Supplementary file 5 – Supplementary Tables 1 to 4

Supplementary Table 1. Evidence based on test content

Number of instances of evidence based on test content across all studies		
Method to generate evidence		
Literature review	4	8%
Existing measures of the construct	8	15%
Expert review	14	27%
Participant involvement:		
Concept mapping	3	6%
Interviews	2	4%
Participant feedback processes about items	4	8%
Construct descriptions (e.g., high/low)	4	8%
Item intent descriptions	1	2%
Examination of administration methods	3	6%
Other method (e.g., item difficulty):		
Item difficulty	5	10%
Items tested against item intents	1	2%
IRT analysis for item selection within domains	1	2%
Item selection based on hospital medical texts	1	2%
Item selection based on HL conceptual model	1	2%
Total instances of evidence based on test content	*52*	*100%*

Supplementary Table 2. Evidence based on response processes

Number of instances of evidence based on response processes across all studies		
Method to generate evidence		
With respondents:		
Cognitive interviews	3	43%
Recording and timing responses to items	3	43%
With users:		
Cognitive interviews	1	14%
Total instances of evidence based on response processes	*7*	*100%*

Supplementary Table 3. Evidence based on internal structure

Number of instances of evidence based on internal structure across all studies		
Method to generate evidence		
Exploratory factor analysis (incl. PCA*)	7	25%
Confirmatory factory analysis (incl. IRT** item discriminations)	7	25%
Multi-group factor analysis	1	4%
Correlation patterns / multi-trait scaling analysis:		
Tetrachoric correlations	1	4%

Authors: Hawkins M; Elsworth GR; Hoban E; Osborne RH. (2019)

Inter-item correlations	1	4%
Item-total correlations	1	4%
Item-remainder correlations	2	7%
Differential item functioning	3	11%
Other method:		
Very Simple Structure	1	4%
Velicer's Minimum Average partial criterion	1	4%
Rasch analysis (overall fit, individual person/item fit)	1	4%
Intra-factor correlations	1	4%
IRT for item discriminations	1	4%
Total instances of evidence based on response processes	*28*	*100%*

*PCA = principal component analysis; **IRT = item response theory

Supplementary Table 4. Evidence based on relations to other variables

Summary of number of instances of evidence based on relations to other variables across all studies		
Type of evidence		
Convergent evidence	57	53%
Discriminant evidence	3	3%
Criterion-referenced evidence	17	16%
Evidence for group differences	30	28%
Evidence for generalisation	0	0%
Total instances of evidence based on relations to other variables	107	100%
Number of instances of evidence based on relations to other variables across all studies		
Convergent evidence (relationships between items and scales of the same or similar structure) (n=38 studies):		
Spearman's correlation coefficient	11	19%
Pearson correlation coefficient	11	19%
Linear regression models	5	9%
Logistic regression models	2	4%
Receiver operating characteristic / Area under the ROC (AUROC)	11	19%
Wilcoxen signed rank test	2	4%
Cross tabulations / calculated agreement and disagreement	2	4%
Goodman-Kruskal gamma correlation	1	2%
Bland-Altman plots	1	2%
Cohen's Kappa	1	2%
Sensitivity and specificity	1	2%
Stratum-specific likelihood ratios	1	2%
Unnamed / unclear correlation calculations with similar measures	8	14%
Total instances of convergent evidence	*57*	*100%*
Discriminant evidence (measures of different constructs are sufficiently uncorrelated) (n=2 studies)		
Comparison of AVE and shared variance between HLQ scales	1	33%
Pearson correlation coefficient	1	33%%
Multiscale factor analysis	1	33%
Total instances of discriminant evidence	*3*	*100%*

Authors: Hawkins M; Elsworth GR; Hoban E; Osborne RH. (2019)

Criterion-referenced evidence (how accurately test scores predict criterion performance) (n=9 studies):		
Spearman's correlation coefficient	2	12%
Pearson correlation coefficient	1	6%
Linear regression models	6	35%
Logistic regression models	2	12%
ROC/AUROC	1	6%
Chi-squared test of independence	3	18%
ANOVA	1	6%
Cohen's d	1	6%
Total instances of criterion-referenced evidence	*17*	*100%*
Evidence for group differences (relationships of test scores with background characteristics such as demographic information) (n=19 studies):	**N**	**%**
Linear regression models	4	13%
Logistic regression models	3	10%
Univariate associations	1	3%
Spearman's correlation coefficient	1	3%
Chi-squared test	3	10%
Analysis of variance (ANOVA)	5	17%
Analysis of covariance (ANCOVA)	1	3%
Cross tabulations	1	3%
Area under the ROC (AUROC)	1	3%
Kruskal-Wallis test	1	3%
Mann-Whitney U test	2	7%
Goodman-Kruskal gamma correlation	1	3%
Independent sample t-test	3	10%
Exploratory partial correlation analysis	1	3%
Bayesian fit statistics	1	3%
Descriptive statistics (sub-group differences)	1	3%
Total instances of evidence of group differences	*30*	*100%*
Evidence for generalisation (degree to which evidence can be generalised to a new situation) (n=0 studies):	**N**	**%**
Only research synthesis-type studies - see validity generalisation in the *Standards*.	0	0%

Authors: Hawkins M; Elsworth GR; Hoban E; Osborne RH. (2019)