw,kw.
93. intrarater.tw,kw.
94. 'intra rater'.tw,kw.
95. intertester.tw,kw.
96. 'inter tester'.tw,kw.
97. intratester.tw,kw.
98. 'intra tester'.tw,kw.
99. interobserver.tw,kw.
100. 'inter observer'.tw,kw.
101. 'intra observer'.tw,kw.
102. interexaminer.tw,kw.
103. 'inter examiner'.tw,kw.
104. intraexaminer.tw,kw.
105. 'intra examiner'.tw,kw.
106. interindividual.tw,kw.
107. 'inter individual'.tw,kw.
108. intraindividual.tw,kw.
109. 'intra individual'.tw,kw.
110. interparticipant.tw,kw.

111. 'inter participant'.tw,kw.
112. intraparticipant.tw,kw.
113. 'intra participant'.tw,kw.
114. (intertechninican or inter-technician or intratechnician or intra-technician).tw,kw.
115. (interassay or inter-assay or intraassay or intra-assay).tw,kw.
116. kappa*.tw,kw.
117. 'coefficient of variation'.tw,kw.
118. repeatab*.tw,kw.
119. ((replicab* or repeated) and (measure* or findings or result* or test*)).tw,kw.
120. tests.tw,kw.
121. (generaliza* or generalisa*).tw,kw.
122. concordance.tw,kw.
123. (intraclass and correlation).tw,kw.
124. discriminative.tw,kw.
125. 'known group'.tw,kw.
126. 'factor analys*'.tw,kw.
127. 'factor structure*'.tw,kw.
128. 'dimension*'.tw,kw.
129. 'multitrait scaling analys*'.tw,kw.
130. (error* and (measure* or correlat* or evaluat* or accuracy or accurate or precision or mean)).tw,kw.
131. 'individual variability'.tw,kw.
132. 'interval variability'.tw,kw.
133. 'rate variability'.tw,kw.
134. (variability and (analysis or values)).tw,kw.
135. (uncertainty and (measurement or measuring)).tw,kw.
136. 'standard error of measurement'.tw,kw.
137. sensitiv*.tw,kw.
138. responsive*.tw,kw.
139. (limit and detection).tw,kw.
140. interpretab*.tw,kw.
141. (small* and (real or detectable) and (change or Difference)).tw,kw.
142. 'meaningful change'.tw,kw.
143. 'item response model'.tw,kw.
144. irt.tw,kw.
145. rasch.tw,kw.
146. 'differential item functioning'.tw,kw.
147. 'cross-cultural equivalence'.tw,kw.
148. 'detect change'.tw,kw.

149. subscale*.tw,kw.

150. item discriminant.tw,kw.

151. interscale correlation*.tw,kw.

152. error*.tw,kw.

153. DIF.tw,kw.

154. "computer adaptive testing".tw,kw.

155. "item bank".tw,kw.

156. or/34-155

157. exp arthritis, rheumatoid/

158. rheumatoid arthritis.tw,kw.

159. 157 or 158

160. 33 and 48 and 156 and 159

161. 160 not ("addresses" or "bibliography" or "case reports" or "comment" or "directory" or "editorial" or "festschrift" or "interview" or "lectures" or "legal cases" or "legislation" or “letter" or "news" or "newspaper article" or "patient education handout" or "popular works" or "congresses" or "consensus development conference" or "consensus development conference, nih" or "practice guideline").pt. not (animals/ not humans.sh.)

162. limit 161 to english

Abbreviations

ACR	American College of Rheumatology
CAT	Computer Adaptive Testing
CTT	Classical Test Theory
COSMIN	COnsensus-based Standards for the selection of health Measurement INstruments
HAQ-DI	Health Assessment Questionnaire Disability Index
HAQ-II	Health Assessment Questionnaire II
IRT	Item Response Theory
FSAM	Functional Status Assessment Measure
MHAQ	Modified Health Assessment Questionnaire
MDHAQ	Multidimensional Health Assessment Questionnaire
PROMIS	Patient-Reported Outcomes Measurement Information System
RADAM	Rheumatoid Arthritis Disease Activity Measure
RISE	Rheumatology Informatics System for Effectiveness

Supplementary Table 1. Characteristics of the patient-reported functional status assessment measures

Functional status measure	**No. of Items and Domains**	**Domains**	**Response options, range**	**Assistive devices/help from others**	**Recall**	**Range/ Interpretation**
ADL-Q (1)	47 Items/ 7 Domains	Easting & drinking, mobility, going to the toilet, dressing, personal hygiene, grooming, communication	7 response categories/ Ability measures expressed in logits.	Both captured in response items	PADL tasks based on ADL performances within past 24hrs IADL tasks based on performance within last 7 days	
ALDS (2, 3)	77 Items	Extensive list of individual questions no domains specified, includes ADLs and IADLs	"Can carry out" or "I cannot carry out the activity". Range 0-100	Not addressed	"Are you able to…" no time specified on tool	Range of scores from 0-100, algorithms for scoring described separately (uses logits)
APaQ (4)	2 Items	Days RA kept a person from usual activities. How often was a person able to perform usual activities completely.	Question 1: 0-30 days. Question 2: range 1-6	Not addressed	30 Days	
(Modified) Barthel Index§ (5, 6)	10 Items	Feeding, washing and dressing, get up out of bed or chair, bathing, ascending and descending stairs, walk 50 yards, control bowel and bladder	Weighted score system, value assigned to each item is 5, 10 or 15, depending on the time and amount of assistance required, except for the item of mobility (15 if full mobility even if use of wheelchair).	Both captured in response items	Current performance	0-100, higher scores indicate increased independence
Bradley et al. (7)	41 Items/5 Domains	Mobility, bending down, dexterity, bending arm, reaching up	Items scored on the WHO disability severity scale with a new category for "performance in an abnormal manner": 8-point scale- 0 if no difficulties were encountered, 1 (difficulty), (2) abnormal performance, (3) aids were required, (4) aids with a helping hand, (5) personal assistance, (6) personal help plus an aid, (7) activity impossible	Assistive devices captured in response items	NS	Aggregated disability score for each functional group
CIAQ-FI (8)	10 Items/Domains NS	Transfer (toilet), grip strength, dressing, standing, waiting, reaching, walking, stairs, housework, move heavy objects	4-point scale: without any difficulty, 1 (with some difficulty), 2 (with much difficulty), 3 (unable to do)	Not addressed	1 week	
CSSRD-FAS (9)	35 Items/5 Domains	Personal care; mobility; transfer; work/chores; assistive devices	4-point scale: 0 (can), 1 (with effort), 2 (with extreme effort), 3 (cannot)	Assistive devices captured in domain	NS	Weights assigned to each domain in an overall summary of functional ability totaling 100%. Weights: Personal care (0.43), mobility (0.17), transfer (0.12), Work/Chores (0.245), Assistive devices (0.035). Total 1.0
EQUAD (10-12)	102 Items/11 Domains	Eating, transfer, toileting, dressing, bathing, cooking, mobility indoors, cleaning, washing/clothes care, mobility outdoors/ shopping, communication.	4-point scale 0 (without any difficulty), 1 (with some difficulty), 2 (with much difficulty), and 3 (unable to do).	Instructed to complete with and without use of devices	Same day	Rasch analysis used to transform ordinal score to obtain linear measure. Higher scores more functional disability.

Functional status measure	**No. of Items and Domains**	**Domains**	**Response options, range**	**Assistive devices/help from others**	**Recall**	**Range/ Interpretation**
FALQ (13)	41 Items (+ 1 open-ended question)/Domains NS	Stand, arising, jump, run, squat, cutting toenails, putting on socks, write, picking up coins, buttoning, opening jars, using cutlery, making a fist, reaching, throwing, lifting heavy things, toileting, personal care (brush teeth, wash face), stooping/bending, sexual intercourse, driving, dancing, hiking, golf, bicycle, bowl, riding horse, tennis, swim, ski, knitting, shopping.	4-point scale: 1 (yes), with no difficulty; 2 (yes but with some difficulty), 3 (cannot do it), 4 (don't know)	Not addressed	NS	Higher scores more functional disability (ignoring scored 4 presumably)
GARS (14, 15)	18 Items/divided into ADL and IADL major domains with 18 questions total	ADLS: Dressing; transfers (in/out bed; up from a chair, on/off toilet); personal care; (wash face/hands; wash/dry whole body; take care of feet and toenails); feeding; mobilizing around the house; stairs; walking outdoors; IADLs: meal preparation, light or heavy household chores; laundry; making beds; shopping	1 (yes, I can do it fully independently without any difficulty); 2 (yes, I can do it fully but with some difficulty); 3 (yes, I can do it fully independently but with great difficulty); 4 (no I cannot do it fully independently, I can only do it with someone's help).	Accounted for partially in scoring. Looks at actual disability (includes use of devices implicit in assessment e.g., if no difficulty walking with a cane then first category of difficulty selected)	Last week	Add sum from each of the items. In some studies, category 3+4 collapsed to make it more comparable to the HAQ. Higher scores, more functional disability
Lee et al. (16)	17 Items/Domains NS	Turn head side to side, comb hair, close drawers, open doors, lift teapot, lift cup with one hand, turn key in lock, cut meat with knife, butter bread, wind watch, walk, walk without help, crutches, walking stick, stairs (up/down), stand with knees straight, stand on toes, bend down to pick something up off floor	0 (no difficulty in performing the movement); 1 (ability to perform the movement but with difficulty); 2 (complete inability to perform the movement). Max total score 40 (there are 3 subsets under walking)	Assistive devices specified for walking items only	NS	Sum of item scores. higher scores, more functional disability.
MAL (17)	19 Items/ 12 Domains	These domains were initially considered (however, unclear what was finally included): Personal care, mobility, using your normal means of transport, household activities, household maintenance, social activities, relaxation, paid employment, hobbies, caring for others, sexual relationships, voluntary work.	Scale 1-5: with 1 representing the least degree of activity limitation and 5 the highest. Minimum of 15 and max of 75.	Unclear (probably no based on available information)	Unclear	Sum of item scores. higher scores, more functional disability.
PS-ADL (18)	39 Items/12 Domains	Eating and drinking, Mobility, Going to the toilet, Dressing, Personal hygiene, Grooming, Communication, Transportation, Cooking, Shopping, Cleaning, and Washing	Difficulty: 0-3: 0 (without difficulty); 1 (without difficulty with assistive devices), 2 (difficulty, with or without assistive devices) and 3 (unable to do). Satisfaction:0-2 (measures whether patient satisfied with their performance), 0 (satisfied), 1 (could have been better), 2 (dissatisfaction with performance)	Assistive devices captured in response items	1 week	Scales for performance (0-3) and satisfaction (0-2) calculated separately with mean for each subscale.

Functional status measure	**No. of Items and Domains**	**Domains**	**Response options, range**	**Assistive devices/help from others**	**Recall**	**Range/ Interpretation**
ROAD (19-21)	12 Items/ 3 Domains	Assesses function in 3 domains: fine movements (close hand, hand shake, buttons, open jars, reach), locomotor activities of the lower extremities (stand, walk, stairs, in/out of a car), activities involving upper and lower extremities (wash/dry body, run errands and shop, housework/job).	Scale from 0-4:0 (without any difficulty), 1 (with slight difficulty), 2 (with some difficulty), 3 (with great difficulty), 4 (unable to do)	Not addressed	1 week	Mathematical normalization procedure done to express scale in range from 0-10 with 0 representing better health status and 10 representing poorer health status. Presented as 3 subscales.
S-VLA (22)	14 Items. For each item patients asked 2 questions: A) how much difficulty do you have with this activity because of your RA? B) when you perform the activity, do you have to make accommodations such as (see paper for limitations).	Basic needs, meal preparation, light housework, heavier housework, gardening, caring for family members, attending social events, getting around in home, walking outside home, leisure outside home, hobbies, physical recreation, traveling out of town,	Scale from 0-4: difficulty =0 and no accommodations score =0; difficulty=0 and any accommodations: score=1; difficulty=1 (some difficulty), regardless of accommodations: score=2; difficulty=2(a lot of difficulty, regardless of accommodations: score=3; difficulty=3 (unable to perform), regardless of accommodations: score=4.	Accommodation accounted for in scoring	NS	Sum of item scores. higher scores, more functional disability.
VAS Physical Function (23)	N/A VAS scale	N/A VAS scale	Circle the number that best describes the difficulty you had in doing daily physical activities due to your rheumatoid arthritis during the last 48 hours 0=”none” 10=”extreme”	N/A	48 hours	Higher scores more functional disability
VAS Function (F) Scale (24)	N/A VAS scale	N/A VAS scale	Anchored at one end "No functional limitations"=0 and at the other end with "severe functional limitations"=100	N/A	NS	Higher scores more functional disability
Alternative HAQ (no assistive devices)	Same as for HAQ-DI	Same as for HAQ-DI	Same as for HAQ-DI	N/A	1 week	Same as HAQ-DI
AHAQ [4] (25)	Same as for HAQ-DI	Same as for HAQ-DI	Same as for HAQ-DI	N/A	1 week	Same as HAQ-DI
HAQ-DI “legacy” (26)[1]	41 Items/8 Domains/20 Specific functions	Dressing and grooming, arising, eating, walking, hygiene, reaching, gripping, and errands and chores	4-point Likert scale: 0 (without difficulty), 1 (with some difficulty), 2 (with much difficulty), and 3 (unable to do).	13 questions/8 questions	1 week	0-3/higher scores indicate more disability
MHAQ[2] (27)	8 Items/8 Domains	Same as for HAQ-DI	Same as for HAQ-DI	Not addressed	3 months	Same as HAQ-DI
MDHAQ (28, 29)	10 Items/10[3] Domains	8 items (same as MHAQ) + “walk 2 miles” and “participate in recreational activities and sports as you would like”	Same as for HAQ-DI	Not addressed	1 week	Same as HAQ-DI
HAQII (30)	10 Items/10 Domains	5 from original HAQ-DI and 5 additional items	Same as for HAQ-DI	Not addressed	1 week	Same as HAQ-DI
PROMIS (includes only those found used during validation English language studies for RA populations)						

Functional status measure	**No. of Items and Domains**	**Domains**	**Response options, range**	**Assistive devices/help from others**	**Recall**	**Range/ Interpretation**
PROMIS PF10a (31)	10 items/ Domains NS	Item themes: Vigorous activities, walking >1 mile, stairs, carrying groceries, bending/kneeling/ stooping, vacuuming/yard work, dressing, shampoo hair, wash and dry body, on/off toilet	Scale of 1-5: 1 (not at all), 2 (very little), 3 (somewhat), 4 (quite a lot), 5 (cannot do) for first 5 items. Scale of 1-5: 1 (without any difficulty), 2 (with a little difficulty), 3(with some difficulty), 4 (with much difficulty), 5(unable to do)	No	Current abilities	0-100 unit scale/higher scores indicate more disability
PROMIS 20-“item static”/ SF, also called PF20a (32)	20 items	Item themes: vacuuming/yard work, open heavy door/ dressing/tying shoelaces/buttons, washing your back, drying back, sit edge of bed, wash and dry body, get in/out of car, squeeze toothpaste, hold plate of food, run short distance, shampoo hair, on/off toilet, transfer bed to chair, vigorous activities running/lifting heavy objects/sports, kneeling/bending/stooping, carrying groceries, physical labor, walking >1mile, climbing stairs	Same as PF10a	No	Current abilities Current abilities	0-100 unit scale/higher scores indicate more disability
PROMIS CAT (33, 34)	Terminated after 1set number of personalized items	PROMIS Item Bank	Same as other PROMIS questions	Potentially variable as assistive devices available in full item bank	Current abilities	0-100 unit scale/higher scores indicate more disability

ADL: Activities of daily living; ADL-Q: Activities of daily living questionnaire; ALDS: Academic Medical Center Linear Disability Score; APaQ: Activity Participation Questionnaire; CIAQ-FI: combined inflammatory arthritis questionnaire-Functional Impairment; CSSRD-FAS: Cooperative Systematic Studies of Rheumatic Diseases group; EQUAD: Evaluation of Daily Activity Questionnaire; FALQ: Functional activity level questionnaire; HAQ: Health Assessment Questionnaire; MAL: Measure of Activity Limitation; MHAQ: Modified HAQ; MD HAQ: Multidimensional HAQ; IADL: Instrumental Activities of daily living; NS: Not Specified; PADL: Personal Activities of Daily Living; PS-ADL: Performance and Satisfaction in Activities of Daily Living; ROAD: Recent-Onset Arthritis Disability questionnaire S-VLA: shortened version of the Valued Life Activities Scale.

[§§] This is considered a modified Barthel Index as the original required administration by a therapist

[1] Orginal HAQ described by Fries in 1980 included 5 domains: death, disability, discomfort, drug toxicity, dollar costs. The “Legacy” HAQ or HAQ DI refer to the disability portion of this original scale.

[2] Original MHAQ description by Pincus in 1983 (27) included questions concerning perceived patient satisfaction regarding activities of daily living as well as perceived change in degree of disability.

[3] Original MDHAQ had 14 items (28) and also included questions about psychological aspects of disease; there is also a 10-ADL MDHAQ (29)

[4] Same as original HAQ but scores were generated for the item categories making up the disability index by taking the mean of the item scores in a category instead of the worst item score (like the original HAQ-DI); the disability index was the mean of the alternative category scores.

Supplementary Table 2. Characteristics of the included studies

Author (year)	**Performance Measure(s)**	**Mean Age Years ± SD (range)**	**Population (setting)**	**N (Country)**	**Measurement Property(ies) Evaluated**
Performance Measure Evaluated: HAQ-DI (and original)*					
Bombardier 1991 (35)	HAQ-DI	Placebo: 51±0.9 Auranofin: 50±0.9	RA (RCT**)	303*** (Canada)	Responsiveness
Brown 1984 (36)	HAQ-DI + pain	53 (SD NR)	RA (subset of community-based "study group")	48 (USA)	Structural validity Hypothesis testing
Buchbinder 1995 (37)	HAQ-DI	53.3±1.1	RA (RCT)	144, HAQ subgroup 78 (Canada)	Responsiveness
Cole 2005 (38)	HAQ-DI	51±13	RA (LCD)	278 (USA & Mexico)	Structural Validity
Fitzpatrick 1989 (39)	HAQ-DI	55 ±11.4	RA (Single practice)	105 (UK)	Hypothesis testing Responsiveness
Fitzpatrick 1993 (40)	HAQ-DI	56 ± 12.1	RA (Single practice)	102 (UK)	Hypothesis testing Responsiveness
Fitzpatrick 1992 (41)	HAQ-DI	56±12.1	RA (Single practice)	101 (UK)	Responsiveness
Fries 1980 (26)	HAQ (original)	For reliability testing: 54 (SD NR) For validity testing: 57 (SD NR)	RA (University Clinics, RTC)	48 (USA)	Internal Consistency, Reliability, Hypothesis testing (convergent validity), Structural validity
Goeppinger 1988 (42)	HAQ (assumed DI)	Virginia whole sample: 60.7 ±13.5 Stanford whole sample: 60.4 ±13.1	RA, OA, Diabetes (Multi-center)	365 (USA)	Internal Consistency, Reliability, Hypothesis testing (convergent validity), Content validity

Author (year)	**Performance Measure(s)**	**Mean Age Years ± SD (range)**	**Population (setting)**	**N (Country)**	**Measurement Property(ies) Evaluated**
Greenwood 2001 (43)	HAQ (assumed DI)	64 (48-83)	RA (Single practice)	247 (UK)	Responsiveness; Interpretability
Häkkinen 2005 (44)	HAQ-DI	58 (21-83)	RA (Physiotherapy practice of referred inpatients)	304	Internal Consistency
Hawley 1992 (45)	HAQ-DI	Group 1: 56.0±12.3 Group 2: 50.8±12.5	RA (Single practice)	6 months: 233 10 years: 157 (USA)	Responsiveness
Hendrikx 2015 (46)	HAQ-DI	59.1±13.0	RA (Single practice LCD)	469 (Netherlands)	Interpretability
Kosinski 2000 (47)	HAQ-DI	Age <45: (n=110, 16%) Age 45-64: (n=289, 56%) Age 65+: (n=194, 28%)	RA (2 RCTs)	693 (USA)	Interpretability
Lassere 2001 **(48)**	HAQ-DI	Study B 61; Study C 56 (SD NR)	RA (Rheumatology clinics, 2 sub-studies)	Study B n=42; Study C n=26 (Australia?)	Reliability; Measurement Error
Linde 2008 (49)	HAQ- DI	P1: median 59 (19-87) P2: median 60 (22-82)	RA (Study cohort, LCD)	Sample 1: 200 Sample 2: 150 (Denmark)	Internal consistency; Reliability, Hypothesis Testing, Responsiveness
Marra 2005 (50)	HAQ (assumed DI)	NR	RA (Rheumatology Clinics)	320 (Canada)	Reliability, Responsiveness
Marra 2005 (51)	HAQ (assumed DI)	61.5±25.9	RA (Rheumatology Clinics)	313 (Canada)	Hypothesis Testing
Pope 2009 (52)	HAQ-DI	60.5±13.6 (17-90)	RA (University-based clinic)	225 (Canada)	Responsiveness

Author (year)	**Performance Measure(s)**	**Mean Age Years ± SD (range)**	**Population (setting)**	**N (Country)**	**Measurement Property(ies) Evaluated**
Redelmeier 1993 (53)	HAQ (assumed DI)	Initial group: 63 (23-71) Replication group: 64 (30-80)	RA, OA, (CSMG*)	46 (USA)	Hypothesis testing
Rohekar 2009 (54)	HAQ (assumed DI)	59.91±11.83	RA (University-based clinic)	122 (Canada)	Reliability
Seror 2010 (55)	HAQ-DI vs. Individualized scales[7]	58±11.9	RA (Study cohort)	370 (France)	Internal Consistency, Hypothesis Testing, Criterion Validity, Responsiveness
Sheehan 2001 **(56)**	HAQ vs ADL scale (from NHANES)	NR	General population, RA	Population 1: 4430 (NHANES); Population 2: 605 RA (USA); Pop 3: 74 RA (Great Britain)	Structural validity
Singer 1982 (57)	HAQ (original) Mathies Tool (reference not English) Singer et al. Tool (published in book, not a journal)	NR	RA (Multiple Hospitals)	46 (Austria)	Reliability, Hypothesis Testing
Sousa 2008 (58)	HAQ-DI	NR	HIV, RA (Study cohort)	901 (USA)	Structural Validity
Sullivan 1987 (59)	HAQ (unspecified)	NR	RA, OA, Gout, Other (Single practice)	(Scotland)	Hypothesis Testing
Taylor 2007 (60)	HAQ-DI	RA: 60.7±14.4	RA, PsA (LCD)	581 (New Zealand)	Internal Consistency, Structural Validity, Hypothesis Testing
Tennant 1996 (61)	HAQ (unspecified, presumed DI)	66.6±SD 7.9	RA, OA (LCD)	506 (UK)	Structural Validity
Verhoeven 2000 (62)	HAQ (unspecified, presumed DI) Functional Status VAS	(23-70)	RA (RCT)	155 (Netherlands)	Responsiveness

Author (year)	Performance Measure(s)	Mean Age Years ± SD (range)	Population (setting)	N (Country)	Measurement Property(ies) Evaluated
Ward 2015 (63)	HAQ (assumed DI)	51.0±13.7	RA (University clinics)	250 (USA)	Responsiveness
Ward 1994 (64)	HAQ -DI	46 (SD NR; 28-73)	RA (University Rheumatology Clinics, CSMG, CBR*)	24 (USA)	Hypothesis Testing, Responsiveness
Wells 2008 (65)	HAQ (assumed DI)	Abatacept 53.5 (12.4) Placebo 52.7 (11.3)	RA (RCT)	Total 391: Abatacept n=258; Placebo n=133 (Canada?)	Responsiveness
Wolfe 2005 (66)	HAQ-DI	51.6±SD 9.4	RA (LCD)	8931 (USA)	Interpretability
AHAQ					
Tomlin 1996 (25)	AHAQ vs HAQ-DI	62.1±12.6	RA (Hospital Rheumatology Clinic)	107 (USA)	Internal Consistency, Hypothesis Testing, Responsiveness
MHAQ					
Callahan 1992 (67)	MHAQ	55.2 (SD NR)	RA (LCD)	982 (USA)	Hypothesis testing (Convergent validity) Responsiveness
Hagen 1999 (68)	MHAQ	NR	RA (LCD)	595 (Norway)	Responsiveness
Kvamme 2010 (69)	MHAQ	For RA patients that completed PASS & MCII: 54.6 ±13.4	RA, PsA, AS (LCD)	4036* (Norway)	Interpretability
Martin 2007 (70)	MHAQ IRT-based scale combining MHAQ and SF-36 PF-10 scale	55 (17-83)	RA (RCT)	339 (US, Non-US)	Internal Consistency, Measurement Error, Hypothesis Testing, Responsiveness
Nagasawa 2010 (71)	HAQ-DI MHAQ	52.8±12.4 (24-71)	RA (Study Cohort)	87 (Japan)	Hypothesis Testing, Criterion Validity
Pincus 1983 (27)	MHAQ vs HAQ-DI	NR	RA, other Rheumatic Diseases (Multiple practices)	263* (USA)	Reliability, Criterion Validity

Author (year)	**Performance Measure(s)**	**Mean Age Years ± SD (range)**	**Population (setting)**	**N (Country)**	**Measurement Property(ies) Evaluated**
Russel 2003 (72)	MHAQ	NR	RA (University-based clinic)	Group 1: 24 Group 2: 60 (Canada)	Reliability, Responsiveness
Stucki 1995 (73)	MHAQ vs HAQ	62 (SD NR)	RA (University-based practice)	56 (Switzerland)	Hypothesis Testing, Responsiveness
Tugwell 2000 (74)	HAQ-DI vs MHAQ	Leflunomide 54.1 ± 12.0; placebo 54.6 ±10.7; MTX 53.3 ± 11.8	RA (RCT)	480 total: Leflunomide n=182; methotrexate n=180; placebo n=118 (USA?)	Responsiveness
Uhlig 2006 (75)	HAQ (assumed DI) vs MHAQ	55.8±12.9	RA (Study cohort)	179 (USA)	Hypothesis Testing, Criterion Validity
Wolfe 2001 (76)	HAQ, MHAQ, RA-HAQ, and DHAQ and HAQ20	58.01 ±12.57	RA (Multiple practices)	2491 (USA)	Structural Validity
Ziebland 1992 (77)	HAQ (assumed DI) vs MHAQ	56 ± 12.2	RA (Study cohort)	100 (UK)	Responsiveness
MDHAQ					
Pincus 1999 (28)	MDHAQ	For 162 RA patients: 54.7 (SD NR)	RA, Fibromyalgia, OA, SLE, Vasculitis, PsA, Scleroderma, Other (Single practice)	688 (USA)	Reliability, Hypothesis Testing
Pincus 2005 (29)	10-ADL MDHAQ to other versions e.g. 14-ADL MDHAQ, 20-ADL and 8-ADL MDHAQs	53±12	RA, Fibromyalgia, Other (Single practice)	144 (USA)	Internal Consistency, Structural validity
HAQ-II					

Author (year)	**Performance Measure(s)**	**Mean Age Years ± SD (range)**	**Population (setting)**	**N (Country)**	**Measurement Property(ies) Evaluated**
Wolfe 2004 (30)	HAQ II vs MHAQ, MDHAQ, HAQ-DI	NR	RA, OA, Fibromyalgia (LCD, Study Cohorts)	Development: 19957 Validation Studies: 14038 RAES Cohort Correlation: 693 (USA, Canada) Hypothesis Testing: 837 (USA)	Internal Consistency, Structural Validity, Hypothesis Testing, Criterion validity, Responsiveness

Author (year)	Performance Measure(s)	Mean Age Years ± SD (range)	Population (setting)	N (Country)	Measurement Property(ies) Evaluated
Performance Measure Evaluated: PROMIS, Multiple					
Bartlett 2015 (33)	PROMIS PF CAT MHAQ	55.5 (13.3)	RA (Academic Rheumatology Practice)	177 (USA)	Internal Consistency, Reliability, Hypothesis Testing, Criterion Validity
Fries 2011 (78)	PROMIS vs HAQ-DI (same study as below by Fries 2011 in J Rheum, reporting different metrics)	65 (SD NR)	RA other diseases (setting not clear)	Responsiveness Testing: 451 (USA) Mode of Administration Testing: 721 (USA)	Responsiveness, Reliability
Fries 2011 (79)	PROMIS PF-10[††], PROMIS PF- 20[††], Legacy HAQ, SF-36 PF-10, Item-Improved HAQ, Item-Improved PF-10	65 (SD NR)	RA (Not clear group as above)	451 (USA)	Responsiveness
Hays 2015 (80)	PROMIS PF 20[†] HAQ-DI SF-36	NR (may be elsewhere)	RA (Study Cohort)	451 (USA)	Responsiveness, Hypothesis testing
Oude Voshaar 2014 (34)	PROMIS CATs with 5, 10, 15 items vs HAQ-DI SF-36	NR	RA (LCD / Simulated Study)	690 (Netherlands; may have used a systematic review for some patients)	Reports on sensitivity to change and measurement precision, methods hard to evaluate using COSMIN, put some information in the footnotes of tables instead, Hypothesis testing
Oude Voshaar 2015 (81)	PROMIS item bank and 20-item SF[††], compared to HAQ-DI and SF-36 PF-10	NR (may be elsewhere)	RA (LCD)	690 (Netherlands)	Content Validity, Hypothesis Testing
Schalet 2016 (82)	PROMIS PF-10 SF & PF-20[§]	NR	RA, Back pain, Cancer, MDD, COPD, CHF (Setting not clear)	1415 (USA)	Responsiveness
Wahl 2017 (31)	PROMIS PF-10a	59 (14)	RA, Other (Rheumatology Clinic)	416 (USA)	Hypothesis testing Responsiveness

SF: Short form

[†] Probably PF 20a as same questions presented but not specified in manuscript

††In these studies further specification of the type of SF not given (e.g. PF-10a, 20a not specified)

§In this study by Schalet et al. in methods for RA only short forms were given (both 10 and 20 item versions), in results they appear to be reported together. Further specification of the type of short form (e.g. PF-10a) not provided.

* Unless otherwise noted, HAQ or HAQ-DI both refer to the Disability Index of the Health Assessment Questionnaire

** Acronyms: CBR: Community-Based Recruitment CSMG: Community Self-Management Group; LCD: Longitudinal Clinical Database or LCD; NR: Not Reported; RA: Rheumatoid Arthritis; RCT: Randomized controlled trial

*** For all studies included in Table 2, the value of 'n' reported refers to total number of patients included in the study but due to subgrouping within that 'n' and choice of study design, not all patients comprising 'n' are represented by a complete dataset or would have completed all of the functional status assessments being evaluated.

Supplementary Table 3. Psychometric properties (Internal consistency, reliability and measurement error) & COSMIN ratings of included studies

Author	**Internal consistency**			**Reliability**					**Measurement error**		
	Results	**Study n**	**COSMIN score**	**Result**	**Design**	**Time interval**	**Study n**	**COSMIN score**	**Result**	**Study n**	**COSMIN score**
HAQ-DI (and original)											
Fries 1980 HAQ (original) (26)	Cronbach's alpha NR (other statistics reported, hard to compare)	20	Poor	Spearman's rho: 0.85; Weighted Kappa 0.52 (moderate)	Inter-rater	0-12 days	20	Poor	N/A	N/A	N/A
Goeppinger 1988 HAQ (assumed DI) (42)	Pearson's *r*=0.46 to 0.63[2] and Cronbach's alpha 0.77 to 0.87	15	Poor	Pearson's *r*: 0.95 (RA only)	Test- retest	7 days	30 (15RA)	Poor	N/A	N/A	N/A
Häkkinen 2005 HAQ-DI (44)	Chronbach's alpha for overall score 0.91 95% CI 0.89 (one side lower limit reported); for subscales ranged from 0.71 to 0.84	304	Poor	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Lassere 2001 HAQ-DI (48)	N/A	N/A	N/A	Study B: ICC 0.91; Study C: ICC 0.95 (no 95%CI reported)	Test-retest	Study B: day 8; Study C: day 2	Study B (24); Study C (26)	Poor	SDD 95% LoA Study B (–0.69 to 0.59); Study C (–0.29 to 0.48)	Study B (24); Study C (26)	Poor
Linde 2008 HAQ- DI (49)	Cronbach's alpha 0.95	200	Poor	ICC 0.97 (95% CI 0.96-0.98)	Test-retest	14 days	150	Poor	95% LoA, mean ± 1.96*SD: 0±0.38	87	Poor
Marra 2005 HAQ (assumed DI) (50)	N/A	N/A	N/A	ICC 0.97 (95% CI 0.93-0.98)	Test-retest	5 weeks	50	Good	N/A	N/A	N/A
Rohekar 2009 HAQ (assumed DI) (54)	NR	NR	NR	ICC 0.897 (95% CI 0.855, 0.927)	Test-retest	1-2 days	122	Poor	N/A	N/A	N/A
Seror 2010 HAQ-DI vs. Individualized scales[7] (55)	Chronbach's alpha (95% CI's) HAQ-DI 0.87 (0.85 to 0.89); Importance questionnaires: Individualized HAQ multiplicative 0.88 (0.85 to 0.90); Individualized HAQ additive 0.88 (0.86 to 0.90); Preference questionnaire: not estimable	370	Poor	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Author	**Internal consistency**			**Reliability**					**Measurement error**		
	Results	**Study n**	**COSMIN score**	**Results**	**Design**	**Time interval**	**Study n**	**COSMIN score**	**Results**	**Study n**	**COSMIN score**
Singer 1982 HAQ (original) (57)	N/A	N/A	N/A	Correlation between HAQ by patient and occupational therapist *r* 0.859 p<0.001	Inter-rater	Same day	30	Poor	N/A	N/A	N/A
Sheehan 2001 HAQ DI vs ADL score from NHANES (56)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	Rasch analysis: HAQ has greater precision and less measurement error in assessing patients with arthritis than the ADL has in the general population.	Population 1: 4430 (NHANES); Population 2: 605 RA (USA); Pop 3: 74 RA (Great Britain)	N/A (difficult to assess using COSMIN criteria)
Taylor 2007 HAQ-DI (60)	Cronbach's alpha NR Table 2 has Fit of data to the Rasch model HAQ for each subscale in RA. InFitMNSQ range from 0.78-1.38. DIF p value significant for Rising, grip and activity	142 RA	Fair	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
AHAQ [7]											
Tomlin 1996 AHAQ vs HAQ-DI(25)	Chronbach's alpha NR Spearman's rho intra-correlation coefficients and Fisher's transformation of the coefficients (Zrho) of category scores with the disability index for HAQ range (rho=: 0.608-0.785) and AHAQ (rho= 0.660-0.806).	107	Poor	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
MHAQ											

Author	Internal consistency			Reliability					Measurement error		
	Results	Study n	COSMIN score	Results	Design	Time interval	Study n	COSMIN score	Results	Study n	COSMIN score
Martin 2007 MHAQ and new IRT based scale combining MHAQ and SF-36 PF-10 scale (70)	Chronbach's alpha NR (used IRT-based methods) Correlation between factors: Solution 1 (2 factor based on original scales MHAQ and SF-36): 0.79 and Solution 2 (single factor): 0.74	339	Excellent	N/A	N/A	N/A	N/A	N/A	Reports on 95% CI (not on SEMs or SDCs or LoA):" For the entire score range, the 95% CI around individual scores was smaller for the combined (total) IRT based scale than for other measures"	339	Fair
Pincus 1983 MHAQ vs HAQ (27)	HAQ: Chronbach's alpha (range 0.710-0.890) MHAQ: Correlations of mean scores between difficulty and satisfaction, change and help *r*=0.694, 0.380, 0.229** (* p<0.001, **p<0.002)[3]	HAQ: 97 MHAQ: 190	Fair	HAQ: Pearson's *r* of 0.78 (P < 0.001) MHAQ: Pearson's *r* of 0.91 1 (P < 0.001) (study also reported on HAQ shown above)	Test-retest	1 month	28	Poor	N/A	N/A	N/A
Russell 2003 (72) MHAQ	N/A	N/A	N/A	ICC 0.89 (95% CI NR)	Test-retest	2 visits 3 weeks apart	24	Poor	SEM 0.14 SD Diff 0.20	24	Poor
MDHAQ											
Pincus 1999 MDHAQ(28)	N/A	N/A	N/A	Kappa scores for all items ranged from 0.65 to 0.81 (all P <0.001)[4]	Test-retest	Pre & post visit (same day)	112	Poor	N/A	N/A	N/A
Pincus 2005 Compares 10-ADL MDHAQ to other versions e.g. 14-ADL MDHAQ, 20-ADL and 8-ADL MDHAQs (29)	"Internal Consistency" alpha (95% CI, lower limit): HAQ 0.90 (0.88); 8-ADL MHAQ 0.90 (0.88); 14-ADL MDHAQ 0.92 (0.90); 10-ADL MDHAQ 0.89 (0.87).	144	Fair	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
HAQ II											
Wolfe et al. 2004 (30) HAQ II vs MHAQ, MDHAQ, HAQ-DI	Cronbach's alpha: HAQ 0.83, MHAQ 0.81 and MD-HAQ 0.85 and HAQ-II 0.88	19927	Fair	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

PROMIS											
Author	Internal consistency			Reliability					Measurement error		
	Results	Study n	COSMIN score	Result	Design	Time interval	Study n	COSMIN score	Results	Study n	COSMIN score
Bartlett 2015 PROMIS PF CAT vs MHAQ (33)	Crohnbach's alpha 0.985 (95% CI: .981, .988)	177	Fair	Spearman’s rho: 0.975	Test-retest	mean 2.2 days	34	Fair	N/A	N/A	N/A
Fries 2011 PROMIS PF-10††, PROMIS PF- 20††, Legacy HAQ, SF-36 PF-10, Item-Improved HAQ, Item-Improved PF-10 (78)	N/A	N/A	N/A	Chronbach’s alpha=0.93, r=0.92[5]. Generalized linear model demonstrated no relevant effect for different modes of administration.	Test-retest comparing modes of administration (paper vs internet)	unclear	721 (n for RA ?)	Fair	N/A	N/A	N/A
Oude Voshaar 2014 PROMIS CAT PF-5,10,15 (34)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A[6]	N/A	N/A
Oude Voshaar 2015 PROMIS item bank and 20-item SF††, compared to HAQ-DI and SF-36 PF-10(81)	N/A	N/A	N/A	Global reliability of HAQ-DI and PF-10 0.89 and 0.90. Precision of full PROMIS PF item set higher than HAQ-DI or PF-10 at all levels (data not shown in study)	IRT methods to report global reliability	N/A given methods	690	N/A	N/A	N/A	N/A

NB: all abbreviations at end of Table 4

[1] Cronhbach’s alpha (usual method of reporting internal consistency) not calculated, instead authors report Spearman’s correlations between each set of questions and with overall disability index. Range with disability index reported here.

[2] Pearson's r calculated for 4 categories with only 2 items and coefficient alpha computed for the remaining 4 categories with >2 items.

[4] Study examined test-retest reliability for each of the 8 items of the MHAQ and 10 new items but data not shown to tease apart MHAQ vs MDHAQ questions.

[5] Fries et al 2011 report these findings for 721 participants (including RA, depression and/or chronic obstructive pulmonary disease) examining impact of mode of administration between paper and pencil, internet-based modes of administration of forms measuring daily life functions, back-neck function and 2 items lower and 2 items upper extremity function. These are framed as “preliminary results” in the manuscript

[6] Oude Voshaar et al 2014 (34) report on measurement precision but can’t be rated based on COSMIN. Concluded that higher precision based on RMSE (root mean square errors) observed for PROMIS CAT (5, 10 & 15 CAT) compared with HAQ DI and SF-36 PF-10

[7] Seror et al. (55) examined individualized scales. At baseline and final visits, patients had to rate the importance they attached to each activity addressed by the 20 HAQ-DI items, and to select the 5 activities they considered the most important. Different individualized scales were evaluated: scales preserving all domains, in which the score for each item is multiplied by or added to its importance; and scales involving for each patient only the 5 most important items.

[8] Tomlin et al. (25) the Alternative HAQ (AHAQ) uses the arithmetic mean of the category scores instead of the worst item scores in that category, followed by the usual arithmetic mean of the category scores.

† Probably PF 20a as same questions presented but not specified in manuscript

†† In these studies further specification of the type of SF not given (e.g. PF-10a, 20a not specified)

§ In this study by Schalet et al. in methods for RA only short forms were given (both 10 and 20 item versions), in results they appear to be reported together. Further specification of the type of short form (e.g. PF-10a) not provided.

Supplementary Table 4. Psychometric properties (Validity, responsiveness and Interpretability) & COSMIN ratings of included studies

Functional status measure	**Validity (hypothesis testing)**			**Responsiveness**			**Interpretability**	
	Design	**Result**	**COSMIN score**	**Treatment**	**Result**	**COSMIN score**	**f/c (%)**	**MIC or MID**
HAQ-DI (and original)								
Bombardier 1991 HAQ-DI (35)	N/A	N/A	N/A	Auranofin	SES:0.25 RE: 1.11/1.09 Comparator TJC	Fair	NR	NR
Brown 1984 HAQ DI +Pain (36)	Hypothesis testing (Convergent); Structural validity	Correlation with AIMS Physical 0.91**; AIMS pain 0.39**; AIMS psychological 0.23; HAQ Pain* (*p<0.05, p<0.01) Factor 1 "Physical" explains 55% of the variance; Factor 2 "Pain" explains 15% of the variance	HT: Fair SV: Fair	N/A	N/A	N/A	NR	NR
Buchbinder 1995 HAQ-DI (37)	N/A	N/A	N/A	Cyclosporin	RE:0.58^{2} r=0.41^{3} (overall), 0.54(treatment), 0.21 (placebo) Comparator TJC	Fair	NR	NR
Cole 2005 HAQ-DI (38)	Structural validity	Single factor (5.47^{4}, 68.4% of variance explained)	Excellent	N/A	N/A	N/A	NR	NR
Fitzpatrick 1989 HAQ-DI (39)	Hypothesis testing (Convergent)	Correlations with MM items (t1,t2): Stiffness (0.41, 0.40); Pain (0.61, 0.64); Grip strength (-0.73, -0.68); Ritchie Index (0.6, 0.589); ESR (0.38, 0.33) all p=0.001; Hgb (-0.23 p<0.01, -0.21, p<0.05)	Fair	Usual care	Change in HAQ score^{5} (any, >0.25): sensitivity of improvement (0.65, 0.30); sensitivity of worsening (0.60, 0.47); specificity of improvement (0.61, 0.84); specificity of worsening (0.73, 0.82). Comparator ARA functional status	Fair	NR	NR
Fitzpatrick 1993 HAQ-DI (40)	Hypothesis testing (Convergent)	Correlations with parts of HAQ. Mobility HAQ: ESR 0.24**, Articular index 0.27**, Grip strength -0.41***, Beck Depression Inventory 0.21*; ADL HAQ: ESR 0.25**, Articular index 0.31**, Grip strength -0.45***, Beck Depression Inventory 0.15 NS; Household HAQ ESR 0.43***, Articular index 0.24**, Grip -0.44***, Beck depression index 0.2**	Fair	Usual care	Correlations at^{6} (t1-t2, t2-t3): ESR (0.26, 0.28); Articular index (0.12, 0.21); Grip strength (-0.23, -0.35). ES: Better (0.48, 0.2); Worse (0.27, 0.11) Comparator patient global	Fair	NR	NR

		*p<0.05, **p<0.01, ***p<0.001						
Functional status measure	**Validity (hypothesis testing)**			**Responsiveness**			**Interpretability**	
	Design	**Result**	**COSMIN score**	**Treatment**	**Result**	**COSMIN score**	**f/c (%)**	**MIC or MID**
Fitzpatrick 1992 HAQ-DI (41)	NR	NR	N/A	Usual care	ES t1-t2 for subscales of the HAQ for patients with improved health by self-assessment Mobility (0.38); ADL (0.28); Household (0.74)	Fair	NR	NR
Fries 1980 HAQ (original) (26)	Structural validity; Hypothesis testing (Convergent)	1St principal[7] component weight 0.58 to 0.93 (65% of interperson variation accounted for in this one dimension) 2nd principal component weight -0.50 to 0.52 (10% of interprerson variation); Spearman's rho= 0.88 Comparator observed function	SV: Poor HT: Poor	N/A	N/A	N/A	NR	NR
Goeppinger 1988 HAQ (assumed DI) (42)	Hypothesis testing (convergent); "concurrent validity"; Content validity	Pearson's *r*: 0.88 Comparator AIMS total health score; Canonical correlation for discriminant function 0.57, 0.65[8] Content validity:" Content analysis suggested the HAQ represented the scope of nursing practice better than the AIMS"	HT: Poor CT: Fair	N/A	N/A	N/A	NR	NR
Greenwood 2001 HAQ-DI (43)	N/A	N/A	N/A	Usual care	Kappa for 3 time periods were 0.72 (n=38), 0.69 (n=37), and 0.76 (n=39) Comparator change in general health	Fair	6%/NR	0.48
Hawley 1992 HAQ-DI (45)	N/A	N/A	N/A	Methotrexate	Group 1 (MTX sub-study at 6 months) ES 0.51. For early disease ES was 0.72 vs late disease 0.37 (using 2yr cutoff). Group 2 (10-year FU study): ES size at 2 yrs=-0.01; ES at 5 yrs=-1.64 and at 10 yrs =-2.39	Poor	NR	NR
Functional status measure	**Validity (hypothesis testing)**			**Responsiveness**			**Interpretability**	
	Design	**Result**	**COSMIN score**	**Treatment**	**Result**	**Cosmin score**	**f/c(%)**	**MIC or MID**
Hendrikx 2015 HAQ-DI (46)	N/A	N/A	N/A	N/A	N/A	N/A	NR	From ROC analysis: MIC HAQ-DI improvement -0.06 (false

								positive change 24%; false neg chance 72%; bootstrap MIC 0.01 95% CI -0.30 :0.88); MIC HAQ-DI deterioration 0.08 (false positive change 30%; false neg chance 54%; bootstrap MIC 0.08 95% CI -0.30 :0.27)
Kosinski 2000 HAQ-DI (47)	N/A	N/A	N/A	N/A	N/A	N/A	NR	Summary of mean HAQ-DI changes at one level of improvement across 5 RA severity measures: PtG (-0.24); PhG (-0.17); pain VAS (-0.22); SJC (-0.19); TJC (-0.13) Ave change (-0.19). Summary of categorical (% better) changes in HAQ-DI scores at one level of improvement across 5 RA severity measures: PtG (31%); PhG (23%); Pain (33%); SJC (26%); TJC (23%); Average change (27%)
Functional status measure	**Validity (hypothesis testing)**			**Responsiveness**			**Interpretability**	
	Design	**Result**	**COSMIN score**	**Treatment**	**Result**	**COSMIN score**	**f/c(%)**	**MIC or MID**
Kvamme 2010 MHAQ (69)	N/A	N/A	N/A	N/A	N/A	N/A	NR	[9]PASS:75% sens 0.63; 80% spec 0.33; Area under

								ROC 0.75, 95% CI 0.73-0.77. MCII: 75% sens cutpoint =0; 80% spec cutpoint -0.25; Area under ROC 0.71, 95% CI 0.69-0.73.
Functional status measure	**Validity (hypothesis testing)**			**Responsiveness**			**Interpretability**	
	Design	**Result**	**COSMIN score**	**Treatment**	**Result**	**COSMIN score**	**f/c(%)**	**MIC or MID**
Linde 2008 HAQ-DI (49)	Hypothesis testing (known-groups, convergent, discriminant)	Known groups: significant differences in HAQ scores between low and moderate DAS28 scores and between low vs moderate and moderate vs high VAS arthritis activity. ES: bone erosions 0.22, disability pension 0.66 (significant). Convergent and discriminant: multitrait-multimethod correlation matrix. For HAQ and SF-36 PF -0.769; HAQ and SF-36 physical role limitations -0.574; HAQ and bodily pain -0.714; VAS pain 0.714; SF36 vitality -0.600; VAS fatigue 0.671; VAS global RA 0.714; RA QoL 0.814; EQ-5D-0.791; 15D -0.741; GH-0.508.	Fair	Usual care	SRM (n=96) improvement (n=26) HAQ-0.10; No change (n=47) -0.26; Deterioration (n=23) 0.13.	Fair	Pop 1 0/25; Pop 2 1/10	NR
Marra 2005 HAQ-DI (50)	N/A	N/A	N/A	Usual care	Transition defined categories: HAQ ES with 95% CI: worse 0.22 (0.04 to 0.38) ; same -0.09 (-0.28 to 0.02); better -0.24 (-0.38 to -0.11); SRM 95% CI worse 0.33 (0.06 to 0.65) ; same -0.20 (-0.56 to -0.10); better -0.39 (-0.69 to -0.30); RE worse 1.21; better 0.71. Patient VAS: HAQ effect sizes with 95% CI: worse 0.34 (0.11 to 0.44) ; same -0.08 (-0.06 to -0.25); better -0.35 (-0.32 to -0.76); SRM 95% CI worse 0.50 (0.28 to 0.88) ; same -0.17 (-0.12 to -0.46); better -0.50 (-	Fair	NR	NR

					0.48 to -0.92); RE worse 0.97; better 0.72			
Functional status measure	**Validity (hypothesis testing)**			**Responsiveness**			**Interpretability**	
	Design	**Result**	**COSMIN score**	**Treatment**	**Result**	**COSMIN score**	**f/c (%)**	**MIC or MID**
Marra 2005 HAQ-DI (51)	Hypothesis testing (Convergent)	Spearman's rho 0.46 with RA severity, 0.45 RA control (both p<0.0001). ES[12] : AE of drug therapy 0.19, hospitalized in last year 0.44, other chronic diseases 0.29, days off work/school due to RA in last year yes/no 0.60, use of allied health/home services for RA in past year (y/n) 0.74, rent or purchase of equipment for RA in past year ES 0.61. All ES significant with the exception of AE to RA therapy. Correlation (Spearman's rho) for overall scores with RA severity: HUI2 global utility -0.66; HUI3 global utility -0.76, SF-6D global utility -0.73, EQ-5D global utility -0.61, RAQoL score 0.76. RA duration in years 0.28, SJC 0.48, TJC 0.46, PtG VAS -0.53, Pain VAS 0.54. All of these latter correlations starting with HUI2 were significant.	Fair	NR	NR	NR	NR	MID 0.15
Pope 2009 HAQ-DI (52)	N/A	N/A	N/A	Usual care (presumably)	Spearman's rho [patient assessment of global change, change in the HAQ-DI] 0.36 (p < 0.001). ES for somewhat improved 0.12 and somewhat worsened 0.20. ES for somewhat better/much better (0.27) and for somewhat worse/much worse (0.27)	Poor	NR	MID estimates for HAQ-DI change mean (SD): much better (n=11)–0.57 (0.67) 95% CI –1.01 to –0.12; somewhat better (n=35)–0.09 (0.42) 95% CI –0.23 to 0.05); same (n=120) 0.03 (0.32) 95% CI–0.030 to 0.09; somewhat worse

								(n=50) 0.15 (0.33) 95% CI 0.060 to 0.25; 0.50 (0.13) 95% CI 0.40 to 0.60.
Functional status measure	**Validity (hypothesis testing)**			**Responsiveness**			**Interpretability**	
	Design	**Result**	**COSMIN score**	**Treatment**	**Result**	**COSMIN score**	**f/c(%)**	**MIC or MID**
Redelmeier 1993 HAQ (assumed DI) (53)	Hypothesis testing (Convergent)	Differences in HAQ scores and subjective comparison ratings were significantly correlated (Spearman Rank Correlation, 0.41; 95% confidence interval, 0.31 to 0.50).	Poor	N/A	N/A	N/A	N/A	Overall estimate of the threshold of symptomatic clinical importance (0.19 HAQ units; 95% confidence interval, 0.10 to 0.28 HAQ units)

Seror 2010 Different individualized scales vs HAQ-DI (55)	Criterion validity; Hypothesis testing (convergent)	All individualized scale scores highly correlated to HAQ-DI (Spearman's *r* ≥0.75) Lower correlations were observed with measures of disease activity: TJC, SJC (Spearman's *r* 0.21 to 0.39) and DAS28 (0.38 to 0.47). The lowest correlations observed with biological features of disease activity, such as ESR and CRP level (0.10 to 0.18).	Excellent; Fair	Leflunomide	SRMs HAQ-DI 0.74 (95% CI 0.64 to 0.86);
					Importance questionnaire: individualized HAQ **multiplicative** 0.69 (95% CI 0.58 to 0.79);
					Individualized HAQ **additive** 0.68 (95% CI 0.58 to 0.80).
					Preference questionnaire 5-item HAQ 0.65 (95% CI 0.55 to 0.77);
					Weighted five-item HAQ 0.64 (95% CI 0.54 to 0.76)
						Poor	NR	NR

Singer 1982 HAQ (original) (57)	Hypothesis testing (Convergent)	Correlation coefficient 0.754 p<0.001 (between Ritchie articular index and disability stated by occupational therapist based on HAQ	Poor	NA	NA	NA	9% had minimal score/0 had highest	NR
Sousa 2008 HAQ-DI (58)	Structural Validity	2nd principal component weight -0.50 to 0.52 (10% of interperson variation)	Fair	NA	NA	NA	NR	NR
Sullivan 1987 HAQ (unspecified) (59)	Hypothesis testing (Convergent)	Correlation between HAQ and observation (r=0 83)	Poor	N/A	N/A	N/A	NR	NR

Functional status measure	Validity (hypothesis testing)			Responsiveness			Interpretability	
	Design	**Result**	**COSMIN score**	**Treatment**	**Result**	**COSMIN score**	**f/c(%)**	**MIC or MID**
Taylor 2007 HAQ-DI (60)	Structural Validity; Hypothesis testing (Convergent)	Rasch model adequately fit the observed HAQ DI data but there was evidence of misfitting items and DIF. The item separation was 2.06 logits in the PsA group and 3.80 logits in the RA group, indicating better span of the disability scale in RA. The HAQ DI was nonlinear at the extremes of the disability scale for both groups, especially at HAQ DI scores <0.875 for the PsA group and <0.375 for the RA group. Authors report results of regression [the person Rasch estimates (logit scale) of HAQ DI were plotted against those of SF36] where slope for regression line is 1.14 [95% CI 0.96, 1.31] Authors conclude that they are measuring the same concept.	SV: Good HT: Fair	N/A	N/A	N/A	Floor effects were significant in the PsA group, in which 30.4% had scores indicating no disability, but only 6.9% of the RA group had scores indicating no disability	N/A
Tennant 1996 HAQ fitted to Rasch model (61) HAQ	Structural	The fit of the HAQ data: "The mean square information-weighted fit statistic INFIT is between -0.7 and +1.3, a range considered to represent an adequate fit of the data to the model. The hierarchical nature of the scale, expressed by item separation, is somewhat restricted at 2.82. This meets basic requirements that a scale should identify at least 2 strata, but suggests that in the HAQ, the underlying scale construct of disability is limited in it range."	Fair	N/A	N/A	N/A	NR	NR

Functional status measure	Validity (hypothesis testing)			Responsiveness			Interpretability	
	Design	Result	COSMIN score	Treatment	Result	COSMIN score	f/c (%)	MIC or MID
Verhoeven 2000 HAQ (unspecified, presumed DI) (62)	N/A	N/A	N/A	COBRA Clinical Trial[10]	**AT 16 weeks Combined treatment** (n=75) mean change -1.1, Standard Error of change 0.1, SRM 1.5, ES 1.5; **SSZ (n=79)** mean change -0.4, Standard error of change 0.1, SRM 0.8, ES 0.6; tvalue 6.2. **At 28 weeks: Combined treatment** (n=75) mean change -1.1, Standard Error of change 0.1, SRM 1.4, ES 1.5; **SSZ (n=79)** mean change -0.6, Standard error of change 0.1, SRM 0.9, ES 0.8; tvalue 4.5.	Poor	NR	NR
Ward 1994 HAQ-DI (64)	Hypothesis testing (convergent)	Partial correlations[21] between the physician determined measures and patient determined measures: HAQ-DI the following SJC=0.56, weighted SJC=0.50, TJC=0.55, weighted TJC=0.61, Physician Global=0.70. All p<0.001; Partial correlations among the patient derived measures, functional measures and lab measures: HAQ-DI and patient global=0.71, pain=0.64, AM stiffness 0.45 all p<0.0001; Disability and ESR 0.30, Hgb -0.12 and platelet 0.15 all p<0.0001. Partial correlations between each of the 2 top candidate measures of each group and the most accurate individual measures by multivariate analysis. HAQ-DI and Physician global =0.87, weighted TJC 0.79, patient global assessment 0.76 and pain 0.74 and ESR 0.44 all p<0.0001. Of the functional measures, the DI was more highly	Poor	Usual care	HAQ-DI SRMs: Physician global 0.6, patient global 0.64, ESR 0.30, Average of the above 0.51. SJC 0.32, weighted SJC 0.33, TJC 0.10, Weighted TJC 0.16, Physician global 0.84, patient global 0.74, pain 0.48, AM stiffness duration 0.39, grip strength 0.14, walk time 0.53, ESR 0.12, Hemoglobin 0.17, Platelet count 0.37.	Poor	13%/0% (baseline)	NR

		correlated with each of the other measures than was grip strength when both were evaluated simultaneously.						
Functional status measure	**Validity (hypothesis testing)**			**Responsiveness**			**Interpretability**	
	Design	**Result**	**COSMIN score**	**Treatment**	**Result**	**COSMIN score**	**f/c (%)**	**MIC or MID**
Ward 2015 HAQ (presumed DI) (63)	N/A	N/A	N/A	Standard care [calls this 'sensitivity to change']	Mean change HAQ: −0.4±0.6; SRM −0.65 (95% CI −0.58 to −0.72); Mean change by improvement category improved -0.63; same -0.08; worsened 0.06 (pANOVA<0.0001).	Fair	NR	NR
Wells 2008 HAQ (presumed DI) (65)	N/A	N/A	N/A	RCT abatacept vs placebo	Relative improvement SRM (95% CI) 0.63 (0.42 to 0.85) RE 1.22	Fair	NR	NR
Wolfe 2005 HAQ-DI (66)	N/A	N/A	N/A	N/A	N/A	N/A	NR	RID rates adjusted for age and sex and refer to patients <65 years. Mean difference (95% CI): Work disabled 0.74 (0.71, 0.76); Social security disability 0.76 (0.72, 0.79); TJR 0.54 (0.49, 0.59); Poverty 0.57 (0.52, 0.61); satisfied with health 0.75 (0.71, 0.79); depend on others for help 0.87 (0.83, 0.91). "As expected, RID are considerably greater than MCID" "Using a health utility score as a common metric, improvements corresponding

								with MCID result in small differences of 0.06, whereas RID differences based on satisfaction with health, independence, and no work disability are as great as 0.27, 0.26, and 0.23, respectively."
MHAQ								
Functional status measure	**Validity (hypothesis testing)**			**Responsiveness**			**Interpretability**	
	Design	**Result**	**COSMIN score**	**Treatment**	**Result**	**COSMIN score**	**f/c(%)**	**MIC or MID**
Callahan 1992 MHAQ (difficulty subscale) (67)	Hypo. Test (convergent)	Positive correlations with dissatisfaction 0.85 and pain 0.79. (both p<0.001)	Fair	Usual care	ES (1year) -0.05 ES (5year) -0.28[3] r: 1-year dissatisfaction 0.68, pain 0.52. 5- years dissatisfaction 0.62, pain 0.58. All p<0.001	Fair	NR	NR
Hagen 1999 MHAQ (68)	N/A	N/A	N/A	Usual care	ANOVA for MHAQ over 5 classes of change scores F=26.6, p<0.001, $R^2$0.15; SRMs (95%CI) for MHAQ: improvement ≥2 (0.8, 95%CI 0.4, 1.1); 1 (0.3, 95%CI, 0.1, 0.5); 0 (-0.1, 95%CI 0.0, -0.3); deterioration -1 (-0.4, 95%CI -0.2, -0.6); ≤2(-1.1, 95%CI -0.6, -1.6). Comparator Patient global disease (-2 to 2)	Fair	NR	NR
Martin 2007 MHAQ and new IRT based scale combining MHAQ and SF-36 PF-10 scale (70)	Structural validity; Hypothesis testing (discriminant)	Solution 1: Correlation between PF10 & MHAQ 0.79 (high) but improved model fit. Discriminant validity reports RV and 6 & 12 months: MHAQ 0.71, 0.70; Total IRT scale 1.0, 1.0[10]	SV: Excellent; HT: Fair	Abatacept vs placebo	MHAQ ES at 6, 12 months: placebo 0.34, 0.25; 10mg/kg 0.72, 0.72; Total scale ES at 6, 12 months: 0.43,0.49; 10mg/kg 0.68, 0.68.	Fair	Pre-treatment MHAQ: 0/29 IRT model: 0/2 Post-treatment 3,6,12mo MHAQ 0all/ 12,17,18 IRT model 0all/2,3,5	NR

Functional status measure (MHAQ continued)	**Validity (hypothesis testing)**			**Responsiveness**			**Interpretability**	
	Design	**Result**	**COSMIN score**	**Treatment**	**Result**	**COSMIN score**	**f/c (%)**	**MIC or MID**
Nagasawa 2010 HAQ-DI & MHAQ (71)	Criterion Validity; Hypothesis testing (Convergent)	Strong correlation between HAQ-DI and the mHAQ score (r = 0.892, p<0.0001); Additional correlations reported between baseline HAQ-DI and mHAQ respectively age (0.144, 0.159), disease duration months (0.029, 0.037), RF titre (0.227, 0.164), TJC (0.443*, 0.412*), SJC (0.254***, 0.144), PtG (0.566*, 0.515*), CRP (0.218**,0.167) , DAS28 (0.562*, 0.494*), MMP-3 (0.052,-0.127), vdH-Sharp score baseline (0.139, 0.118). *p<0.0001, **p<0.01, ***p<0.05	CV: Fair HT: Fair	N/A	N/A	N/A	NR	NR
Pincus 1983 HAQ-DI vs MHAQ (27)	Criterion validity; Hypothesis testing (Convergent)	Correlations between MHAQ vs HAQ in included activities: *r*=0.708-0.840 (all p<0.001); Chronbach's alpha 0.710-0.890 (no correlations reported for overall scoring). MHAQ: Correlations of mean scores between difficulty and satisfaction, change and help *r*=0.694, 0.380, 0.229** (* p<0.001, **p<0.002)[3]	CV: Fair	N/A	N/A	N/A	NR	NR
Russell 2003 MHAQ (72)	N/A	N/A	N/A	Infliximab	MHAQ ES 0.62; SRM 0.74; 58% improved by >2 SEM; 48% improved by 95% Bland-Altman Limits of Agreement.	Poor	NR	SDD 0.27

Stucki 1995 MHAQ vs HAQ “original” (73)	Hypothesis testing (Convergent)	Comparison of the rank correlation of the HAQ and the difficulty section of the MHAQ with clinical and lab parameters. Disease activity physician (HAQ=0.55**, MHAQ 0.45**); DAS (HAQ=0.53**, MHAQ=0.53**); Mallya index (HAQ 0.74**, MHAQ, 0.59**); SJC (HAQ=0.25, MHAQ=0.21); TJC (HAQ=0.55**, MHAQ=0.51**); Grip strength (HAQ=-0.62**, MHAQ=-0.51**); strength index (HAQ=-0.61**, MHAQ=-0.52**); pain (HAQ=0.54**, MHAQ=0.52**); AM stiffness (HAQ=0.55**, MHAQ=0.36*); ESR (HAQ=0.23, MHAQ=0.33*), Hgb (HAQ=-0.17, MHAQ=-0.11) *p<0.05, **p<0.01	Fair	Unclear [presumably usual care]	Pearson correlation between change in HAQ and changes in: physician's estimate of disease activity (r= 0.27, p < 0-05); Mallya index (r= 0.30 p < 0.05); pain (r= 0.44, p < 0-01); strength index (r=-0.36, p < 0-0 1); patient's perception of change (r = 0.29, p<0-05). The correlations with morning stiffness, DAS, swollen and tender joint counts, ESR, and hemoglobin were not significant. In a parametric analysis (assuming interval characteristic of the MHAQ) change in MHAQ correlated only with change in pain (r =0.32, p < 0-05)	Fair	NR	NR

Functional status measure (MHAQ continued)	Validity (hypothesis testing)			Responsiveness			Interpretability	
	Design	Result	COSMIN score	Design	Result	COSMIN score	f/c (%)	MIC or MID
Tugwell 2000 HAQ-DI vs MHAQ (74)	N/A	N/A	N/A	RCT of Leflunomide vs placebo or methotrexate	To detect a treatment effect of leflunomide vs placebo: MHAQ SES:-0.69, RE 1.37, Z statistic 0.80, p=0.422; HAQ-DI SES -0.80; RE 1.84; Z statistic 1.60, p= 0.110. To detect a treatment effect of methotrexate vs placebo MHAQ SES:-0.43, RE 0.91, Z statistic 0.17, p=0.884; HAQ-DI SES -0.43; RE 0.91; Z statistic 0.1, p= 0.879. Comparator for both: TJC	Fair	NR	NR
Uhlig 2006 HAQ (presumed DI) vs MHAQ (75)	Hypothesis testing (Convergent); Criterion validity	Pearson correlation coefficients (all p<0.01) HAQ adjusted & i) AIMS physical component 0.82 ii) SF 36P= 0.79; HAQ not adjusted and i) AIMS PC 0.82 ii) SF36P 0.78; MHAQ and i) AIMS physical component 0.82 and SF 36P=0.71. Also all domains of SF-36 examined with correlations in Table 5 (data not abstracted); For the following only significantly correlated findings from Table 5 in the following order HAQadjusted/ HAQunadjusted/MHAQ: SJC(66) 0.43/ 0.39/ 0.33; TJC(68) 0.43/ 0.41 /0.30; Ritchie score 0.61/ 0.63 /0.58; CRP 0.32 /0.28/ 0.29; Grip strength 0.55/ 0.52/ 0.42; Fatigue 0.42 /0.40 /0.38; Patient global 0.28/ 0.28 /0.27; pain 0.58 0.58 0.62(all above p<0.001); ESR 0.20** 0.17*** 0.14NS; Modified Sharp Score 0.25** 0.20*** 0.13NS **p<0.01; ***p<0.05	HT: Fair CV: Fair	N/A	N/A	N/A	NR	NR

		Criterion validity: Pearson correlation coefficients HAQ adjusted and MHAQ=0.85; HAQ not adjusted and MHAQ=0.88						

Functional status measure (MHAQ continued)	Validity (hypothesis testing)			Responsiveness			Interpretability	
	Design	**Result**	**COSMIN score**	**Design**	**Result**	**COSMIN score**	**f/c (%)**	**MIC or MID**
Wolfe 2001 HAQ-DI, MHAQ, RA-HAQ, and DHAQ and HAQ20 (76)	Structural Validity	The HAQ had one non-fitting item, "take a tub bath," but the non-fit was large, 1.57 and 1.51 for the INFIT and OUTFIT, respectively. The MHAQ had 2 slightly non-fitting items "turn taps on and off" and "lift a full cup or glass to the mouth." The INFIT and OUTFIT statistics for these items ranged between 1.20 and 1.29.	Fair	N/A	N/A Reports RE's instead: Compared to the MHAQ, the HAQ relative efficiency is 1.28, and compared to the RA-HAQ it is 1.37.	N/A	Percent with 0 for: HAQ 4.00%; MHAQ 12.84%; RA-HAQ 12.26%; HAQ no assistive devices 6.48%; HAQ difficult 8 items 5.51%; HAQ 20 items 5.00% Percent with highest scores for: HAQ 0.12%; MHAQ 0%; RA-HAQ 0.04% (Table 3)	NR
Ziebland 1992 HAQ (assumed DI) vs MHAQ (77)	N/A	N/A	N/A	Presumed usual care	Pearson's *r* for change scores for HAQ and Ritchie 0.18, Grip strength 0.41**, Pain 0.26*, AM stiffness 0.20, ESR 0.29*, Hgb 0.1, Global transition item 0.4** . For MHAQ and Ritchie 0.40**, Grip strength 0.40**, Pain 0.47**, AM stiffness 0.35**, ESR 0.51**, Hgb 0.32*, Global transition item 0.77** (*p<0.01. **p<0.001)	Fair	NR	NR

MDHAQ								
Functional status measure	**Validity (hypothesis testing)**			**Responsiveness**			**Interpretability**	
	Design	**Result**	**COSMIN score**	**Treatment**	**Result**	**COSMIN score**	**f/c(%)**	**MIC or MID**
Pincus 2005 Compares 10-ADL MDHAQ to other versions e.g. 14-ADL MDHAQ, 20-ADL and 8-ADL MDHAQs (29)	Structural validity	The HAQ and MHAQ PF scales formed one factor only. The 14-ADL MDHAQ scale formed 3 factors. The 10-ADL MDHAQ scale formed 2 factors.	Poor	N/A	N/A	N/A	NR	NR
Pincus 1999 MDHAQ, MHAQ and HAQ-DI (28)	Hypothesis testing (convergent)	Spearman's rho for MHAQ correlations with: Age 0.08*, duration of disease 0.12**, formal education level -0.24, Advanced ADL 0.75, psychological items 0.50, pain (VAS) 0.57; Fatigue (VAS) 0.46, helplessness index 0.51, AIMS anxiety 0.33, AIMS depression 0.43, complete Beck depression index 0.49, Center for Epidemiologic Studies Depression Scale 0.45 (all $p<0.001$ except *$p<0.05$ and **$p<0.01$)[13]. MHAQ correlations with Sleep 0.51; Stress 0.44; Anxiety 0.35; Depression 0.37 (all $p<0.001$)	Fair	N/A	N/A	N/A	MHAQ 22%/NR HAQ:16%/NR MDHAQ not clearly reported	NR

Wolfe 2004 HAQ-II vs HAQ, MHAQ, SF36, MDHAQ **(Responsiveness evaluated only in HAQ and HAQ-II)** (30)	Hypothesis Testing (Convergent) Structural Validity Criterion validity	Correlations for HAQ-II, HAQ and MHAQ with SF-36 -0.85, - 0.80, -0.72; EuroQol utility (0–1 scale) with HAQ-II, HAQ and MHAQ -0.67 - 0.64 -0.69. RADAI score (0–10) with HAQ-II, HAQ and MHAQ 0.65 0.63 0.66 Rheumatology Distress Index (0–100 scale) 0.61 0.59 0.61; Global disease severity (0–10 VAS) 0.61 0.58 0.59 ; Pain (0–10 VAS) 0.61 0.59 0.61 ; Fatigue (0–10 VAS) 0.56 0.54 0.52 ; SF-6D utility (0–1 scale) -0.56 -0.54 -0.48;Work Limitations Questionnaire index (0–100 scale) 0.56 0.54 0.55; QOL scale (0–100 VAS) -0.54 -0.51 - 0.52; AIMS depression scale (0–10) 0.44 0.42 0.47; Sleep disturbance (0–10 scale) 0.41 0.40 0.42; AIMS anxiety scale (0–10) 0.38 0.36 0.41; Social security disability (%) 0.34 0.32 0.34 ; GI severity (0–10 scale) 0.33 0.31 0.34; Total direct medical costs, $ 0.24 0.23 0.20;Total joint replacement, % 0.18 0.20 0.13 . In RA Validation Study (n=693) for HAQ-II, HAQ and MHAQ Pain (0–10 VAS) 0.66 0.66 0.67; PtG (0–10 VAS) 0.62 0.60 0.61; Fatigue (0–10 VAS) 0.57 0.56 0.55; DAS28 0.51 0.54 0.50; PhG; severity (0–10 VAS) 0.48 0.50 0.50; Disability (stopped work) 0.41 0.42 0.35; TJC (range 0–28) 0.37 0.39 0.40; ESR 0.25 0.27 0.22; SJC (range 0–28) 0.24 0.27 0.25; Joint surgery, no/yes 0.20 0.23	HT: Fair SV: Fair CV: Fair	Usual care (presumed)	ES HAQ-II was 23.0 [95% CI 18.4–27.4). ES for HAQ was 24.8 (95% CI 20.0–29.5). These differences were not significant (P =0.298).	Fair	validation study (n=14038) percent with lowest scores (0): HAQ 10.1%; HAQ-II 5.8%; M-HAQ 24.5%; SF-36 3.4% MD-HAQ from a separate sample (n=15,543) 4.4% with scores of 0. validation study (n=14038) percent with highest scores (3): HAQ 0.2%; HAQ-II 0.1%; M-HAQ 0.2%; SF-36 3.0%	NR

		0.11. There were no significant differences in the correlations among the questionnaires. Structural Validity "The HAQ-II had the longest scale, as measured in logits, indicating that it captured more of the continuum of disability than did the other questionnaires. The MD-HAQ also had a long scale, by virtue of the difficult items "participate in sports and games" and "walk 2 miles." However, these items misfit the Rasch model, indicating a lack of unidimensionality and/or inaccurate assessment. The HAQ also had items that did not fit the Rasch model. Within the HAQ hygiene category, the items "Take a tub bath" and "shampoo hair" misfit the model. This, in turn, led to the misfitting of the hygiene category." We also noted gaps in the scales of all the HAQ family questionnaires except for the HAQ-II." Correlations with HAQ: HAQ-II 0.91, MHAQ 0.84						

PROMIS Physical Function subscale								
Functional status measure	**Validity (hypothesis testing)**			**Responsiveness**			**Interpretability**	
	Design	**Result**	**COSMIN score**	**Treatment**	**Result**	**COSMIN score**	**f/c(%)**	**MIC or MID**
Bartlett 2015 PROMIS PF CAT MHAQ (33)	Hypothesis testing (Convergent)	Pearson's r with PROMIS subscales*:* Pain intensity -.561, Pain interfere -.709, Fatigue -.635, Sleep disturbance -.376, Sleep impairment -.432, depression -.398, Anxiety -.361, Anger -.229, Ability to participate Social 0.698, Satisfaction with role activities 0.627 (all p≤0.01). Correlation with legacy measures Pain VAS -.593, Patient Global VAS-.688. Pearson's r with legacy MHAQ -0.752.	Fair	N/A	N/A	N/A	NR (Reports 46% scored 0 on MHAQ)	NR
Fries 2011 PROMIS (78) Note: this study appears to be related to study below also by Fries from the same year.	N/A	N/A	N/A	Usual care (presumably)	"All instruments were sensitive to change in PF status, with p-values for changes in PF scores ranging from 0.001 to 0.05 and SRM and ES computations mirroring these results. The most responsive were the PROMIS 20-item Short Forms. Under study conditions, IRT-Improved instruments could detect a 1.2 % difference with 80 % power, while reference instruments could detect only a 2.4 % difference (p <0.01). Sample sizes required for the best IRT-improved instruments were only 24% of the worst Legacy comparator (100 vs. 427)."[14]	Fair	NR[15]	NR
Fries 2011 PROMIS PF-10††, PROMIS PF- 20††, Legacy HAQ, SF-36 PF-10, Item-Improved HAQ, Item-Improved PF-10 (79)	N/A	N/A	N/A	Usual care (presumably)	All PF scales were responsive to change in function over 12 months (P<0.05). SRM's: Legacy PF-10 0.10, Legacy HAQ 0.14, Item Improved PF-10	Fair	NR	MDD: Legacy PF-10 2.43, Legacy HAQ 1.40, Item Improved PF-10 2.16, Item improved HAQ

					0.09, Item improved HAQ 0.13, PROMIS PF 10 0.13, PROMIS PF 20 0.13. Cohen's ES: Legacy PF-10 0.06, Legacy HAQ 0.06, Item Improved PF-10 0.05, Item improved HAQ 0.05, PROMIS PF 10 0.05, PROMIS PF 20 0.05. Also reports Guyatt's ES.			1.14, PROMIS PF 10 1.47, PROMIS PF 20 1.24. MIC/MID not reported
PROMIS continued								
Hays 2015 PROMIS PF 20† HAQ-DI, SF-36 (80)	Hypothesis testing (Convergent)	Correlation (unspecified type) between PF-20 and SF-36 (0.84) and HAQ-DI (-0.89)	Fair	Usual care (presumably)	Product–moment (Spearman) correlations PF-20: 0.35 (0.33) at 12mo and 0.34 (0.33) at 6mo. HAQ: 0.29 (0.25) at 12mo, 0.29 (0.25) at 6mo Comparator: anchor item[17]	Fair		MID for PF-20 was 2 points (about 0.20 of an SD)
Oude Voshaar 2014 PROMIS CATs with 5, 10, 15 items Vs HAQ-DI, SF-36 (34)	Hypothesis testing (convergent)	Correlations between IRT-based and standard scores were 0.97 (HAQ) 0.95 (SF-36 PF 10)	Fair	N/A	N/A[18]	N/A	NR	NR
Oude Voshaar 2015 PROMIS item bank and 20-item SF††, compared to HAQ-DI and SF-36 PF-10 (81)	Content validity; Hypothesis testing (convergent)	All the evaluated items[19] of the HAQ-DI, PF-10 and PROMIS PF item bank refer to health concepts that are relevant indicators of PF in RA. Pearson *r*: PROMIS PF (approx 40 items) and HAQ-DI (0.76), SF-36 PF-10 (0.84). Pearson's *r* between PROMIS PF (approx 40 items) HAQ-DI and Pain (-0.52, 0.52), General health (-0.53, 0.48), Disease activity (-0.46, 0.50), Fatigue (-0.47, 0.46), Stiffness (-0.63, 0.62), Age (0.14, -0.07*) (all significant at p<0.05 level except item indicated by *)[20]	Excellent; Good	N/A	N/A	N/A	Reports 53% scored 0 on HAQ (not reported specifically for other measures)	NR
Schalet et al 2016 PROMIS PF-10 SF & PF-20§ (82)	N/A	N/A	N/A	Usual care (presumably)	SRMs: RA better (0.21), about the same (-0.12), worse (-0.19)	Fair	NR	NR

					Comparator: general health anchor SRMs: RA better (0.29), about the same (0.03) , worse (-0.46) Comparator: general PF anchor			
Wahl 2017 PROMIS PF-10a (31)	Hypothesis testing (convergent, discriminant & known groups)	*r*: strong correlations with HAQ (-0.874) and patient global assessment of RA activity (-0.720), and moderate correlation with pain scores (-0.631). No correlation with SJC or TJC (r<-0.446) Known-groups: ES (Cohen's d) was large in the group dichotomized by disease activity (0.93), moderate by age (0.62), no difference by seropositive status, history of erosive disease or joint replacement.	Good	Usual care (presumably)	Mean PF-10a scores differed significantly between groups (P<0.001). SRM moderate in the improvement group (0.73), small in the groups with stable disease (20.02) and clinical deterioration (20.43). Linear mixed-effects modeling showed that changes in CDAI scores over time were associated with changes in PF-10a scores over time (P<0.001).	Good	HAQ 0/19% vs PF-10a <1%/8% (p<0.0001)	NR

ADL: Activities of Daily Living; AE: Adverse Effects; ANOVA: Analysis of Variance; ARA: American Rheumatism Association; AUC: Area Under the Curve; Ave: Average; C: Ceiling; COBRA Clinical Trial: Combinatietherapie Bij Reumatoide Artritis; DIF: Differential Item Functioning; ES: Effect Size; ESR: Erythrocyte sedimentation rate; F: Floor; Hgb: Hemoglobin; ICF: International Classification of Functioning, Disability and Health; IRT: Item response theory; MIC: Minimal Important Change; MID: Minimal Important Difference; MDD: Minimal Detectable difference; MM: Mallya and Mace Index; Mo: Month; N/A: Not applicable; NR: Not reported; NS: Not Significant;PF: Physical function; PhG: Physician Global; PtG: Patient Global; Pop: population; RID: Really Important Difference; RE: Relative efficiency statistic; RV: Relative validity; SDD: Smallest Detectable Difference; SES: Standardized Effect Size; SRMs: Standardized Response Means; Sens: Sensitivity; Spe: Specificity; SJC: Swollen Joint Count; TJC: Tender Joint Count; VAS: Visual Analog Scale; vdH-Sharp score: van der Heijde modification of the Sharp score

[1] Using ratio of effect sizes/Using Analysis of Covariance

[2] Standard effect sizes also likely calculated for the HAQ but not reported in manuscript (reference to all non-reported SES's as "similar magnitude for the remaining outcome measures."

[3] Also reports ES stratified by disease duration (<2years and ≥2 years) and based on the presence of "second-line therapy"

[4] Eigenvalue

[5] While HAQ isn't dichotomous, the authors created cut-off points in order to examine sensitivity and specificity. Cutoffs chosen: to indicate that the majority of patients had not changed and one that would indicate change of health status for the majority of patients (>0.25 change for former and any score change for latter).

[6] Correlations and Effect Sizes examined at 2 time points. Study also calculates correlations and effect sizes to parts of the HAQ (not overall score, e.g. mobility, activities of daily living and household activities), too many comparisons to abstract and hard to compare results to other studies so not shown here

[7] Study used Principal Component Analysis

[8] Study used discriminant analysis to determine concurrent validity (not a type of validity that COSMIN recognizes so quality not rated); examined how well HAQ classified individuals into disease groups comparing 30 persons with arthritis compared to diabetes, 78% correctly classified, exercise repeated with 2 groups of 30 persons and 80% correctly classified.

[9] Patient acceptable symptom state (PASS) and Minimal Clinically Important Improvement (MCII) cutpoints with 2 methodological approaches for health-related quality of life and health status measures after 3 months of DMARD treatment in patients with RA

[10] Relative validity (RV) coefficients calculated from ANOVA and ANCOVA to quantify gain (or loss) in validity of the IRT-scored scales compared to MHAQ (and physical function-10, PF-10 measures). The MHAQ was about 70% as efficient as the overall IRT-based score of physical functioning in discriminating among American College of Rheumatology (ACR) groups. Not shown above but RV analysis also used to examine treatment groups and MHAQ was 25% less efficient than the overall IRT-based score of physical functioning in discriminating among treatment groups. Also reported RV scores for upper and lower extremity and PF-10 (all not reported here).

[11] Study also reports effect sizes for 2mg/kg abatacept dose and for upper and lower extremity scales as well as PF-10.

[12] Effect sizes calculated for dichotomous measures of RA severity

[13] Pincus 1999 study also reports on correlations to a 6-item advanced ADL score (not reported here, similar correlations)

[14] Fries 2011 study states that "Our objective was to compare responsiveness between change scores on subsets of PROMIS items and change scores on Legacy instruments to these alternative PRO measures

and to test whether more informative items would reduce sample size requirements"; however, the comparisons that were made were not clearly outlined in the methods or in the results and the paragraph abstracted is the totality of the results presented.

[15] Although Fries 2011(78) study discusses Floor and Ceiling issues, % at highest and lowest responses not reported. Figure 2 in the paper shows sample size-power estimates for different population characteristics. Further details published in next study by same author shown here.

[16] Fries 2011 (79) sample size requirements that are sufficient to detect a change score of 2.5 units on a 0 to 100 scale were also reported (not abstracted here).

[17] Anchor item: "We would like to know about any changes in how you are feeling now compared with how you were feeling 6 months ago. How has your ability to carry out your everyday physical activities such as walking, climbing stairs, carrying groceries, or moving a chair got a lot better, got a little better, stayed the same, got a little worse, or got a lot worse?'

[18] Oude Voshaar 2014 (34)Reports on sensitivity to change but methods not applicable to COSMIN reporting (e.g. evaluated by the ability of a test to detect simulated change of scores of small to moderate magnitude (standardized ESs 0.2, 0.35, 0.50). Concluded that "Substantially improved sensitivity to change was observed for the CAT-10 compared with the HAQ DI and PF-10, particularly in detecting moderate effect sizes."

[19] Oude Voshaar 2014 (81) compares 3 physical function (PF) measures: HAQ-DI, SF-36 PF-10 and PROMIS Item Bank

[20] Oude Voshaar 2014 (81) also evaluates relative validity of the instruments in differentiating between patients in remission and active disease using DAS28. The conclusion was that the HAQ-DI and PROMIS items were about equally efficient, while the PF-10 was less efficient in distinguishing between levels of disease activity. Data not presented in table as method of analysis not

[†] Probably PF 20a as same questions presented but not specified in manuscript

[††] In these studies further specification of the type of SF not given (e.g. PF-10a, 20a not specified)

[§] In this study by Schalet et al. in methods for RA only short forms were given (both 10 and 20 item versions), in results they appear to be reported together. Further specification of the type of short form (e.g. PF-10a) not provided.

easily evaluable using COSMIN.

[21] Partial correlations represent the pooled within person correlations between measures derived from pooled time series analyses.

Medline Search Strategy

The Medline search strategy is described below. This strategy uses MeSH terms and keywords across three themes: #1 construct search (for assessment of functional status), #2 population search (rheumatoid arthritis) and #3 instrument search (including terms for instruments of interest e.g., questionnaires, etc.). The Boolean search operator "AND" was used to combine the 3 search themes

1. exp Health status/
2. 'Health level*'.tw,kw.
3. 'Health Status*'.tw,kw.
4. 'Level* of health'.tw,kw.
5. exp Disability evaluation/
6. (Disability adj2 assessment*).tw,kw.
7. (functional adj2 assessment*).tw,kw.
8. (Disability adj2 evaluation*).tw,kw.
9. exp Health status indicator/
10. 'Health status index*'.tw,kw.
11. 'Health status indic*'.tw,kw.
12. exp Severity of illness index/
13. 'Severity of illness ind*'.tw,kw.
14. exp Activities of daily living/
15. daily life activit*.tw,kw.
16. ADL*.tw,kw.
17. (Activit* adj2 living).tw,kw.
18. exp patient outcome assessment/
19. 'Patient-centered outcome* research'.tw,kw.
20. 'Patient reported outcome*'.tw,kw.
21. 'Patient perspective*'.tw,kw.
22. 'outcome* research'.tw,kw.
23. (outcome* adj2 assessment*).tw,kw.
24. 'functional status'.tw,kw.
25. 'function* impair*'.tw,kw.
26. 'Health assessment questionnaire'.tw,kw.
27. HAQ*.tw,kw.
28. MHAQ.tw,kw.
29. MDHAQ.tw,kw.
30. PROMIS.tw,kw.
31. 'Short Form 36'.tw,kw.
32. SF-36.tw,kw.

33. or/1-32

34. exp "Surveys and Questionnaires"/

35. Survey*.tw,kw.

36. Questionnaire*.tw,kw.

37. Index*.tw,kw.

38. Scale*.tw,kw.

39. Instrument*.tw,kw.

40. tool*.tw,kw.

41. diar*.tw,kw.

42. assessment*.tw,kw.

43. 'self-report*'.tw,kw.

44. measure*.tw,kw.

45. prom.tw,kw.

46. checklist*.tw,kw.

47. rating.tw,kw.

48. or/34-47

49. instrumentation.fs.

50. methods.fs.

51. validation studies.pt.

52. comparative study.pt.

53. exp Validation studies/

54. exp "Outcome Assessment (Health Care)"/

55. outcome measure*.tw,kw.

56. validation Stud*.tw,kw.

57. Validate.tw,kw.

58. Validity.tw,kw.

59. valid*.tw,kw.

60. (homogeneity or homogeneous).tw,kw.

61. ((minimal* or clinic*) and (important or significant or detectable) and (change or difference)).tw,kw.

62. 'minimal* real difference*'.tw,kw.

63. 'ceiling effect'.tw,kw.

64. 'floor effect'.tw,kw.

65. detect* change*.tw,kw.

66. exp "reproducibility of results"/

67. reproducib*.tw,kw.

68. (reliab* or unreliab*).tw,kw.

69. (reliab* and (test or retest)).tw,kw.

70. responsiveness*.tw,kw.

71. 'test-retest'.tw,kw.
72. (test adj1 retest).tw,kw.
73. discriminant analysis.tw,kw.
74. exp observer variation/
75. 'observer variation'.tw,kw.
76. exp Psychometrics/
77. Psychometr*.tw,kw.
78. clinometr*.tw,kw.
79. clinimetr*.tw,kw.
80. coefficient.tw,kw.
81. 'internal consistency'.tw,kw.
82. (cronbach* and alpha*).tw,kw.
83. 'item correlation*'.tw,kw.
84. 'item selection*'.tw,kw.
85. 'item reduction*'.tw,kw.
86. agreement.tw,kw.
87. precision.tw,kw.
88. imprecision.tw,kw.
89. 'precise values'.tw,kw.
90. stability.tw,kw.
91. interrater.tw,kw.
92. 'inter rater'.tw,kw.
93. intrarater.tw,kw.
94. 'intra rater'.tw,kw.
95. intertester.tw,kw.
96. 'inter tester'.tw,kw.
97. intratester.tw,kw.
98. 'intra tester'.tw,kw.
99. interobserver.tw,kw.
100. 'inter observer'.tw,kw.
101. 'intra observer'.tw,kw.
102. interexaminer.tw,kw.
103. 'inter examiner'.tw,kw.
104. intraexaminer.tw,kw.
105. 'intra examiner'.tw,kw.
106. interindividual.tw,kw.
107. 'inter individual'.tw,kw.
108. intraindividual.tw,kw.
109. 'intra individual'.tw,kw.

110. interparticipant.tw,kw.
111. 'inter participant'.tw,kw.
112. intraparticipant.tw,kw.
113. 'intra participant'.tw,kw.
114. (intertechninican or inter-technician or intratechnician or intra-technician).tw,kw.
115. (interassay or inter-assay or intraassay or intra-assay).tw,kw.
116. kappa*.tw,kw.
117. 'coefficient of variation'.tw,kw.
118. repeatab*.tw,kw.
119. ((replicab* or repeated) and (measure* or findings or result* or test*)).tw,kw.
120. tests.tw,kw.
121. (generaliza* or generalisa*).tw,kw.
122. concordance.tw,kw.
123. (intraclass and correlation).tw,kw.
124. discriminative.tw,kw.
125. 'known group'.tw,kw.
126. 'factor analys*'.tw,kw.
127. 'factor structure*'.tw,kw.
128. 'dimension*'.tw,kw.
129. 'multitrait scaling analys*'.tw,kw.
130. (error* and (measure* or correlat* or evaluat* or accuracy or accurate or precision or mean)).tw,kw.
131. 'individual variability'.tw,kw.
132. 'interval variability'.tw,kw.
133. 'rate variability'.tw,kw.
134. (variability and (analysis or values)).tw,kw.
135. (uncertainty and (measurement or measuring)).tw,kw.
136. 'standard error of measurement'.tw,kw.
137. sensitiv*.tw,kw.
138. responsive*.tw,kw.
139. (limit and detection).tw,kw.
140. interpretab*.tw,kw.
141. (small* and (real or detectable) and (change or Difference)).tw,kw.
142. 'meaningful change'.tw,kw.
143. 'item response model'.tw,kw.
144. irt.tw,kw.
145. rasch.tw,kw.
146. 'differential item functioning'.tw,kw.
147. 'cross-cultural equivalence'.tw,kw.

148. 'detect change'.tw,kw.

149. subscale*.tw,kw.

150. item discriminant.tw,kw.

151. interscale correlation*.tw,kw.

152. error*.tw,kw.

153. DIF.tw,kw.

154. "computer adaptive testing".tw,kw.

155. "item bank".tw,kw.

156. or/34-155

157. exp arthritis, rheumatoid/

158. rheumatoid arthritis.tw,kw.

159. 157 or 158

160. 33 and 48 and 156 and 159

161. 160 not ("addresses" or "bibliography" or "case reports" or "comment" or "directory" or "editorial" or "festschrift" or "interview" or "lectures" or "legal cases" or "legislation" or “letter" or "news" or "newspaper article" or "patient education handout" or "popular works" or "congresses" or "consensus development conference" or "consensus development conference, nih" or "practice guideline").pt. not (animals/ not humans.sh.)

162. limit 161 to english

Appendix References

1. Waehrens EE, Bliddal H, Danneskiold-Samsoe B, Lund H, Fisher AG. Differences between questionnaire- and interview-based measures of activities of daily living (ADL) ability and their association with observed ADL ability in women with rheumatoid arthritis, knee osteoarthritis, and fibromyalgia. Scandinavian Journal of Rheumatology. 2012;41:95-102.
2. Weisscher N, Glas CA, Vermeulen M, De Haan RJ. The use of an item response theory-based disability item bank across diseases: accounting for differential item functioning. Journal of Clinical Epidemiology. 2010;63:543-9.
3. Weisscher N, Wijbrandts CAW, de Haan R, Glas CA, Vermeulen M, Tak PP. The Academic Medical Center Linear Disability Score item bank: psychometric properties of a new generic disability measure in rheumatoid arthritis. Journal of Rheumatology. 2007;34:1222-8.
4. Li T, Wells G, Westhovens R, Tugwell P. Validation of a simple activity participation measure for rheumatoid arthritis clinical trials. Rheumatology. 2009;48:170-5.
5. Bakheit AMO, Harries SR, Hull RG. Validity of a self-administered version of the Barthel Index in patients with rheumatoid arthritis. Clinical Rehabilitation. 1995;9:234-7.
6. Bakheit AMO, Harries SR, Hull RG. A Comparison between the Stanford Health Assessment Questionnaire and the Barthel Index in Patients with Rheumatoid Arthritis. British Journal of Occupational Therapy. 1995;58:253-5.
7. Badley EM, Wagstaff S, Wood PH. Measures of functional ability (disability) in arthritis in relation to impairment of range of joint movement. Annals of the Rheumatic Diseases. 1984;43:563-9.
8. El Miedany Y, El Gaafary M, Youssef SS, Palmer D. Incorporating patient reported outcome measures in clinical practice: Development and validation of a questionnaire for inflammatory arthritis. Clinical and Experimental Rheumatology. 2010;28:734-44.
9. Egger MJ, Ward LR, Karg MB, Williams HJ, Reading JC, Alarcon GS, et al. Reliability and validity of the CSSRD Functional Assessment Survey in rheumatoid arthritis. Arthritis and Rheumatism. 1995;8:21-7.
10. Nordenskiold U. Daily activities in women with rheumatoid arthritis. Aspects of patient education, assistive devices and methods for disability and impairment assessment. Scandinavian Journal of Rehabilitation Medicine - Supplementum. 1997;37:1-72.
11. Nordenskiold U, Grimby G, Dahlin-Ivanoff S. Questionnaire to evaluate the effects of assistive devices and altered working methods in women with rheumatoid arthritis. Clinical Rheumatology. 1998;17:6-16.
12. Nordenskiold U, Grimby G, Hedberg M, Wright B, Linacre JM. The structure of an instrument for assessing the effects of assistive devices and altered working methods in women with rheumatoid arthritis. Arthritis Care & Research. 1996;9:358-67.
13. Liang M, Schurman DJ, Fries J. A patient-administered questionnaire for arthritis assessment. Clinical Orthopaedics and Related Research. 1978;NO.131:123-9.
14. Doeglas D, Krol B, Guillemin F, Suurmeijer T, Sanderman R, Smedstad LM, et al. The assessment of functional status in rheumatoid arthritis: a cross cultural, longitudinal comparison of the Health Assessment Questionnaire and the Groningen Activity Restriction Scale. J Rheumatol. 1995;22:1834-43.
15. Suurmeijer TPBM, Doeglas DM, Moum T, Briancon S, Krol B, Sanderman R, et al. The Groningen Activity Restriction Scale for measuring disability: Its utility in international comparisons. American Journal of Public Health. 1994;84:1270-3.

16. Lee P, Jasani MK, Dick WC, Buchanan WW. Evaluation of a functional index in rheumatoid arthritis. Scandinavian Journal of Rheumatology. 1973;2:71-7.
17. Goodacre L, Smith J, Meddis D, Goodacre J. Development and validation of a patient-centred Measure of Activity Limitation (MAL) in rheumatoid arthritis. Rheumatology. 2007;46:703-8.
18. Archenholtz B, Dellhag B. Validity and reliability of the instrument Performance and Satisfaction in Activities of Daily Living (PS-ADL) and its clinical applicability to adults with rheumatoid arthritis. Scandinavian Journal of Occupational Therapy. 2008;15:13-22.
19. Salaffi F, Ciapetti A, Gasparini S, Migliore A, Scarpellini M, Corsaro SM, et al. Comparison of the Recent-Onset Arthritis Disability questionnaire with the Health Assessment Questionnaire disability index in patients with rheumatoid arthritis. Clinical and Experimental Rheumatology. 2010;28:855-65.
20. Salaffi F, Franchignoni F, Giordano A, Ciapetti A, Gasparini S, Ottonello M, et al. Classical test theory and Rasch analysis validation of the Recent-Onset Arthritis Disability questionnaire in rheumatoid arthritis patients. Clinical Rheumatology. 2013;32:211-7.
21. Salaffi F, Stancati A, Neri R, Grassi W, Bombardieri S. Measuring functional disability in early rheumatoid arthritis: the validity, reliability and responsiveness of the Recent-Onset Arthritis Disability (ROAD) index. Clinical & Experimental Rheumatology. 2005;23:S31-42.
22. Katz PP, Radvanski DC, Allen D, Buyske S, Schiff S, Nadkarni A, et al. Development and validation of a short form of the valued life activities disability questionnaire for rheumatoid arthritis. Arthritis care & research. 2011;63:1664-71.
23. Tubach F, Ravaud P, Martin-Mola E, Awada H, Bellamy N, Bombardier C, et al. Minimum clinically important improvement and patient acceptable symptom state in pain and function in rheumatoid arthritis, ankylosing spondylitis, chronic back pain, hand osteoarthritis, and hip and knee osteoarthritis: Results from a prospective multinational study. Arthritis Care Res (Hoboken). 2012;64:1699-707.
24. Wolfe F, Michaud K, Pincus T. Preliminary evaluation of a visual analog function scale for use in rheumatoid arthritis. Journal of Rheumatology. 2005;32:1261-6.
25. Tomlin GS, Holm MB, Rogers JC, Kwoh CK. Comparison of standard and alternative health assessment questionnaire scoring procedures for documenting functional outcomes in patients with rheumatoid arthritis. Journal of Rheumatology. 1996;23:1524-30.
26. Fries JF, Spitz P, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. Arthritis & Rheumatism. 1980;23:137-45.
27. Pincus T, Summey JA, Soraci Jr SA, Wallston KA, Hummon NP. Assessment of patient satisfaction in activities of daily living using a modified Stanford Health Assessment questionnaire. Arthritis and Rheumatism. 1983;26:1346-53.
28. Pincus T, Swearingen C, Wolfe F. Toward a multidimensional Health Assessment Questionnaire (MDHAQ): assessment of advanced activities of daily living and psychological status in the patient-friendly health assessment questionnaire format. Arthritis & Rheumatism. 1999;42:2220-30.
29. Pincus T, Sokka T, Kautiainen H. Further development of a physical function scale on a Multidimensional Health Assessment Questionnaire for standard care of patients with rheumatic diseases. Journal of Rheumatology. 2005;32:1432-9.
30. Wolfe F, Michaud K, Pincus T. Development and validation of the health assessment questionnaire II: a revised version of the health assessment questionnaire. Arthritis Rheum. 2004;50:3296-305.

31. Wahl E, Gross A, Chernitskiy V, Trupin L, Gensler L, Chaganti K, et al. Validity and Responsiveness of a 10-Item Patient-Reported Measure of Physical Function in a Rheumatoid Arthritis Clinic Population. Arthritis care & research. 2017;69:338-46.
32. Fries JF, Cella D, Rose M, Krishnan E, Bruce B. Progress in assessing physical function in arthritis: PROMIS short forms and computerized adaptive testing. J Rheumatol. 2009;36:2061-6.
33. Bartlett SJ, Orbai AM, Duncan T, DeLeon E, Ruffing V, Clegg-Smith K, et al. Reliability and Validity of Selected PROMIS Measures in People with Rheumatoid Arthritis. PLoS ONE [Electronic Resource]. 2015;10:e0138543.
34. Oude Voshaar MAH, Ten Klooster PM, Glas CAW, Vonkeman HE, Krishnan E, Van De Laar MAFJ. Relative Performance of Commonly Used Physical Function Questionnaires in Rheumatoid Arthritis and a Patient-Reported Outcomes Measurement Information System Computerized Adaptive Test. Arthritis and Rheumatology. 2014;66:2900-8.
35. Bombardier C, Raboud J. A comparison of health-related quality-of-life measures for rheumatoid arthritis research. The Auranofin Cooperating Group. Controlled Clinical Trials. 1991;12:243S-56S.
36. Brown JH, Kazis LE, Spitz PW, Gertman P, Fries JF, Meenan RF. The dimensions of health outcomes: a cross-validated examination of health status measurement. American Journal of Public Health. 1984;74:159-61.
37. Buchbinder R, Bombardier C, Yeung M, Tugwell P. Which outcome measures should be used in rheumatoid arthritis clinical trials? Clinical and quality-of-life measures' responsiveness to treatment in a randomized controlled trial. Arthritis & Rheumatism. 1995;38:1568-80.
38. Cole JC, Motivala SJ, Khanna D, Lee JY, Paulus HE, Irwin MR. Validation of single-factor structure and scoring protocol for the Health Assessment Questionnaire-Disability Index. Arthritis & Rheumatism. 2005;53:536-42.
39. Fitzpatrick R, Newman S, Lamb R, Shipley M. A comparison of measures of health status in rheumatoid arthritis. British Journal of Rheumatology. 1989;28:201-6.
40. Fitzpatrick R, Ziebland S, Jenkinson C, Mowat A. A comparison of the sensitivity to change of several health status instruments in rheumatoid arthritis. Journal of Rheumatology. 1993;20:429-36.
41. Fitzpatrick R, Ziebland S, Jenkinson C, Mowat A, Mowat A. Importance of sensitivity to change as a criterion for selecting health status measures. Quality in Health Care. 1992;1:89-93.
42. Goeppinger J, Doyle MA, Charlton SL, Lorig K. A nursing perspective on the assessment of function in persons with arthritis. Res Nurs Health. 1988;11:321-31.
43. Greenwood MC, Doyle DV, Ensor M. Does the Stanford Health Assessment Questionnaire have potential as a monitoring tool for subjects with rheumatoid arthritis? Annals of the Rheumatic Diseases. 2001;60:344-8.
44. Hakkinen A, Kautiainen H, Hannonen P, Ylinen J, Arkela-Kautiainen M, Sokka T. Pain and joint mobility explain individual subdimensions of the health assessment questionnaire (HAQ) disability index in patients with rheumatoid arthritis. Annals of the Rheumatic Diseases. 2005;64:59-63.
45. Hawley DJ, Wolfe F. Sensitivity to change of the health assessment questionnaire (HAQ) and other clinical and health status measures in rheumatoid arthritis: results of short-term clinical trials and observational studies versus long-term observational studies.[Erratum appears in Arthritis Care Res 1992 Dec;5(4):229]. Arthritis Care & Research. 1992;5:130-6.
46. Hendrikx J, Fransen J, Kievit W, van Riel PLCM. Individual patient monitoring in daily clinical practice: a critical evaluation of minimal important change. Quality of Life Research. 2015;24:607-16.

47. Kosinski M, Zhao SZ, Dedhiya S, Osterhaus JT, Ware Jr JE. Determining minimally important changes in generic and disease-specific health-related quality of life questionnaires in clinical trials of rheumatoid arthritis. Arthritis and Rheumatism. 2000;43:1478-87.
48. Lassere MN, van der Heijde D, Johnson KR, Boers M, Edmonds J. Reliability of measures of disease activity and disease damage in rheumatoid arthritis: implications for smallest detectable difference, minimal clinically important difference, and analysis of treatment effects in randomized controlled trials. Journal of Rheumatology. 2001;28:892-903.
49. Linde L, Sørensen J, Ostergaard M, Hørslev-Petersen K, Hetland ML. Health-related quality of life: validity, reliability, and responsiveness of SF-36, EQ-15D, EQ-5D, RAQoL, and HAQ in patients with rheumatoid arthritis. Journal of Rheumatology. 2008;35:1528-37.
50. Marra CA, Rashidi AA, Guh D, Kopec JA, Abrahamowicz M, Esdaile JM, et al. Are indirect utility measures reliable and responsive in rheumatoid arthritis patients? Qual Life Res. 2005;14:1333-44.
51. Marra CA, Woolcott JC, Kopec JA, Shojania K, Offer R, Brazier JE, et al. A comparison of generic, indirect utility measures (the HUI2, HUI3, SF-6D, and the EQ-5D) and disease-specific instruments (the RAQoL and the HAQ). Social Science & Medicine. 2005;60:1571-82.
52. Pope JE, Khanna D, Norrie D, Ouimet JM. The minimally important difference for the health assessment questionnaire in rheumatoid arthritis clinical practice is smaller than in randomized controlled trials. Journal of Rheumatology. 2009;36:254-9.
53. Redelmeier DA, Lorig K. Assessing the clinical importance of symptomatic improvements. An illustration in rheumatology. Archives of Internal Medicine. 1993;153:1337-42.
54. Rohekar G, Pope J. Test-retest reliability of patient global assessment and physician global assessment in rheumatoid arthritis. J Rheumatol. 2009;36:2178-82.
55. Seror R, Tubach F, Baron G, Guillemin F, Ravaud P, Seror R, et al. Measure of function in rheumatoid arthritis: individualised or classical scales? Annals of the Rheumatic Diseases. 2010;69:97-101.
56. Sheehan TJ, DeChello LM, Garcia R, Fifield J, Rothfield N, Reisine S. Measuring disability: application of the Rasch model to activities of daily living (ADL/IADL). Journal of Outcome Measurement. 2001;5:839-63.
57. Singer F, Kolarz G, Mayrhofer F, Scherak O, Thumb N. The use of questionnaires in the evaluation of the functional capacity in rheumatoid arthritis. Clin Rheumatol. 1982;1:251-61.
58. Sousa KH, Kwok OM, Ryu E, Cook SW. Confirmation of the validity of the HAQ-DI in two populations living with chronic illnesses. Journal of Nursing Measurement. 2008;16:31-42.
59. Sullivan FM, Eagers RC, Lynch K, Barber JH. Assessment of disability caused by rheumatic diseases in general practice. Annals of the Rheumatic Diseases. 1987;46:598-600.
60. Taylor WJ, McPherson KM. Using Rasch analysis to compare the psychometric properties of the Short Form 36 Physical Function score and the Health Assessment Questionnaire Disability Index in people with psoriatic arthritis and rheumatoid arthritis. Arthritis Care & Research. 2007;57:723-9.
61. Tennant A, Hillman M, Fear J, Pickering A, Chamberlain MA. Are we making the most of the Stanford Health Assessment Questionnaire? British Journal of Rheumatology. 1996;35:574-8.
62. Verhoeven AC, Boers M, van Der Linden S. Responsiveness of the core set, response criteria, and utilities in early rheumatoid arthritis. Ann Rheum Dis. 2000;59:966-74.
63. Ward MM, Guthrie LC, Alba MI. Clinically important changes in individual and composite measures of rheumatoid arthritis activity: thresholds applicable in clinical trials. Annals of the Rheumatic Diseases. 2015;74:1691-6.
64. Ward MM. Clinical measures in rheumatoid arthritis: which are most useful in assessing patients? J Rheumatol. 1994;21:17-27.

65. Wells G, Li T, Maxwell L, Maclean R, Tugwell P. Responsiveness of patient reported outcomes including fatigue, sleep quality, activity limitation, and quality of life following treatment with abatacept for rheumatoid arthritis. Annals of the Rheumatic Diseases. 2008;67:260-5.
66. Wolfe F, Michaud K, Strand V. Expanding the definition of clinical differences: from minimally clinically important differences to really important differences. Analyses in 8931 patients with rheumatoid arthritis. Journal of Rheumatology. 2005;32:583-9.
67. Callahan LF, McCoy A, Smith W. Comparison and sensitivity to change of self-report scales to assess difficulty, dissatisfaction, and pain in performing activities of daily living over one and five years in rheumatoid arthritis. Arthritis Care & Research. 1992;5:137-45.
68. Hagen KB, Smedstad LM, Uhlig T, Kvien TK. The responsiveness of health status measures in patients with rheumatoid arthritis: comparison of disease-specific and generic instruments. Journal of Rheumatology. 1999;26:1474-80.
69. Kvamme MK, Kristiansen IS, Lie E, Kvien TK. Identification of cutpoints for acceptable health status and important improvement in patient-reported outcomes, in rheumatoid arthritis, psoriatic arthritis, and ankylosing spondylitis. Journal of Rheumatology. 2010;37:26-31.
70. Martin M, Kosinski M, Bjorner JB, Ware Jr JE, MacLean R, Li T. Item response theory methods can improve the measurement of physical function by combining the modified health assessment questionnaire and the SF-36 physical function scale. Quality of Life Research. 2007;16:647-60.
71. Nagasawa H, Kameda H, Sekiguchi N, Amano K, Takeuchi T. Differences between the Health Assessment Questionnaire Disability Index (HAQ-DI) and the modified HAQ (mHAQ) score before and after infliximab treatment in patients with rheumatoid arthritis. Modern Rheumatology. 2010;20:337-42.
72. Russell AS, Conner-Spady B, Mintz A, Mallon C, Maksymowych WP. The responsiveness of generic health status measures as assessed in patients with rheumatoid arthritis receiving infliximab. Journal of Rheumatology. 2003;30:941-7.
73. Stucki G, Stucki S, Bruhlmann P, Michel BA. Ceiling effects of the Health Assessment Questionnaire and its modified version in some ambulatory rheumatoid arthritis patients. Annals of the Rheumatic Diseases. 1995;54:461-5.
74. Tugwell P, Wells G, Strand V, Maetzel A, Bombardier C, Crawford B, et al. Clinical improvement as reflected in measures of function and health- related quality of life following treatment with leflunomide compared with methotrexate in patients with rheumatoid arthritis: Sensitivity and relative efficiency to detect a treatment effect in a twelve-month, placebo-controlled trial. Arthritis and Rheumatism. 2000;43:506-14.
75. Uhlig T, Haavardsholm EA, Kvien TK. Comparison of the Health Assessment Questionnaire (HAQ) and the modified HAQ (MHAQ) in patients with rheumatoid arthritis. Rheumatology. 2006;45:454-8.
76. Wolfe F. Which HAQ is best? A comparison of the HAQ, MHAQ and RA-HAQ, a difficult 8 item HAQ (DHAQ), and a rescored 20 item HAQ (HAQ20): analyses in 2,491 rheumatoid arthritis patients following leflunomide initiation. Journal of Rheumatology. 2001;28:982-9.
77. Ziebland S, Fitzpatrick R, Jenkinson C, Mowat A, Mowat A. Comparison of two approaches to measuring change in health status in rheumatoid arthritis: the Health Assessment Questionnaire (HAQ) and modified HAQ. Annals of the Rheumatic Diseases. 1992;51:1202-5.
78. Fries J, Rose M, Krishnan E. The PROMIS of better outcome assessment: responsiveness, floor and ceiling effects, and Internet administration. Journal of Rheumatology. 2011;38:1759-64.
79. Fries JF, Krishnan E, Rose M, Lingala B, Bruce B. Improved responsiveness and reduced sample size requirements of PROMIS physical function scales with item response theory. Arthritis Research & Therapy. 2011;13:R147.

80. Hays RD, Spritzer KL, Fries JF, Krishnan E. Responsiveness and minimally important difference for the patient-reported outcomes measurement information system (PROMIS) 20-item physical functioning short form in a prospective observational study of rheumatoid arthritis. Annals of the Rheumatic Diseases. 2015;74:104-7.

81. Oude Voshaar MAH, ten Klooster PM, Glas CAW, Vonkeman HE, Taal E, Krishnan E, et al. Validity and measurement precision of the PROMIS physical function item bank and a content validity-driven 20-item short form in rheumatoid arthritis compared with traditional measures. Rheumatology. 2015;54:2221-9.

82. Schalet BD, Hays RD, Jensen SE, Beaumont JL, Fries JF, Cella D. Validity of PROMIS physical function measured in diverse clinical samples. Journal of Clinical Epidemiology. 2016;73:112-8.