APPENDIX B
Knee Methodological Quality

Red type = Author's comments and methodological quality ratings
Strikethrough indicates that the grading criteria are not applicable

Battaglia/Am J Sports Med/2007

STEP1: Evaluated measurement properties in the article

Check if present	Property	Location
	Reliability	Box A
	Agreement/Measurement Error	Box B
✔	Hypothesis Testing / Construct validity	Box C
	Criterion/Predictive Validity	Box D
	Responsiveness	Box E

Step 2. Determining if a study meets the standards for good methodological quality

C. Hypothesis Testing/Construct Validity N=102

Design requirements	Excellent	Good	Fair	Poor
Was the percentage of missing items given? none	Percentage of missing items described	Percentage of missing items NOT described	NA	NA
Was there a description of how missing items were handled? none	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
Was the sample size included in the analysis adequate?	Adequate sample size (≥100 per analysis)	Good sample size (50-99 per analysis)	Moderate sample size (30-49 per analysis)	Small sample size (<30 per analysis)
Were hypotheses regarding correlations or mean differences formulated a priori (i.e. before data collection)?	hypotheses formulated a priori	Minimal number of hypotheses formulate a priori	Hypotheses vague or not formulated but possible to deduce what was expected	Unclear what was expected
Was the expected *direction* of correlations or mean differences included in the hypotheses?	Expected direction of the correlations or differences stated	Expected direction of the correlations or differences NOT stated –		
Was the expected	Expected	Expected		

absolute or relative *magnitude* of correlations or mean differences included in the hypotheses?	magnitude of the correlations or differences stated	magnitude of the correlations or differences NOT stated		
for convergent validity: Was an adequate description provided of the comparator instrument(s)?	Adequate description of the constructs measured by the comparator instrument(s)	Adequate description of most of the constructs measured by the comparator instrument(s)	Poor description of the constructs measured by the comparator instrument(s)	NO description of the constructs measured by the comparator instrument(s)
for convergent validity: Were the measurement properties of the comparator instrument(s) adequately described?	Adequate measurement properties of the comparator instrument(s) in a population similar to the study population	Adequate measurement properties of the comparator instrument(s) but not sure if these apply to the study population	Some information on measurement properties (or a reference to a study on measurement properties) of the comparator instrument(s) in any study population	No information on the measurement properties of the comparator instrument(s)
Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study (e.g. only data presented on a comparison with an instrument that measures another construct)	Other important methodological flaws in the design or execution of the study
Were design and statistical methods adequate for the hypotheses to be tested?	Statistical methods applied appropriate	Assumable that statistical methods were appropriate, e.g. Pearson correlations applied, but distribution of scores or mean (SD) not presented	Statistical methods applied NOT optimal	Statistical methods applied NOT appropriate

Bjorklund/Knee Surg Sports Traumatol/2006

STEP1: Evaluated measurement properties in the article

Check if present	Property	Location
✔	Reliability	Box A
	Agreement/Measurement Error	Box B
✔	Hypothesis Testing / Construct	Box C

	validity	
	Criterion/Predictive Validity	Box D

Step 2. Determining if a study meets the standards for good methodological quality
A. Reliability

Design requirements	Excellent	Good	Fair	Poor
Was the percentage of missing subjects given? none	Percentage of missing subjects described	Percentage of missing subjects NOT described		
Was there a description of how missing subjects were handled? none	Described how missing subjects were handled	Not described but it can be deduced how missing subjects were handled	Not clear how missing subjects were handled	
Was the sample size included in the analysis adequate?	Adequate sample size (≥100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30)
Were at least two measurements available? Intra and inter intra only	At least two measurements			Only one measurement
Were the administrations independent? Intra	Independent measurements	Assumable that the measurements were independent	Doubtful whether the measurements were independent	Measurements NOT independent
Was the time interval stated? Intra	Time interval stated		Time interval NOT stated	
Were patients stable in the interim period on the construct to be measured? Intra	Patients were stable (evidence provided)	Assumable that patients were stable	Unclear if patients were stable	Patients were NOT stable
Was the time interval appropriate? Intra	Time interval appropriate		Doubtful whether time interval was appropriate	Time interval NOT appropriate
Were the test conditions similar for both measurements? e.g. type of administration, environment, instructions	Test conditions were similar (evidence provided)	Assumable that test conditions were similar	Unclear if test conditions were similar	Test conditions were NOT similar
Were there any important flaws in the design or methods of the study? Intra	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study
~~for continuous scores: Was an intraclass correlation coefficient (ICC) calculated?~~	~~ICC calculated and model or formula of the ICC is~~	~~ICC calculated but model or formula of the ICC not~~	~~Pearson or Spearman correlation coefficient~~	~~No ICC or Pearson or Spearman correlations~~

	~~described~~	~~described or not optimal. Pearson or Spearman correlation coefficient calculated with evidence provided that no systematic change has occurred~~	~~calculated WITHOUT evidence provided that no systematic change has occurred or WITH evidence that systematic change has occurred~~	~~calculated~~
For dichotomous/nominal/ordinal scores: Was kappa calculated?	Kappa calculated			Only percentage agreement calculated
discriminantfor ordinal scores: Was a weighted kappa calculated?	Weighted Kappa calculated		Unweighted Kappa calculated	Only percentage agreement calculated
for ordinal scores: Was the weighting scheme described? e.g. linear, quadratic	Weighting scheme described	Weighting scheme NOT described		

C. Hypothesis Testing / Construct Validity

Design requirements	Excellent	Good	Fair	Poor
Was the percentage of missing items given? none	Percentage of missing items described	Percentage of missing items NOT described	NA	NA
Was there a description of how missing items were handled? none	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
Was the sample size included in the analysis adequate?	Adequate sample size (≥100 per analysis)	Good sample size (50-99 per analysis)	Moderate sample size (30-49 per analysis)	Small sample size (<30 per analysis)
Were hypotheses regarding correlations or mean differences formulated a priori (i.e. before data collection)?	hypotheses formulated a priori	Minimal number of hypotheses formulate a priori	Hypotheses vague or not formulated but possible to deduce what was expected	Unclear what was expected
Was the expected *direction* of correlations or	Expected direction of the correlations or	Expected direction of the correlations or		

mean differences included in the hypotheses?	differences stated	differences NOT stated –		
Was the expected absolute or relative *magnitude* of correlations or mean differences included in the hypotheses?	Expected magnitude of the correlations or differences stated	Expected magnitude of the correlations or differences NOT stated		
for convergent validity: Was an adequate description provided of the comparator instrument(s)?	Adequate description of the constructs measured by the comparator instrument(s)	Adequate description of most of the constructs measured by the comparator instrument(s)	Poor description of the constructs measured by the comparator instrument(s)	NO description of the constructs measured by the comparator instrument(s)
for convergent validity: Were the measurement properties of the comparator instrument(s) adequately described?	Adequate measurement properties of the comparator instrument(s) in a population similar to the study population	Adequate measurement properties of the comparator instrument(s) but not sure if these apply to the study population	Some information on measurement properties (or a reference to a study on measurement properties) of the comparator instrument(s) in any study population	No information on the measurement properties of the comparator instrument(s) NOTE: First time tested in this study
Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study (e.g. only data presented on a comparison with an instrument that measures another construct)	Other important methodological flaws in the design or execution of the study
Were design and statistical methods adequate for the hypotheses to be tested?	Statistical methods applied appropriate	Assumable that statistical methods were appropriate, e.g. Pearson correlations applied, but distribution of scores or mean (SD) not presented	Statistical methods applied NOT optimal	Statistical methods applied NOT appropriate

Bjorklund/Knee Surg Sports Traumatol/2009

STEP1: Evaluated measurement properties in the article

Check if present	Property	Location
	Reliability	Box A
	Agreement/Measurement Error	Box B
✔	Hypothesis Testing / Construct validity	Box C
	Criterion/Predictive Validity	Box D
✔	Responsiveness	Box E
	Predictive/Criterion validity	Box F

Step 2. Determining if a study meets the standards for good methodological quality
C. Hypothesis Testing / Construct Validity

Design requirements	Excellent	Good	Fair	Poor
Was the percentage of missing items given? none	Percentage of missing items described	Percentage of missing items NOT described	NA	NA
Was there a description of how missing items were handled? none	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
Was the sample size included in the analysis adequate?	Adequate sample size (≥100 per analysis)	Good sample size (50-99 per analysis)	Moderate sample size (30-49 per analysis)	Small sample size (<30 per analysis)
Were hypotheses regarding correlations or mean differences formulated a priori (i.e. before data collection)?	hypotheses formulated a priori	Minimal number of hypotheses formulate a priori	Hypotheses vague or not formulated but possible to deduce what was expected	Unclear what was expected
Was the expected *direction* of correlations or mean differences included in the hypotheses?	Expected direction of the correlations or differences stated	Expected direction of the correlations or differences NOT stated –		
Was the expected absolute or relative *magnitude* of correlations or mean differences included in the hypotheses?	Expected magnitude of the correlations or differences stated	Expected magnitude of the correlations or differences NOT stated		
for convergent validity: Was an	Adequate description of the	Adequate description of	Poor description of the constructs	NO description of the constructs

adequate description provided of the comparator instrument(s)?	constructs measured by the comparator instrument(s)	most of the constructs measured by the comparator instrument(s)	measured by the comparator instrument(s)	measured by the comparator instrument(s)
for convergent validity: Were the measurement properties of the comparator instrument(s) adequately described?	Adequate measurement properties of the comparator instrument(s) in a population similar to the study population IKDC, SF-36	Adequate measurement properties of the comparator instrument(s) but not sure if these apply to the study population	Some information on measurement properties (or a reference to a study on measurement properties) of the comparator instrument(s) in any study population	No information on the measurement properties of the comparator instrument(s) NOTE: First time tested in this study
Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study (e.g. only data presented on a comparison with an instrument that measures another construct)	Other important methodological flaws in the design or execution of the study
Were design and statistical methods adequate for the hypotheses to be tested?	Statistical methods applied appropriate	Assumable that statistical methods were appropriate, e.g. Pearson correlations applied, but distribution of scores or mean (SD) not presented	Statistical methods applied NOT optimal	Statistical methods applied NOT appropriate

E. Responsiveness

Design requirements	Excellent	Good	Fair	Poor
Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
Was the sample size included in the analysis adequate?	Adequate sample size (≥100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30)
Was a longitudinal design with at	Longitudinal design used			No longitudinal design used

least two measurement used?				
Was the time interval stated?	Time interval adequately described			Time interval NOT described
If anything occurred in the interim period (e.g. intervention, other relevant events), was it adequately described?	Anything that occurred during the interim period (e.g. treatment) adequately described	Assumable what occurred during the interim period	Unclear or NOT described what occurred during the interim period	If anything occurred in the interim period (e.g. intervention, other relevant events), was it adequately described?
Was a proportion of the patients changed (i.e. improvement or deterioration)?	Part of the patients were changed (evidence provided)	NO evidence provided, but assumable that part of the patients were changed	Unclear if part of the patients were changed	Patients were NOT changed
For constructs for which a gold standard was not available:				
~~Were hypotheses about changes in scores formulated a priori (i.e. before data collection)?~~	~~Hypotheses formulated a priori~~		~~Hypotheses vague or not formulated but possible to deduce what was expected~~	~~Unclear what was expected~~
~~Was the expected *direction* of correlations or mean differences of the change scores of HR-PRO instruments included in these hypotheses?~~	~~Expected direction of the correlations or differences stated~~	~~Expected direction of the correlations or differences NOT stated~~		
~~Were the expected absolute or relative *magnitude* of correlations or mean differences of the change scores of HR-PRO instruments included in these hypotheses?~~	~~Expected magnitude of the correlations or differences stated~~	~~Expected magnitude of the correlations or differences NOT stated~~		
~~Was an adequate description provided of the comparator instrument(s)?~~	~~Adequate description of the constructs measured by the comparator instrument(s)~~		~~Poor description of the constructs measured by the comparator instrument(s)~~	~~NO description of the constructs measured by the comparator instrument(s)~~
~~Were the measurement~~	~~Adequate measurement~~	~~Adequate measurement~~	~~Some information on measurement~~	~~NO information on the~~

~~properties of the comparator instrument(s) adequately described?~~	~~properties of the comparator instrument(s) in a population similar to the study population~~	~~properties of the comparator instrument(s) but not sure if these apply to the study population~~	~~properties (or a reference to a study on measurement properties) of the comparator instrument(s) in any study population~~	~~measurement properties of the comparator instrument(s)~~
~~Were there any important flaws in the design or methods of the study?~~	~~No other important methodological flaws in the design or execution of the study~~		~~Other minor methodological flaws in the design or execution of the study (e.g. only data presented on a comparison with an instrument that measures another construct)~~	~~Other important methodological flaws in the design or execution of the study~~
~~Were design and statistical methods adequate for the hypotheses to be tested?~~	~~Statistical methods applied appropriate~~		~~Statistical methods applied NOT optimal~~	~~Statistical methods applied NOT appropriate~~
For constructs for which a gold standard was available:				
Can the criterion for change be considered as a reasonable gold standard?	Criterion used can be considered an adequate 'gold standard' (evidence provided)	No evidence provided, but assumable that the criterion used can be considered an adequate 'gold standard'	Unclear whether the criterion used can be considered an adequate 'gold standard'	Criterion used can NOT be considered an adequate 'gold standard'
Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study
for continuous scores: Were correlations between change scores, or the area under the Receiver Operator Curve (ROC) curve calculated?	Correlations or Area under the ROC Curve (AUC) calculated			Correlations or AUC NOT calculated
~~for dichotomous scales: Were sensitivity and specificity (changed versus~~	~~Sensitivity and specificity calculated~~			~~Sensitivity and specificity NOT calculated~~

~~not changed)~~ ~~determined?~~				

Carter/Br J Sports Med/ 1997

STEP1: Evaluated measurement properties in the article

Check if present	Property	Location
	Reliability	Box A
	Agreement/Measurement Error	Box B
✔	Hypothesis Testing / Construct validity	Box C
	Criterion/Predictive Validity	Box D
✔	Responsiveness	Box E
	Predictive validity/Criterion/Predictive Validity	Box F

Step 2. Determining if a study meets the standards for good methodological quality
C. Hypothesis Testing / Construct Validity

Design requirements	Excellent	Good	Fair	Poor
Was the percentage of missing items given? none	Percentage of missing items described	Percentage of missing items NOT described	NA	NA
Was there a description of how missing items were handled? none	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
Was the sample size included in the analysis adequate?	Adequate sample size (≥100 per analysis)	Good sample size (50-99 per analysis) 50 + 23 controls	Moderate sample size (30-49 per analysis)	Small sample size (<30 per analysis)
Were hypotheses regarding correlations or mean differences formulated a priori (i.e. before data collection)?	hypotheses formulated a priori	Minimal number of hypotheses formulate a priori	Hypotheses vague or not formulated but possible to deduce what was expected	Unclear what was expected
Was the expected *direction* of correlations or mean differences included in the hypotheses?	Expected direction of the correlations or differences stated	Expected direction of the correlations or differences NOT stated –		
Was the expected absolute or	Expected magnitude of the	Expected magnitude of the		

relative *magnitude* of correlations or mean differences included in the hypotheses?	correlations or differences stated	correlations or differences NOT stated		
for convergent validity: Was an adequate description provided of the comparator instrument(s)?	Adequate description of the constructs measured by the comparator instrument(s)	Adequate description of most of the constructs measured by the comparator instrument(s)	Poor description of the constructs measured by the comparator instrument(s)	NO description of the constructs measured by the comparator instrument(s)
for convergent validity: Were the measurement properties of the comparator instrument(s) adequately described?	Adequate measurement properties of the comparator instrument(s) in a population similar to the study population IKDC, SF-36	Adequate measurement properties of the comparator instrument(s) but not sure if these apply to the study population	Some information on measurement properties (or a reference to a study on measurement properties) of the comparator instrument(s) in any study population	No information on the measurement properties of the comparator instrument(s)
Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study (e.g. only data presented on a comparison with an instrument that measures another construct)	Other important methodological flaws in the design or execution of the study
Were design and statistical methods adequate for the hypotheses to be tested?	Statistical methods applied appropriate	Assumable that statistical methods were appropriate, e.g. Pearson correlations applied, but distribution of scores or mean (SD) not presented	Statistical methods applied NOT optimal	Statistical methods applied NOT appropriate

E. Responsiveness

Design requirements	Excellent	Good	Fair	Poor
Was the percentage of missing items given? none	Percentage of missing items described	Percentage of missing items NOT described		
Was there a description of how	Described how missing items	Not described but it can be	Not clear how missing items were	

missing items were handled? none	were handled	deduced how missing items were handled	handled	
Was the sample size included in the analysis adequate?	Adequate sample size (≥100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30)
Was a longitudinal design with at least two measurement used?	Longitudinal design used			No longitudinal design used
Was the time interval stated?	Time interval adequately described			Time interval NOT described
If anything occurred in the interim period (e.g. intervention, other relevant events), was it adequately described?	Anything that occurred during the interim period (e.g. treatment) adequately described	Assumable what occurred during the interim period	Unclear or NOT described what occurred during the interim period	If anything occurred in the interim period (e.g. intervention, other relevant events), was it adequately described?
Was a proportion of the patients changed (i.e. improvement or deterioration)?	Part of the patients were changed (evidence provided)	NO evidence provided, but assumable that part of the patients were changed	Unclear if part of the patients were changed	Patients were NOT changed
For constructs for which a gold standard was not available:				
~~Were hypotheses about changes in scores formulated a priori (i.e. before data collection)?~~	~~Hypotheses formulated a priori~~		~~Hypotheses vague or not formulated but possible to deduce what was expected~~	~~Unclear what was expected~~
~~Was the expected *direction* of correlations or mean differences of the change scores of HR-PRO instruments included in these hypotheses?~~	~~Expected direction of the correlations or differences stated~~	~~Expected direction of the correlations or differences NOT stated~~		
~~Were the expected absolute or relative *magnitude* of correlations or mean differences of the change scores of HR-PRO instruments included in these~~	~~Expected magnitude of the correlations or differences stated~~	~~Expected magnitude of the correlations or differences NOT stated~~		

~~hypotheses?~~				
~~Was an adequate description provided of the comparator instrument(s)?~~	~~Adequate description of the constructs measured by the comparator instrument(s)~~		~~Poor description of the constructs measured by the comparator instrument(s)~~	~~NO description of the constructs measured by the comparator instrument(s)~~
~~Were the measurement properties of the comparator instrument(s) adequately described?~~	~~Adequate measurement properties of the comparator instrument(s) in a population similar to the study population~~	~~Adequate measurement properties of the comparator instrument(s) but not sure if these apply to the study population~~	~~Some information on measurement properties (or a reference to a study on measurement properties) of the comparator instrument(s) in any study population~~	~~NO information on the measurement properties of the comparator instrument(s)~~
~~Were there any important flaws in the design or methods of the study?~~	~~No other important methodological flaws in the design or execution of the study~~		~~Other minor methodological flaws in the design or execution of the study (e.g. only data presented on a comparison with an instrument that measures another construct)~~	~~Other important methodological flaws in the design or execution of the study~~
~~Were design and statistical methods adequate for the hypotheses to be tested?~~	~~Statistical methods applied appropriate~~		~~Statistical methods applied NOT optimal~~	~~Statistical methods applied NOT appropriate~~
For constructs for which a gold standard was available:				
Can the criterion for change be considered as a reasonable gold standard?	Criterion used can be considered an adequate 'gold standard' (evidence provided)	No evidence provided, but assumable that the criterion used can be considered an adequate 'gold standard'	Unclear whether the criterion used can be considered an adequate 'gold standard'	Criterion used can NOT be considered an adequate 'gold standard'
Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study
for continuous scores: Were correlations between change scores, or the area	Correlations or Area under the ROC Curve (AUC) calculated			Correlations or AUC NOT calculated

under the Receiver Operator Curve (ROC) curve calculated?				
~~for dichotomous scales: Were sensitivity and specificity (changed versus not changed) determined?~~	~~Sensitivity and specificity calculated~~			~~Sensitivity and specificity NOT calculated~~

Crossley/J Orthop Res/ 2007

STEP1: Evaluated measurement properties in the article

Check if present	Property	Location
✔	Reliability	Box A
	Agreement/Measurement Error	Box B
	Hypothesis Testing / Construct validity	Box C
✔	Criterion/Predictive Validity	Box D
	Responsiveness	Box E

Step 2. Determining if a study meets the standards for good methodological quality

A. Reliability n=10

Design requirements	Excellent	Good	Fair	Poor
Was the percentage of missing subjects given? none	Percentage of missing subjects described	Percentage of missing subjects NOT described		
Was there a description of how missing subjects were handled? none	Described how missing subjects were handled	Not described but it can be deduced how missing subjects were handled	Not clear how missing subjects were handled	
Was the sample size included in the analysis adequate?	Adequate sample size (≥100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30)
Were at least two measurements available? Intra and inter intra only	At least two measurements			Only one measurement
Were the administrations independent? Intra	Independent measurements	Assumable that the measurements were independent	Doubtful whether the measurements were independent	Measurements NOT independent
Was the time interval stated? Intra	Time interval stated		Time interval NOT stated	
Were patients stable in the interim period on the construct to be measured? Intra	Patients were stable (evidence provided)	Assumable that patients were stable	Unclear if patients were stable	Patients were NOT stable
Was the time interval	Time interval		Doubtful	Time interval

appropriate? Intra	appropriate-7-10 days		whether time interval was appropriate	NOT appropriate
Were the test conditions similar for both measurements? e.g. type of administration, environment, instructions	Test conditions were similar (evidence provided)	Assumable that test conditions were similar	Unclear if test conditions were similar	Test conditions were NOT similar
Were there any important flaws in the design or methods of the study? Intra	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study
for continuous scores: Was an intraclass correlation coefficient (ICC) calculated?	ICC calculated and model or formula of the ICC is described	ICC calculated but model or formula of the ICC not described or not optimal. Pearson or Spearman correlation coefficient calculated with evidence provided that no systematic change has occurred	Pearson or Spearman correlation coefficient calculated WITHOUT evidence provided that no systematic change has occurred or WITH evidence that systematic change has occurred	No ICC or Pearson or Spearman correlations calculated
~~For dichotomous/nominal/ordinal scores: Was kappa calculated?~~	~~Kappa calculated~~			~~Only percentage agreement calculated~~
~~for ordinal scores: Was a weighted kappa calculated?~~	~~Weighted Kappa calculated~~		~~Unweighted Kappa calculated~~	~~Only percentage agreement calculated~~
~~for ordinal scores: Was the weighting scheme described? e.g. linear, quadratic~~	~~Weighting scheme described~~	~~Weighting scheme NOT described~~		

D. Criterion/Predictive Validity (Predictive in this case n= 58 but only 27 used in analysis

Design requirements	Excellent	Good	Fair	Poor
Was the percentage of missing items given? none	Percentage of missing items described	Percentage of missing items NOT described	NA	NA
Was there a description of how missing	Described how missing items were handled	Not described but it can be deduced how missing items were	Not clear how missing items were handled	

items were handled? none		handled- did not use controls in the final linear regression model		
Was the sample size included in the analysis adequate?	Adequate sample size (≥100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30) N=27
Can the criterion used or employed be considered as a reasonable 'gold standard'?	Criterion used can be considered an adequate 'gold standard' (evidence provided)	No evidence provided, but assumable that the criterion used can be considered an adequate 'gold standard'	Unclear whether the criterion used can be considered an adequate 'gold standard'	Criterion used can NOT be considered an adequate 'gold standard'
Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study	Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study	Were there any important flaws in the design or methods of the study?
for continuous scores: Were correlations, or the area under the receiver operating curve calculated?	Correlations or AUC calculated			Correlations or AUC NOT calculated
~~for dichotomous scores: Were sensitivity and specificity determined?~~	~~Sensitivity and specificity calculated~~			~~Sensitivity and specificity NOT calculated~~

Eastlack/MSSE/ 1999

STEP1: Evaluated measurement properties in the article

Check if present	Property	Location
	Reliability	Box A
	Agreement/Measurement Error	Box B
✔	Hypothesis Testing / Construct validity	Box C
✔	Criterion/Predictive Validity	Box D
	Responsiveness	Box E

Step 2. Determining if a study meets the standards for good methodological quality

C. Hypothesis Testing / Construct Validity n=45

Design requirements	Excellent	Good	Fair	Poor
Was the percentage of missing items	Percentage of missing items described	Percentage of missing items NOT described	NA	NA

given? none				
Was there a description of how missing items were handled? none	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
Was the sample size included in the analysis adequate?	Adequate sample size (≥100 per analysis)	Good sample size (50-99 per analysis)	Moderate sample size (30-49 per analysis)	Small sample size (<30 per analysis)
Were hypotheses regarding correlations or mean differences formulated a priori (i.e. before data collection)?	hypotheses formulated a priori	Minimal number of hypotheses formulate a priori	Hypotheses vague or not formulated but possible to deduce what was expected	Unclear what was expected
Was the expected *direction* of correlations or mean differences included in the hypotheses?	Expected direction of the correlations or differences stated	Expected direction of the correlations or differences NOT stated –		
Was the expected absolute or relative *magnitude* of correlations or mean differences included in the hypotheses?	Expected magnitude of the correlations or differences stated	Expected magnitude of the correlations or differences NOT stated		
for convergent validity: Was an adequate description provided of the comparator instrument(s)?	Adequate description of the constructs measured by the comparator instrument(s)	Adequate description of most of the constructs measured by the comparator instrument(s)	Poor description of the constructs measured by the comparator instrument(s)	NO description of the constructs measured by the comparator instrument(s)
for convergent validity: Were the measurement properties of the comparator instrument(s) adequately described?	Adequate measurement properties of the comparator instrument(s) in a population similar to the study population	Adequate measurement properties of the comparator instrument(s) but not sure if these apply to the study population	Some information on measurement properties (or a reference to a study on measurement properties) of the comparator instrument(s) in any study population	No information on the measurement properties of the comparator instrument(s) NOTE: First time tested in this study
Were there any important flaws	No other important		Other minor methodological	Other important methodological

in the design or methods of the study?	methodological flaws in the design or execution of the study		flaws in the design or execution of the study (e.g. only data presented on a comparison with an instrument that measures another construct)	flaws in the design or execution of the study
Were design and statistical methods adequate for the hypotheses to be tested?	Statistical methods applied appropriate	Assumable that statistical methods were appropriate, e.g. Pearson correlations applied, but distribution of scores or mean (SD) not presented	Statistical methods applied NOT optimal	Statistical methods applied NOT appropriate

D. Criterion/Predictive Validity (Predictive in this case)

Design requirements	Excellent	Good	Fair	Poor
Was the percentage of missing items given? none	Percentage of missing items described	Percentage of missing items NOT described	NA	NA
Was there a description of how missing items were handled? none	Described how missing items were handled	Not described but it can be deduced how missing items were handled- did not use controls in the final linear regression model	Not clear how missing items were handled	
Was the sample size included in the analysis adequate?	Adequate sample size (≥100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30)
Can the criterion used or employed be considered as a reasonable 'gold standard'?	Criterion used can be considered an adequate 'gold standard' (evidence provided)	No evidence provided, but assumable that the criterion used can be considered an adequate 'gold standard'	Unclear whether the criterion used can be considered an adequate 'gold standard'	Criterion used can NOT be considered an adequate 'gold standard'
Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study	Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study	Were there any important flaws in the design or methods of the study?
for continuous scores: Were	Correlations or AUC calculated –			Correlations or AUC NOT

correlations, or the area under the receiver operating curve calculated?	Multiple regression used for predictive validity			calculated
for dichotomous scores: Were sensitivity and specificity determined?	Sensitivity and specificity calculated			Sensitivity and specificity NOT calculated

Gauffin/Int J Sports Med/1990

STEP1: Evaluated measurement properties in the article

Check if present	Property	Location
	Reliability	Box A
	Agreement/Measurement Error	Box B
✔	Hypothesis Testing / Construct validity	Box C
	Criterion/Predictive Validity	Box D
	Responsiveness	Box E

Step 2. Determining if a study meets the standards for good methodological quality
C. Hypothesis Testing / Construct Validity n=30? Never said how many in ref group and no indication whether all were used in analyses

Design requirements	Excellent	Good	Fair	Poor
Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described	NA	NA
Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
Was the sample size included in the analysis adequate?	Adequate sample size (≥100 per analysis)	Good sample size (50-99 per analysis)	Moderate sample size (30-49 per analysis)	Small sample size (<30 per analysis)
Were hypotheses regarding correlations or mean differences formulated a priori (i.e. before data collection)?	hypotheses formulated a priori	Minimal number of hypotheses formulate a priori	Hypotheses vague or not formulated but possible to deduce what was expected	Unclear what was expected

Was the expected *direction* of correlations or mean differences included in the hypotheses?	Expected direction of the correlations or differences stated	Expected direction of the correlations or differences NOT stated –		
Was the expected absolute or relative *magnitude* of correlations or mean differences included in the hypotheses?	Expected magnitude of the correlations or differences stated	Expected magnitude of the correlations or differences NOT stated		
for convergent validity: Was an adequate description provided of the comparator instrument(s)?	Adequate description of the constructs measured by the comparator instrument(s)	Adequate description of most of the constructs measured by the comparator instrument(s)	Poor description of the constructs measured by the comparator instrument(s)	NO description of the constructs measured by the comparator instrument(s)
for convergent validity: Were the measurement properties of the comparator instrument(s) adequately described?	Adequate measurement properties of the comparator instrument(s) in a population similar to the study population	Adequate measurement properties of the comparator instrument(s) but not sure if these apply to the study population	Some information on measurement properties (or a reference to a study on measurement properties) of the comparator instrument(s) in any study population	No information on the measurement properties of the comparator instrument(s)
Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study (e.g. only data presented on a comparison with an instrument that measures another construct)	Other important methodological flaws in the design or execution of the study
Were design and statistical methods adequate for the hypotheses to be tested?	Statistical methods applied appropriate	Assumable that statistical methods were appropriate, e.g. Pearson correlations applied, but distribution of scores or mean (SD) not presented	Statistical methods applied NOT optimal	Statistical methods applied NOT appropriate

Holm/Clin J Sports Med/2004

STEP1: Evaluated measurement properties in the article

Check if present	Property	Location
	Reliability	Box A
	Agreement/Measurement Error	Box B
	Hypothesis Testing / Construct validity	Box C
	Criterion/Predictive Validity	Box D
✔	Responsiveness	Box E

Step 2. Determining if a study meets the standards for good methodological quality

E. Responsiveness n=27

Design requirements	Excellent	Good	Fair	Poor
Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
Was the sample size included in the analysis adequate?	Adequate sample size (≥100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30)
Was a longitudinal design with at least two measurement used?	Longitudinal design used			No longitudinal design used
Was the time interval stated?	Time interval adequately described			Time interval NOT described
If anything occurred in the interim period (e.g. intervention, other relevant events), was it adequately described?	Anything that occurred during the interim period (e.g. treatment) adequately described	Assumable what occurred during the interim period	Unclear or NOT described what occurred during the interim period	If anything occurred in the interim period (e.g. intervention, other relevant events), was it adequately described?
Was a proportion of the patients changed (i.e. improvement or deterioration)?	Part of the patients were changed (evidence provided)	NO evidence provided, but assumable that part of the patients were changed	Unclear if part of the patients were changed	Patients were NOT changed
For constructs for which a gold standard was not available:				
Were hypotheses about changes in scores formulated	Hypotheses formulated a priori		Hypotheses vague or not formulated but possible to	Unclear what was expected

a priori (i.e. before data collection)?			deduce what was expected	
Was the expected *direction* of correlations or mean differences of the change scores of HR-PRO instruments included in these hypotheses?	Expected direction of the correlations or differences stated	Expected direction of the correlations or differences NOT stated		
Were the expected absolute or relative *magnitude* of correlations or mean differences of the change scores of HR-PRO instruments included in these hypotheses?	Expected magnitude of the correlations or differences stated	Expected magnitude of the correlations or differences NOT stated		
Was an adequate description provided of the comparator instrument(s)?	Adequate description of the constructs measured by the comparator instrument(s)		Poor description of the constructs measured by the comparator instrument(s)	NO description of the constructs measured by the comparator instrument(s)
Were the measurement properties of the comparator instrument(s) adequately described?	Adequate measurement properties of the comparator instrument(s) in a population similar to the study population	Adequate measurement properties of the comparator instrument(s) but not sure if these apply to the study population	Some information on measurement properties (or a reference to a study on measurement properties) of the comparator instrument(s) in any study population	NO information on the measurement properties of the comparator instrument(s)
Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study (e.g. only data presented on a comparison with an instrument that measures another construct)	Other important methodological flaws in the design or execution of the study
Were design and statistical methods adequate for the hypotheses to be tested?	Statistical methods applied appropriate		Statistical methods applied NOT optimal	Statistical methods applied NOT appropriate
For constructs for which a gold standard was available:				

~~Can the criterion for change be considered as a reasonable gold standard?~~	~~Criterion used can be considered an adequate 'gold standard' (evidence provided)~~	~~No evidence provided, but assumable that the criterion used can be considered an adequate 'gold standard'~~	~~Unclear whether the criterion used can be considered an adequate 'gold standard'~~	~~Criterion used can NOT be considered an adequate 'gold standard'~~
~~Were there any important flaws in the design or methods of the study?~~	~~No other important methodological flaws in the design or execution of the study~~		~~Other minor methodological flaws in the design or execution of the study~~	~~Other important methodological flaws in the design or execution of the study~~
~~for continuous scores: Were correlations between change scores, or the area under the Receiver Operator Curve (ROC) curve calculated?~~	~~Correlations or Area under the ROC Curve (AUC) calculated~~			~~Correlations or AUC NOT calculated~~
~~for dichotomous scales: Were sensitivity and specificity (changed versus not changed) determined?~~	~~Sensitivity and specificity calculated~~			~~Sensitivity and specificity NOT calculated~~

Hurd/Am J Sports Med/2008

STEP1: Evaluated measurement properties in the article

Check if present	Property	Location
	Reliability	Box A
	Agreement/Measurement Error	Box B
	Hypothesis Testing / Construct validity	Box C
✔	Criterion/Predictive Validity	Box D
	Responsiveness	Box E

Step 2. Determining if a study meets the standards for good methodological quality
D. Criterion/Predictive Validity (Predictive in this case) N= 345

Design requirements	Excellent	Good	Fair	Poor
Was the percentage of missing items given? none	Percentage of missing items described	Percentage of missing items NOT described	NA	NA
Was there a description of	Described how missing items were	Not described but it can be deduced how	Not clear how missing items	

how missing items were handled? none	handled	missing items were handled- did not use controls in the final linear regression model	were handled	
Was the sample size included in the analysis adequate?	Adequate sample size (≥100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30)
Can the criterion used or employed be considered as a reasonable 'gold standard'?	Criterion used can be considered an adequate 'gold standard' (evidence provided)	No evidence provided, but assumable that the criterion used can be considered an adequate 'gold standard'	Unclear whether the criterion used can be considered an adequate 'gold standard'	Criterion used can NOT be considered an adequate 'gold standard'
Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study	Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study	Were there any important flaws in the design or methods of the study?
for continuous scores: Were correlations, or the area under the receiver operating curve calculated?	Correlations or AUC calculated – since it was predictive validity, a hierarchical regression method was used			Correlations or AUC NOT calculated
for dichotomous scores: Were sensitivity and specificity determined?	Sensitivity and specificity calculated			Sensitivity and specificity NOT calculated

Koutras/Int J Sports Med/2009

STEP1: Evaluated measurement properties in the article

Check if present	Property	Location
	Reliability	Box A
	Agreement/Measurement Error	Box B
	Hypothesis Testing / Construct validity	**Box C**
	Criterion/Predictive Validity	Box D
✔	Responsiveness	Box E

Step 2. Determining if a study meets the standards for good methodological quality

E. Responsiveness n=20

Design requirements	Excellent	Good	Fair	Poor
Was the percentage of	Percentage of missing items	Percentage of missing items		

missing items given?	described	NOT described		
Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
Was the sample size included in the analysis adequate?	Adequate sample size (≥100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30)
Was a longitudinal design with at least two measurement used?	Longitudinal design used			No longitudinal design used
Was the time interval stated?	Time interval adequately described			Time interval NOT described
If anything occurred in the interim period (e.g. intervention, other relevant events), was it adequately described?	Anything that occurred during the interim period (e.g. treatment) adequately described	Assumable what occurred during the interim period	Unclear or NOT described what occurred during the interim period	If anything occurred in the interim period (e.g. intervention, other relevant events), was it adequately described?
Was a proportion of the patients changed (i.e. improvement or deterioration)?	Part of the patients were changed (evidence provided)	NO evidence provided, but assumable that part of the patients were changed	Unclear if part of the patients were changed	Patients were NOT changed
For constructs for which a gold standard was not available:				
~~Were hypotheses about changes in scores formulated a priori (i.e. before data collection)?~~	~~Hypotheses formulated a priori~~		~~Hypotheses vague or not formulated but possible to deduce what was expected~~	~~Unclear what was expected~~
~~Was the expected *direction* of correlations or mean differences of the change scores of HR-PRO instruments included in these hypotheses?~~	~~Expected direction of the correlations or differences stated~~	~~Expected direction of the correlations or differences NOT stated~~		
~~Were the expected absolute or relative *magnitude* of correlations or mean differences~~	~~Expected magnitude of the correlations or differences stated~~	~~Expected magnitude of the correlations or differences NOT stated~~		

~~of the change scores of HR-PRO instruments included in these hypotheses?~~				
~~Was an adequate description provided of the comparator instrument(s)?~~	~~Adequate description of the constructs measured by the comparator instrument(s)~~		~~Poor description of the constructs measured by the comparator instrument(s)~~	~~NO description of the constructs measured by the comparator instrument(s)~~
~~Were the measurement properties of the comparator instrument(s) adequately described?~~	~~Adequate measurement properties of the comparator instrument(s) in a population similar to the study population~~	~~Adequate measurement properties of the comparator instrument(s) but not sure if these apply to the study population~~	~~Some information on measurement properties (or a reference to a study on measurement properties) of the comparator instrument(s) in any study population~~	~~NO information on the measurement properties of the comparator instrument(s)~~
~~Were there any important flaws in the design or methods of the study?~~	~~No other important methodological flaws in the design or execution of the study~~		~~Other minor methodological flaws in the design or execution of the study (e.g. only data presented on a comparison with an instrument that measures another construct)~~	~~Other important methodological flaws in the design or execution of the study~~
~~Were design and statistical methods adequate for the hypotheses to be tested?~~	~~Statistical methods applied appropriate~~		~~Statistical methods applied NOT optimal~~	~~Statistical methods applied NOT appropriate~~
For constructs for which a gold standard was available:				
Can the criterion for change be considered as a reasonable gold standard?	Criterion used can be considered an adequate 'gold standard' (evidence provided)	No evidence provided, but assumable that the criterion used can be considered an adequate 'gold standard'	Unclear whether the criterion used can be considered an adequate 'gold standard'	Criterion used can NOT be considered an adequate 'gold standard'
Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study
for continuous	Correlations or			Correlations or

scores: Were correlations between change scores, or the area under the Receiver Operator Curve (ROC) curve calculated?	Area under the ROC Curve (AUC) calculated			AUC NOT calculated
~~for dichotomous scales: Were sensitivity and specificity (changed versus not changed) determined?~~	~~Sensitivity and specificity calculated~~			~~Sensitivity and specificity NOT calculated~~

Myer/JOSPT/2011

STEP1: Evaluated measurement properties in the article

Check if present	Property	Location
	Reliability	Box A
	Agreement/Measurement Error	Box B
✔	Hypothesis Testing / Construct validity	Box C
	Criterion/Predictive Validity	Box D
	Responsiveness	Box E

Step 2. Determining if a study meets the standards for good methodological quality

C. Hypothesis Testing / Construct **Validity N= 18**

Design requirements	Excellent	Good	Fair	Poor
Was the percentage of missing items given? none	Percentage of missing items described	Percentage of missing items NOT described	NA	NA
Was there a description of how missing items were handled? unsure	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
Was the sample size included in the analysis adequate?	Adequate sample size (≥100 per analysis)	Good sample size (50-99 per analysis)	Moderate sample size (30-49 per analysis)	Small sample size (<30 per analysis)
Were hypotheses regarding correlations or mean differences	hypotheses formulated a priori	Minimal number of hypotheses formulate a priori	Hypotheses vague or not formulated but possible to deduce what was	Unclear what was expected

formulated a priori (i.e. before data collection)?			expected	
Was the expected *direction* of correlations or mean differences included in the hypotheses?	Expected direction of the correlations or differences stated	Expected direction of the correlations or differences NOT stated –		
Was the expected absolute or relative *magnitude* of correlations or mean differences included in the hypotheses?	Expected magnitude of the correlations or differences stated	Expected magnitude of the correlations or differences NOT stated		
~~for convergent validity: Was an adequate description provided of the comparator instrument(s)?~~	~~Adequate description of the constructs measured by the comparator instrument(s)~~	~~Adequate description of most of the constructs measured by the comparator instrument(s)~~	~~Poor description of the constructs measured by the comparator instrument(s)~~	~~NO description of the constructs measured by the comparator instrument(s)~~
~~for convergent validity: Were the measurement properties of the comparator instrument(s) adequately described?~~	~~Adequate measurement properties of the comparator instrument(s) in a population similar to the study population~~	~~Adequate measurement properties of the comparator instrument(s) but not sure if these apply to the study population~~	~~Some information on measurement properties (or a reference to a study on measurement properties) of the comparator instrument(s) in any study population~~	~~No information on the measurement properties of the comparator instrument(s)~~
Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study (e.g. only data presented on a comparison with an instrument that measures another construct)	Other important methodological flaws in the design or execution of the study
Were design and statistical methods adequate for the hypotheses to be tested?	Statistical methods applied appropriate	Assumable that statistical methods were appropriate, e.g. Pearson correlations applied, but distribution of scores or mean (SD) not	Statistical methods applied NOT optimal	Statistical methods applied NOT appropriate

		presented		

Nagano/ Open Sports Medicine Journal/2010

STEP1: Evaluated measurement properties in the article

Check if present	Property	Location
✔	Reliability	Box A
	Agreement/Measurement Error	Box B
	Hypothesis Testing / Construct validity	Box C
✔	Criterion/Predictive Validity	Box D
	Responsiveness	Box E

Step 2. Determining if a study meets the standards for good methodological quality

A. Reliability n=14 subjects

Design requirements	Excellent	Good	Fair	Poor
Was the percentage of missing subjects given? Not addressed	Percentage of missing subjects described	Percentage of missing subjects NOT described		
Was there a description of how missing subjects were handled?	Described how missing subjects were handled	Not described but it can be deduced how missing subjects were handled	Not clear how missing subjects were handled	
Was the sample size included in the analysis adequate?	Adequate sample size (≥100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30)
Were at least two measurements available? Intra and inter intra only	At least two measurements			Only one measurement
Were the administrations independent? Intra	Independent measurements	Assumable that the measurements were independent	Doubtful whether the measurements were independent	Measurements NOT independent
Was the time interval stated? Intra	Time interval stated		Time interval NOT stated	
Were patients stable in the interim period on the construct to be measured? Intra	Patients were stable (evidence provided)	Assumable that patients were stable	Unclear if patients were stable	Patients were NOT stable
Was the time interval appropriate? Intra	Time interval appropriate		Doubtful whether time interval was appropriate	Time interval NOT appropriate
Were the test conditions similar for both measurements? e.g. type of administration, environment, instructions	Test conditions were similar (evidence provided)	Assumable that test conditions were similar	Unclear if test conditions were similar	Test conditions were NOT similar
Were there any important	No other		Other minor	Other

flaws in the design or methods of the study? Intra	important methodological flaws in the design or execution of the study		methodological flaws in the design or execution of the study	important methodological flaws in the design or execution of the study
for continuous scores: Was an intraclass correlation coefficient (ICC) calculated?	ICC calculated and model or formula of the ICC is described	ICC calculated but model or formula of the ICC not described or not optimal. Pearson or Spearman correlation coefficient calculated with evidence provided that no systematic change has occurred	Pearson or Spearman correlation coefficient calculated WITHOUT evidence provided that no systematic change has occurred or WITH evidence that systematic change has occurred	No ICC or Pearson or Spearman correlations calculated
~~For dichotomous/nominal/ordinal scores: Was kappa calculated?~~	~~Kappa calculated~~			~~Only percentage agreement calculated~~
~~for ordinal scores: Was a weighted kappa calculated?~~	~~Weighted Kappa calculated~~		~~Unweighted Kappa calculated~~	~~Only percentage agreement calculated~~
~~for ordinal scores: Was the weighting scheme described? e.g. linear, quadratic~~	~~Weighting scheme described~~	~~Weighting scheme NOT described~~		

D. Criterion/Predictive Validity (Predictive in this case) N= 59 subjects; 114 legs

Design requirements	Excellent	Good	Fair	Poor
Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described	NA	NA
Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled- did not use controls in the final linear regression model	Not clear how missing items were handled Not clear if there were missing items	
Was the sample size included in the analysis adequate?	Adequate sample size (≥100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30)
Can the criterion used or	Criterion used can be considered an	No evidence provided, but	Unclear whether the criterion used	Criterion used can NOT be

employed be considered as a reasonable 'gold standard'?	adequate 'gold standard' (evidence provided)	assumable that the criterion used can be considered an adequate 'gold standard'	can be considered an adequate 'gold standard'	considered an adequate 'gold standard'
Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study	Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study	Were there any important flaws in the design or methods of the study?
for continuous scores: Were correlations, or the area under the receiver operating curve calculated?	Correlations or AUC calculated 1. correlation between video peak knee flexion and the anterior reach of the YBT 2. No correlation between YBT and peak knee valgus 3. since it was predictive validity, a stepwise regression method was used			Correlations or AUC NOT calculated
~~for dichotomous scores: Were sensitivity and specificity determined?~~	~~Sensitivity and specificity calculated~~			~~Sensitivity and specificity NOT calculated~~

Noyes/Am J Sports Med/1991

STEP1: Evaluated measurement properties in the article

Check if present	Property	Location
	Reliability	Box A
	Agreement/Measurement Error	Box B
✔	Hypothesis Testing / Construct validity	Box C
✔	Criterion/Predictive Validity	Box D
	Responsiveness	Box E

Step 2. Determining if a study meets the standards for good methodological quality
C. Hypothesis Testing / Construct Validity n=67

Design requirements	Excellent	Good	Fair	Poor
Was the percentage of missing items given? none	Percentage of missing items described	Percentage of missing items NOT described	NA	NA

Was there a description of how missing items were handled? none	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
Was the sample size included in the analysis adequate?	Adequate sample size (≥100 per analysis)	Good sample size (50-99 per analysis)	Moderate sample size (30-49 per analysis)	Small sample size (<30 per analysis)
Were hypotheses regarding correlations or mean differences formulated a priori (i.e. before data collection)?	hypotheses formulated a priori	Minimal number of hypotheses formulate a priori	Hypotheses vague or not formulated but possible to deduce what was expected	Unclear what was expected
Was the expected *direction* of correlations or mean differences included in the hypotheses?	Expected direction of the correlations or differences stated	Expected direction of the correlations or differences NOT stated –		
Was the expected absolute or relative *magnitude* of correlations or mean differences included in the hypotheses?	Expected magnitude of the correlations or differences stated	Expected magnitude of the correlations or differences NOT stated		
for convergent validity: Was an adequate description provided of the comparator instrument(s)?	Adequate description of the constructs measured by the comparator instrument(s)	Adequate description of most of the constructs measured by the comparator instrument(s)	Poor description of the constructs measured by the comparator instrument(s)	NO description of the constructs measured by the comparator instrument(s)
for convergent validity: Were the measurement properties of the comparator instrument(s) adequately described?	Adequate measurement properties of the comparator instrument(s) in a population similar to the study population	Adequate measurement properties of the comparator instrument(s) but not sure if these apply to the study population	Some information on measurement properties (or a reference to a study on measurement properties) of the comparator instrument(s) in any study population	No information on the measurement properties of the comparator instrument(s)
Were there any important flaws	No other important methodological		Other minor methodological	Other important methodological

in the design or methods of the study?	flaws in the design or execution of the study		flaws in the design or execution of the study (e.g. only data presented on a comparison with an instrument that measures another construct)	flaws in the design or execution of the study
Were design and statistical methods adequate for the hypotheses to be tested?	Statistical methods applied appropriate- single and then multiple linear regression for limb symmetry as dependent variable and only significant correlation was a low one (.49) with isokinetic testing at 60 degrees/second	Assumable that statistical methods were appropriate, e.g. Pearson correlations applied, but distribution of scores or mean (SD) not presented	Statistical methods applied NOT optimal	Statistical methods applied NOT appropriate

D. Criterion/Predictive Validity N= 67 subjects

Design requirements	Excellent	Good	Fair	Poor
Was the percentage of missing items given? none	Percentage of missing items described	Percentage of missing items NOT described	NA	NA
Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled- did not use controls in the final linear regression model	Not clear how missing items were handled Not clear if there were missing items	
Was the sample size included in the analysis adequate?	Adequate sample size (≥100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30)
Can the criterion used or employed be considered as a reasonable 'gold standard'?	Criterion used can be considered an adequate 'gold standard' (evidence provided)	No evidence provided, but assumable that the criterion used can be considered an adequate 'gold standard'	Unclear whether the criterion used can be considered an adequate 'gold standard'	Criterion used can NOT be considered an adequate 'gold standard'
Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study	Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study	Were there any important flaws in the design or methods of the study?

for continuous scores: Were correlations, or the area under the receiver operating curve calculated?	Correlations or AUC calculated			Correlations or AUC NOT calculated
for dichotomous scores: Were sensitivity and specificity determined?	Sensitivity and specificity calculated Single hop 52;97; LR+ 17.3 Timed hop 49;92;LR+ 6.125			Sensitivity and specificity NOT calculated

Purdam/PT in Sport/2003

STEP1: Evaluated measurement properties in the article

Check if present	Property	Location
✔	Reliability	Box A
	Agreement/Measurement Error	Box B
	Hypothesis Testing / Construct validity	Box C
	Criterion/Predictive Validity	Box D
✔	Responsiveness	Box E

Step 2. Determining if a study meets the standards for good methodological quality

A. Reliability n=46 subjects

Design requirements	Excellent	Good	Fair	Poor
Was the percentage of missing subjects given?	Percentage of missing subjects described	Percentage of missing subjects NOT described		
Was there a description of how missing subjects were handled?	Described how missing subjects were handled	Not described but it can be deduced how missing subjects were handled	Not clear how missing subjects were handled	
Was the sample size included in the analysis adequate?	Adequate sample size (≥100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30)
Were at least two measurements available? Intra and inter intra only	At least two measurements			Only one measurement
Were the administrations independent? Intra	Independent measurements	Assumable that the measurements were independent	Doubtful whether the measurements were independent	Measurements NOT independent
Was the time interval stated? Intra	Time interval stated		Time interval NOT stated	
Were patients stable in the interim period on the	Patients were stable	Assumable that patients	Unclear if patients were	Patients were NOT stable

construct to be measured? Intra	(evidence provided)	were stable	stable	
Was the time interval appropriate? Intra	Time interval appropriate		Doubtful whether time interval was appropriate	Time interval NOT appropriate
Were the test conditions similar for both measurements? e.g. type of administration, environment, instructions	Test conditions were similar (evidence provided)	Assumable that test conditions were similar	Unclear if test conditions were similar	Test conditions were NOT similar
Were there any important flaws in the design or methods of the study? Intra	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study
~~for continuous scores: Was an intraclass correlation coefficient (ICC) calculated?~~	~~ICC calculated and model or formula of the ICC is described~~	~~ICC calculated but model or formula of the ICC not described or not optimal. Pearson or Spearman correlation coefficient calculated with evidence provided that no systematic change has occurred~~	~~Pearson or Spearman correlation coefficient calculated WITHOUT evidence provided that no systematic change has occurred or WITH evidence that systematic change has occurred~~	~~No ICC or Pearson or Spearman correlations calculated~~
For dichotomous/nominal/ordinal scores: Was kappa calculated?	Kappa calculated			Only percentage agreement calculated
for ordinal scores: Was a weighted kappa calculated?	Weighted Kappa calculated		Unweighted Kappa calculated	Only percentage agreement calculated – actually TEM calculated
for ordinal scores: Was the weighting scheme described? e.g. linear, quadratic	Weighting scheme described	Weighting scheme NOT described	With TEM- not sure how to answer this	

E. Responsiveness n=56, then 50 eligible, then 15/17 control and 13/15 case subjects completing 3 testing sessions + 18 more added for reliability study

Design requirements	Excellent	Good	Fair	Poor
Was the percentage of	Percentage of missing items	Percentage of missing items		

missing items given?	described	NOT described		
Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
Was the sample size included in the analysis adequate?	Adequate sample size (≥100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30) N=28
Was a longitudinal design with at least two measurement used?	Longitudinal design used			No longitudinal design used
Was the time interval stated?	Time interval adequately described			Time interval NOT described
If anything occurred in the interim period (e.g. intervention, other relevant events), was it adequately described?	Anything that occurred during the interim period (e.g. treatment) adequately described	Assumable what occurred during the interim period	Unclear or NOT described what occurred during the interim period	If anything occurred in the interim period (e.g. intervention, other relevant events), was it adequately described?
Was a proportion of the patients changed (i.e. improvement or deterioration)?	Part of the patients were changed (evidence provided)	NO evidence provided, but assumable that part of the patients were changed	Unclear if part of the patients were changed	Patients were NOT changed
For constructs for which a gold standard was not available:				
~~Were hypotheses about changes in scores formulated a priori (i.e. before data collection)?~~	~~Hypotheses formulated a priori~~		~~Hypotheses vague or not formulated but possible to deduce what was expected~~	~~Unclear what was expected~~
~~Was the expected *direction* of correlations or mean differences of the change scores of HR-PRO instruments included in these hypotheses?~~	~~Expected direction of the correlations or differences stated~~	~~Expected direction of the correlations or differences NOT stated~~		
~~Were the expected absolute or relative *magnitude* of correlations or mean differences~~	~~Expected magnitude of the correlations or differences stated~~	~~Expected magnitude of the correlations or differences NOT stated~~		

~~of the change scores of HR-PRO instruments included in these hypotheses?~~				
~~Was an adequate description provided of the comparator instrument(s)?~~	~~Adequate description of the constructs measured by the comparator instrument(s)~~		~~Poor description of the constructs measured by the comparator instrument(s)~~	~~NO description of the constructs measured by the comparator instrument(s)~~
~~Were the measurement properties of the comparator instrument(s) adequately described?~~	~~Adequate measurement properties of the comparator instrument(s) in a population similar to the study population~~	~~Adequate measurement properties of the comparator instrument(s) but not sure if these apply to the study population~~	~~Some information on measurement properties (or a reference to a study on measurement properties) of the comparator instrument(s) in any study population~~	~~NO information on the measurement properties of the comparator instrument(s)~~
~~Were there any important flaws in the design or methods of the study?~~	~~No other important methodological flaws in the design or execution of the study~~		~~Other minor methodological flaws in the design or execution of the study (e.g. only data presented on a comparison with an instrument that measures another construct)~~	~~Other important methodological flaws in the design or execution of the study~~
~~Were design and statistical methods adequate for the hypotheses to be tested?~~	~~Statistical methods applied appropriate~~		~~Statistical methods applied NOT optimal~~	~~Statistical methods applied NOT appropriate~~
For constructs for which a gold standard was available:				
Can the criterion for change be considered as a reasonable gold standard?	Criterion used can be considered an adequate 'gold standard' (evidence provided)	No evidence provided, but assumable that the criterion used can be considered an adequate 'gold standard'	Unclear whether the criterion used can be considered an adequate 'gold standard'	Criterion used can NOT be considered an adequate 'gold standard'
Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study
~~for continuous~~	~~Correlations or~~	Change in 0-10		~~Correlations or~~

~~scores: Were correlations between change scores, or the area under the Receiver Operator Curve (ROC) curve calculated?~~	~~Area under the ROC Curve (AUC) calculated~~	pain scale is interval data		~~AUC NOT calculated~~
~~for dichotomous scales: Were sensitivity and specificity (changed versus not changed) determined?~~	~~Sensitivity and specificity calculated~~			~~Sensitivity and specificity NOT calculated~~

Ross/Knee Sport Taumatol/2002

STEP1: Evaluated measurement properties in the article

Check if present	Property	Location
✔	Reliability	Box A
	Agreement/Measurement Error	Box B
✔	Hypothesis Testing / Construct validity	Box C
	Criterion/Predictive Validity	Box D
	Responsiveness	Box E

Step 2. Determining if a study meets the standards for good methodological quality

A. Reliability n=50 subjects but only 10 for reliability

Design requirements	Excellent	Good	Fair	Poor
Was the percentage of missing subjects given?	Percentage of missing subjects described	Percentage of missing subjects NOT described		
Was there a description of how missing subjects were handled?	Described how missing subjects were handled	Not described but it can be deduced how missing subjects were handled	Not clear how missing subjects were handled	
Was the sample size included in the analysis adequate?	Adequate sample size (≥100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30)
Were at least two measurements available? Intra and inter intra only	At least two measurements			Only one measurement
Were the administrations independent? Intra	Independent measurements	Assumable that the measurements were independent	Doubtful whether the measurements were independent	Measurements NOT independent
Was the time interval stated? Intra	Time interval stated		Time interval NOT stated	
Were patients stable in the interim period on the	Patients were stable	Assumable that patients	Unclear if patients were	Patients were NOT stable

construct to be measured? Intra	(evidence provided)	were stable	stable	
Was the time interval appropriate? Intra	Time interval appropriate		Doubtful whether time interval was appropriate	Time interval NOT appropriate
Were the test conditions similar for both measurements? e.g. type of administration, environment, instructions	Test conditions were similar (evidence provided)	Assumable that test conditions were similar	Unclear if test conditions were similar	Test conditions were NOT similar
Were there any important flaws in the design or methods of the study? Intra	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study
for continuous scores: Was an intraclass correlation coefficient (ICC) calculated?	ICC calculated and model or formula of the ICC is described	ICC calculated but model or formula of the ICC not described or not optimal. Pearson or Spearman correlation coefficient calculated with evidence provided that no systematic change has occurred	Pearson or Spearman correlation coefficient calculated WITHOUT evidence provided that no systematic change has occurred or WITH evidence that systematic change has occurred	~~No ICC or Pearson or Spearman correlations calculated~~
~~For dichotomous/nominal/ordinal scores: Was kappa calculated?~~	~~Kappa calculated~~			~~Only percentage agreement calculated~~
~~for ordinal scores: Was a weighted kappa calculated?~~	~~Weighted Kappa calculated~~		~~Unweighted Kappa calculated~~	~~Only percentage agreement calculated – actually TEM calculated~~
~~for ordinal scores: Was the weighting scheme described? e.g. linear, quadratic~~	~~Weighting scheme described~~	~~Weighting scheme NOT described~~	~~With TEM- not sure how to answer this~~	

C. Hypothesis Testing / Construct Validity

Design requirements	Excellent	Good	Fair	Poor
Was the percentage of missing items	Percentage of missing items described	Percentage of missing items NOT described	NA	NA

given?				
Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
Was the sample size included in the analysis adequate?	Adequate sample size (≥100 per analysis)	Good sample size (50-99 per analysis)	Moderate sample size (30-49 per analysis)	Small sample size (<30 per analysis)
Were hypotheses regarding correlations or mean differences formulated a priori (i.e. before data collection)?	hypotheses formulated a priori	Minimal number of hypotheses formulate a priori	Hypotheses vague or not formulated but possible to deduce what was expected	Unclear what was expected
Was the expected *direction* of correlations or mean differences included in the hypotheses?	Expected direction of the correlations or differences stated- functional tests correlate positively with self-report	Expected direction of the correlations or differences NOT stated –		
Was the expected absolute or relative *magnitude* of correlations or mean differences included in the hypotheses?	Expected magnitude of the correlations or differences stated	Expected magnitude of the correlations or differences NOT stated		
for convergent validity: Was an adequate description provided of the comparator instrument(s)?	Adequate description of the constructs measured by the comparator instrument(s)	Adequate description of most of the constructs measured by the comparator instrument(s)	Poor description of the constructs measured by the comparator instrument(s)	NO description of the constructs measured by the comparator instrument(s)
for convergent validity: Were the measurement properties of the comparator instrument(s) adequately described?	Adequate measurement properties of the comparator instrument(s) in a population similar to the study population- KOS, ADLS, SAS used but only the	Adequate measurement properties of the comparator instrument(s) but not sure if these apply to the study population	Some information on measurement properties (or a reference to a study on measurement properties) of the comparator instrument(s) in any study population	No information on the measurement properties of the comparator instrument(s) NOTE: First time tested in this study

	rel/valid of KOS cited			
Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study (e.g. only data presented on a comparison with an instrument that measures another construct)	Other important methodological flaws in the design or execution of the study
Were design and statistical methods adequate for the hypotheses to be tested?	Statistical methods applied appropriate	Assumable that statistical methods were appropriate, e.g. Pearson correlations applied, but distribution of scores or mean (SD) not presented	Statistical methods applied NOT optimal	Statistical methods applied NOT appropriate

Ross/Knee Sport Taumatol/2010

STEP1: Evaluated measurement properties in the article

Check if present	Property	Location
	Reliability	Box A
	Agreement/Measurement Error	Box B
✔	Hypothesis Testing / Construct validity	Box C
	Criterion/Predictive Validity	Box D
	Responsiveness	Box E

Step 2. Determining if a study meets the standards for good methodological quality

C. Hypothesis Testing / Construct Validity n=48

Design requirements	Excellent	Good	Fair	Poor
Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described	NA	NA
Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
Was the sample size included in the analysis adequate?	Adequate sample size (≥100 per analysis)	Good sample size (50-99 per analysis)	Moderate sample size (30-49 per analysis)	Small sample size (<30 per analysis)

Were hypotheses regarding correlations or mean differences formulated a priori (i.e. before data collection)?	hypotheses formulated a priori	Minimal number of hypotheses formulate a priori	Hypotheses vague or not formulated but possible to deduce what was expected	Unclear what was expected
Was the expected *direction* of correlations or mean differences included in the hypotheses?	Expected direction of the correlations or differences stated- FABQ correlate positively with self-report	Expected direction of the correlations or differences NOT stated –		
Was the expected absolute or relative *magnitude* of correlations or mean differences included in the hypotheses?	Expected magnitude of the correlations or differences stated	Expected magnitude of the correlations or differences NOT stated		
for convergent validity: Was an adequate description provided of the comparator instrument(s)?	Adequate description of the constructs measured by the comparator instrument(s)	Adequate description of most of the constructs measured by the comparator instrument(s)	Poor description of the constructs measured by the comparator instrument(s)	NO description of the constructs measured by the comparator instrument(s)
for convergent validity: Were the measurement properties of the comparator instrument(s) adequately described?	Adequate measurement properties of the comparator instrument(s) in a population similar to the study population- KOS, SAS, ADLS used and the reliability/validity of KOS cited	Adequate measurement properties of the comparator instrument(s) but not sure if these apply to the study population	Some information on measurement properties (or a reference to a study on measurement properties) of the comparator instrument(s) in any study population	No information on the measurement properties of the comparator instrument(s) NOTE: First time tested in this study
Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study (e.g. only data presented on a comparison with an	Other important methodological flaws in the design or execution of the study –

			instrument that measures another construct)- they adapted FABQ to knee so metrics of FABQ not known in knee	
Were design and statistical methods adequate for the hypotheses to be tested?	Statistical methods applied appropriate	Assumable that statistical methods were appropriate, e.g. Pearson correlations applied, but distribution of scores or mean (SD) not presented	Statistical methods applied NOT optimal	Statistical methods applied NOT appropriate

Svensson/Knee Sport Taumatol/2006

STEP1: Evaluated measurement properties in the article

Check if present	Property	Location
	Reliability	Box A
	Agreement/Measurement Error	Box B
	Hypothesis Testing / Construct validity	Box C
	Criterion/Predictive Validity	Box D
✔	Responsiveness	Box E

Step 2. Determining if a study meets the standards for good methodological quality

E. Responsiveness n=59

Design requirements	Excellent	Good	Fair	Poor
Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
Was the sample size included in the analysis adequate?	Adequate sample size (≥100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30)
Was a longitudinal design with at least two measurement used?	Longitudinal design used			No longitudinal design used
Was the time interval stated?	Time interval adequately described			Time interval NOT described

If anything occurred in the interim period (e.g. intervention, other relevant events), was it adequately described?	Anything that occurred during the interim period (e.g. treatment) adequately described	Assumable what occurred during the interim period	Unclear or NOT described what occurred during the interim period	If anything occurred in the interim period (e.g. intervention, other relevant events), was it adequately described?
Was a proportion of the patients changed (i.e. improvement or deterioration)?	Part of the patients were changed (evidence provided)- actually, all were improved at 2 year follow-up	NO evidence provided, but assumable that part of the patients were changed	Unclear if part of the patients were changed	Patients were NOT changed
For constructs for which a gold standard was not available:				
Were hypotheses about changes in scores formulated a priori (i.e. before data collection)?	Hypotheses formulated a priori		Hypotheses vague or not formulated but possible to deduce what was expected	Unclear what was expected
Was the expected *direction* of correlations or mean differences of the change scores of HR-PRO instruments included in these hypotheses?	Expected direction of the correlations or differences stated	Expected direction of the correlations or differences NOT stated		
Were the expected absolute or relative *magnitude* of correlations or mean differences of the change scores of HR-PRO instruments included in these hypotheses?	Expected magnitude of the correlations or differences stated	Expected magnitude of th√e correlations or differences NOT stated		
Was an adequate description provided of the comparator instrument(s)?	Adequate description of the constructs measured by the comparator instrument(s)		Poor description of the constructs measured by the comparator instrument(s)	NO description of the constructs measured by the comparator instrument(s)
Were the measurement properties of the comparator instrument(s) adequately	Adequate measurement properties of the comparator instrument(s) in a population	Adequate measurement properties of the comparator instrument(s) but not sure if	Some information on measurement properties (or a reference to a study on measurement	NO information on the measurement properties of the comparator instrument(s)

described?	similar to the study population	these apply to the study population	properties) of the comparator instrument(s) in any study population	
Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study (e.g. only data presented on a comparison with an instrument that measures another construct)	Other important methodological flaws in the design or execution of the study
Were design and statistical methods adequate for the hypotheses to be tested?	Statistical methods applied appropriate		Statistical methods applied NOT optimal	Statistical methods applied NOT appropriate
~~**For constructs for which a gold standard was available:**~~				
~~Can the criterion for change be considered as a reasonable gold standard?~~	~~Criterion used can be considered an adequate 'gold standard' (evidence provided)~~	~~No evidence provided, but assumable that the criterion used can be considered an adequate 'gold standard'~~	~~Unclear whether the criterion used can be considered an adequate 'gold standard'~~	~~Criterion used can NOT be considered an adequate 'gold standard'~~
~~Were there any important flaws in the design or methods of the study?~~	~~No other important methodological flaws in the design or execution of the study~~		~~Other minor methodological flaws in the design or execution of the study~~	~~Other important methodological flaws in the design or execution of the study~~
~~for continuous scores: Were correlations between change scores, or the area under the Receiver Operator Curve (ROC) curve calculated?~~	~~Correlations or Area under the ROC Curve (AUC) calculated~~			~~Correlations or AUC NOT calculated~~
~~for dichotomous scales: Were sensitivity and specificity (changed versus not changed) determined?~~	~~Sensitivity and specificity calculated~~			~~Sensitivity and specificity NOT calculated~~

STEP1: Evaluated measurement properties in the article

Check if present	Property	Location
	Reliability	Box A
	Agreement/Measurement Error	Box B
✔	Hypothesis Testing / Construct validity	**Box C**
	Criterion/Predictive Validity	Box D
	Responsiveness	Box E

Step 2. Determining if a study meets the standards for good methodological quality
C. Hypothesis Testing / Construct Validity n=48

Design requirements	Excellent	Good	Fair	Poor
Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described	NA	NA
Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
Was the sample size included in the analysis adequate?	Adequate sample size (≥100 per analysis)	Good sample size (50-99 per analysis)	Moderate sample size (30-49 per analysis)	Small sample size (<30 per analysis)
Were hypotheses regarding correlations or mean differences formulated a priori (i.e. before data collection)?	hypotheses formulated a priori	Minimal number of hypotheses formulate a priori	Hypotheses vague or not formulated but possible to deduce what was expected	Unclear what was expected
Was the expected *direction* of correlations or mean differences included in the hypotheses?	Expected direction of the correlations or differences stated-	Expected direction of the correlations or differences NOT stated –		
Was the expected absolute or relative *magnitude* of correlations or mean differences included in the hypotheses?	Expected magnitude of the correlations or differences stated	Expected magnitude of the correlations or differences NOT stated		
~~for convergent validity: Was an adequate~~	~~Adequate description of the constructs~~	~~Adequate description of most of the~~	~~Poor description of the constructs measured by the~~	~~NO description of the constructs measured by the~~

~~description provided of the comparator instrument(s)?~~	~~measured by the comparator instrument(s)~~	~~constructs measured by the comparator instrument(s)~~	~~comparator instrument(s)~~	~~comparator instrument(s)~~
~~for convergent validity: Were the measurement properties of the comparator instrument(s) adequately described?~~	~~Adequate measurement properties of the comparator instrument(s) in a population similar to the study population- KOS, SAS, ADLS used and the rel/valid of KOS cited~~	~~Adequate measurement properties of the comparator instrument(s) but not sure if these apply to the study population~~	~~Some information on measurement properties (or a reference to a study on measurement properties) of the comparator instrument(s) in any study population~~	~~No information on the measurement properties of the comparator instrument(s) NOTE: First time tested in this study~~
Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study (e.g. only data presented on a comparison with an instrument that measures another construct)	Other important methodological flaws in the design or execution of the study –
Were design and statistical methods adequate for the hypotheses to be tested?	Statistical methods applied appropriate	Assumable that statistical methods were appropriate, e.g. Pearson correlations applied, but distribution of scores or mean (SD) not presented	Statistical methods applied NOT optimal- correlation coefficients would be more appropriate	Statistical methods applied NOT appropriate

Witvrouw/Scand J Med Sci Sports/2002

STEP1: Evaluated measurement properties in the article

Check if present	Property	Location
✔	Reliability	Box A
	Agreement/Measurement Error	Box B
✔	Hypothesis Testing / Construct validity	Box C
	Criterion/Predictive Validity	Box D
	Responsiveness	Box E

Step 2. Determining if a study meets the standards for good methodological quality

A. Reliability n=25 for reliability

Design requirements	Excellent	Good	Fair	Poor
Was the percentage of missing subjects given? none	Percentage of missing subjects	Percentage of missing subjects NOT		

	described	described		
Was there a description of how missing subjects were handled?	Described how missing subjects were handled	Not described but it can be deduced how missing subjects were handled	Not clear how missing subjects were handled	
Was the sample size included in the analysis adequate?	Adequate sample size (≥100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30)
Were at least two measurements available? Intra and inter intra only	At least two measurements			Only one measurement
Were the administrations independent? Intra	Independent measurements	Assumable that the measurements were independent	Doubtful whether the measurements were independent	Measurements NOT independent
Was the time interval stated? Intra	Time interval stated		Time interval NOT stated	
Were patients stable in the interim period on the construct to be measured? Intra	Patients were stable (evidence provided)	Assumable that patients were stable	Unclear if patients were stable	Patients were NOT stable
Was the time interval appropriate? Intra	Time interval appropriate		Doubtful whether time interval was appropriate	Time interval NOT appropriate
Were the test conditions similar for both measurements? e.g. type of administration, environment, instructions	Test conditions were similar (evidence provided)	Assumable that test conditions were similar	Unclear if test conditions were similar	Test conditions were NOT similar
Were there any important flaws in the design or methods of the study? Intra	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study
for continuous scores: Was an intraclass correlation coefficient (ICC) calculated?	ICC calculated and model or formula of the ICC is described	ICC calculated but model or formula of the ICC not described or not optimal. Pearson or Spearman correlation coefficient calculated with evidence provided that no systematic	Pearson or Spearman correlation coefficient calculated WITHOUT evidence provided that no systematic change has occurred or WITH evidence that systematic	~~No ICC or Pearson or Spearman correlations calculated~~

		change has occurred	change has occurred	
~~For dichotomous/nominal/ordinal scores: Was kappa calculated?~~	~~Kappa calculated~~			~~Only percentage agreement calculated~~
~~for ordinal scores: Was a weighted kappa calculated?~~	~~Weighted Kappa calculated~~		~~Unweighted Kappa calculated~~	~~Only percentage agreement calculated – actually TEM calculated~~
~~for ordinal scores: Was the weighting scheme described? e.g. linear, quadratic~~	~~Weighting scheme described~~	~~Weighting scheme NOT described~~	~~With TEM- not sure how to answer this~~	

C. Hypothesis Testing / Construct Validity n=30

Design requirements	Excellent	Good	Fair	Poor
Was the percentage of missing items given? none	Percentage of missing items described	Percentage of missing items NOT described	NA	NA
Was there a description of how missing items were handled?	Described how missing items were handled- all in final analysis	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
Was the sample size included in the analysis adequate?	Adequate sample size (≥100 per analysis)	Good sample size (50-99 per analysis)	Moderate sample size (30-49 per analysis)	Small sample size (<30 per analysis)
Were hypotheses regarding correlations or mean differences formulated a priori (i.e. before data collection)?	hypotheses formulated a priori	Minimal number of hypotheses formulate a priori	Hypotheses vague or not formulated but possible to deduce what was expected	Unclear what was expected
Was the expected *direction* of correlations or mean differences included in the hypotheses?	Expected direction of the correlations or differences stated-	Expected direction of the correlations or differences NOT stated –		
Was the expected absolute or relative *magnitude* of correlations or	Expected magnitude of the correlations or differences stated	Expected magnitude of the correlations or differences NOT stated		

mean differences included in the hypotheses?				
for convergent validity: Was an adequate description provided of the comparator instrument(s)?	Adequate description of the constructs measured by the comparator instrument(s)	Adequate description of most of the constructs measured by the comparator instrument(s)	Poor description of the constructs measured by the comparator instrument(s)	NO description of the constructs measured by the comparator instrument(s)
for convergent validity: Were the measurement properties of the comparator instrument(s) adequately described?	Adequate measurement properties of the comparator instrument(s) in a population similar to the study population-	Adequate measurement properties of the comparator instrument(s) but not sure if these apply to the study population	Some information on measurement properties (or a reference to a study on measurement properties) of the comparator instrument(s) in any study population	No information on the measurement properties of the comparator instrument(s) NOTE: First time tested in this study
Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study (e.g. only data presented on a comparison with an instrument that measures another construct)	Other important methodological flaws in the design or execution of the study –
Were design and statistical methods adequate for the hypotheses to be tested?	Statistical methods applied appropriate	Assumable that statistical methods were appropriate, e.g. Pearson correlations applied, but distribution of scores or mean (SD) not presented	Statistical methods applied NOT optimal- correlation coefficients would be more appropriate	Statistical methods applied NOT appropriate

Zouita/Annals of Phys & Rehab Medicine/2009

STEP1: Evaluated measurement properties in the article

Check if present	Property	Location
	Reliability	Box A
	Agreement/Measurement Error	Box B
✔	Hypothesis Testing / Construct validity	Box C
	Criterion/Predictive Validity	Box D
	Responsiveness	Box E

Step 2. Determining if a study meets the standards for good methodological quality
C. Hypothesis Testing / Construct Validity n=46

Design requirements	Excellent	Good	Fair	Poor
Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described	NA	NA
Was there a description of how missing items were handled?	Described how missing items were handled- all in final analysis	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
Was the sample size included in the analysis adequate?	Adequate sample size (≥100 per analysis)	Good sample size (50-99 per analysis)	Moderate sample size (30-49 per analysis)	Small sample size (<30 per analysis)
Were hypotheses regarding correlations or mean differences formulated a priori (i.e. before data collection)?	hypotheses formulated a priori	Minimal number of hypotheses formulate a priori	Hypotheses vague or not formulated but possible to deduce what was expected	Unclear what was expected
Was the expected *direction* of correlations or mean differences included in the hypotheses?	Expected direction of the correlations or differences stated-	Expected direction of the correlations or differences NOT stated –		
Was the expected absolute or relative *magnitude* of correlations or mean differences included in the hypotheses?	Expected magnitude of the correlations or differences stated	Expected magnitude of the correlations or differences NOT stated		
for convergent validity: Was an adequate description provided of the comparator instrument(s)?	Adequate description of the constructs measured by the comparator instrument(s) – same test in a healthy population	Adequate description of most of the constructs measured by the comparator instrument(s)	Poor description of the constructs measured by the comparator instrument(s)	NO description of the constructs measured by the comparator instrument(s)
for convergent validity: Were the measurement properties of the comparator	Adequate measurement properties of the comparator instrument(s) in a	Adequate measurement properties of the comparator instrument(s)	Some information on measurement properties (or a reference to a study on	No information on the measurement properties of the comparator

instrument(s) adequately described?	population similar to the study population-	but not sure if these apply to the study population	measurement properties) of the comparator instrument(s) in any study population	instrument(s) NOTE: First time tested in this study
Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study (e.g. only data presented on a comparison with an instrument that measures another construct)	Other important methodological flaws in the design or execution of the study –
Were design and statistical methods adequate for the hypotheses to be tested?	Statistical methods applied appropriate	Assumable that statistical methods were appropriate, e.g. Pearson correlations applied, but distribution of scores or mean (SD) not presented	Statistical methods applied NOT optimal- correlation coefficients would be more appropriate	Statistical methods applied NOT appropriate

Barber/CORR/1990

STEP1: Evaluated measurement properties in the article

Check if present	Property	Location
	Reliability	Box A
	Agreement/Measurement Error	Box B
✔	Hypothesis Testing / Construct validity	Box C
	Criterion/Predictive Validity	Box D
	Responsiveness	Box E

Step 2. Determining if a study meets the standards for good methodological quality

C. Hypothesis Testing / Construct Validity n=93

Design requirements	Excellent	Good	Fair	Poor
Was the percentage of missing items given? none	Percentage of missing items described	Percentage of missing items NOT described	NA	NA
Was there a description of how missing items were handled?	Described how missing items were handled- all in final analysis	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	

Was the sample size included in the analysis adequate?	Adequate sample size (≥100 per analysis)	Good sample size (50-99 per analysis)	Moderate sample size (30-49 per analysis)	Small sample size (<30 per analysis)
Were hypotheses regarding correlations or mean differences formulated a priori (i.e. before data collection)?	hypotheses formulated a priori	Minimal number of hypotheses formulate a priori	Hypotheses vague or not formulated but possible to deduce what was expected	Unclear what was expected
Was the expected *direction* of correlations or mean differences included in the hypotheses?	Expected direction of the correlations or differences stated-	Expected direction of the correlations or differences NOT stated –		
Was the expected absolute or relative *magnitude* of correlations or mean differences included in the hypotheses?	Expected magnitude of the correlations or differences stated	Expected magnitude of the correlations or differences NOT stated		
~~for convergent validity: Was an adequate description provided of the comparator instrument(s)?~~	~~Adequate description of the constructs measured by the comparator instrument(s) – same test in a healthy population~~	~~Adequate description of most of the constructs measured by the comparator instrument(s)~~	~~Poor description of the constructs measured by the comparator instrument(s)~~	~~NO description of the constructs measured by the comparator instrument(s)~~
~~for convergent validity: Were the measurement properties of the comparator instrument(s) adequately described?~~	~~Adequate measurement properties of the comparator instrument(s) in a population similar to the study population~~	~~Adequate measurement properties of the comparator instrument(s) but not sure if these apply to the study population~~	~~Some information on measurement properties (or a reference to a study on measurement properties) of the comparator instrument(s) in any study population~~	~~No information on the measurement properties of the comparator instrument(s) NOTE: First time tested in this study~~
Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study (e.g. only data presented on a comparison with an instrument that	Other important methodological flaws in the design or execution of the study –

			measures another construct)	
Were design and statistical methods adequate for the hypotheses to be tested?	Statistical methods applied appropriate	Assumable that statistical methods were appropriate, e.g. Pearson correlations applied, but distribution of scores or mean (SD) not presented	Statistical methods applied NOT optimal-correlation coefficients would be more appropriate	Statistical methods applied NOT appropriate

Brosky/JOSPT/1999

STEP1: Evaluated measurement properties in the article

Check if present	Property	Location
✔	Reliability	Box A
	Agreement/Measurement Error	Box B
	Hypothesis Testing / Construct validity	Box C
	Criterion/Predictive Validity	Box D
	Responsiveness	Box E

Step 2. Determining if a study meets the standards for good methodological quality

A. Reliability n=15 for reliability

Design requirements	Excellent	Good	Fair	Poor
Was the percentage of missing subjects given? none	Percentage of missing subjects described	Percentage of missing subjects NOT described		
Was there a description of how missing subjects were handled? none	Described how missing subjects were handled	Not described but it can be deduced how missing subjects were handled	Not clear how missing subjects were handled	
Was the sample size included in the analysis adequate?	Adequate sample size (≥100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30) n=15
Were at least two measurements available? Intra and inter intra only	At least two measurements			Only one measurement
Were the administrations independent? Intra	Independent measurements	Assumable that the measurements were independent	Doubtful whether the measurements were independent	Measurements NOT independent
Was the time interval stated? Intra	Time interval stated		Time interval NOT stated	
Were patients stable in the interim period on the construct to be measured? Intra	Patients were stable (evidence provided)	Assumable that patients were stable	Unclear if patients were stable	Patients were NOT stable

Was the time interval appropriate? Intra	Time interval appropriate		Doubtful whether time interval was appropriate	Time interval NOT appropriate
Were the test conditions similar for both measurements? e.g. type of administration, environment, instructions	Test conditions were similar (evidence provided)	Assumable that test conditions were similar	Unclear if test conditions were similar	Test conditions were NOT similar
Were there any important flaws in the design or methods of the study? Intra	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study
for continuous scores: Was an intraclass correlation coefficient (ICC) calculated?	ICC calculated and model or formula of the ICC is described	ICC calculated but model or formula of the ICC not described or not optimal. Pearson or Spearman correlation coefficient calculated with evidence provided that no systematic change has occurred	Pearson or Spearman correlation coefficient calculated WITHOUT evidence provided that no systematic change has occurred or WITH evidence that systematic change has occurred	~~No ICC or Pearson or Spearman correlations calculated~~
~~For dichotomous/nominal/ordinal scores: Was kappa calculated?~~	~~Kappa calculated~~			~~Only percentage agreement calculated~~
~~for ordinal scores: Was a weighted kappa calculated?~~	~~Weighted Kappa calculated~~		~~Unweighted Kappa calculated~~	~~Only percentage agreement calculated – actually TEM calculated~~
~~for ordinal scores: Was the weighting scheme described? e.g. linear, quadratic~~	~~Weighting scheme described~~	~~Weighting scheme NOT described~~	~~With TEM- not sure how to answer this~~	

Grindem/Am J Sports Med/2011

STEP1: Evaluated measurement properties in the article

Check if present	Property	Location
	Reliability	Box A
	Agreement/Measurement Error	Box B
	Hypothesis Testing / Construct validity	Box C

✔	Criterion/Predictive Validity	Box D
	Responsiveness	Box E

Step 2. Determining if a study meets the standards for good methodological quality
D. Criterion/predictive Validity n=91 but 10 were lost in follow-up so n=81

Design requirements	Excellent	Good	Fair	Poor
Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described	NA	NA
Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled-	Not clear how missing items were handled Not clear if there were missing items	
Was the sample size included in the analysis adequate?	Adequate sample size (≥100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30)
Can the criterion used or employed be considered as a reasonable 'gold standard'?	Criterion used can be considered an adequate 'gold standard' (evidence provided)	No evidence provided, but assumable that the criterion used can be considered an adequate 'gold standard'	Unclear whether the criterion used can be considered an adequate 'gold standard'	Criterion used can NOT be considered an adequate 'gold standard'
Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study	Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study	Were there any important flaws in the design or methods of the study?
for continuous scores: Were correlations, or the area under the receiver operating curve calculated?	Correlations or AUC calculated-Single hop is only test correlated with self-reported IKDC function at 1 year			Correlations or AUC NOT calculated
for dichotomous scores: Were sensitivity and specificity determined?	Sensitivity and specificity calculated Single hop 71;71; LR+ 2.52 LR- 0.40			Sensitivity and specificity NOT calculated

Logerstedt/Am J Sports Med/2012

STEP1: Evaluated measurement properties in the article

Check if present	Property	Location
	Reliability	Box A
	Agreement/Measurement Error	Box B
	Hypothesis Testing / Construct	Box C

	validity	
✔	Criterion/Predictive Validity	Box D
	Responsiveness	Box E

Step 2. Determining if a study meets the standards for good methodological quality
D. Criterion/predictive Validity pre-op to 1 year n=79; 6 mos to 1 year n=85

Design requirements	Excellent	Good	Fair	Poor
Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described	NA	NA
Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled-	Not clear how missing items were handled Not clear if there were missing items	
Was the sample size included in the analysis adequate?	Adequate sample size (≥100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30)
Can the criterion used or employed be considered as a reasonable 'gold standard'?	Criterion used can be considered an adequate 'gold standard' (evidence provided)	No evidence provided, but assumable that the criterion used can be considered an adequate 'gold standard'	Unclear whether the criterion used can be considered an adequate 'gold standard'	Criterion used can NOT be considered an adequate 'gold standard'
Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study	Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study	Were there any important flaws in the design or methods of the study?
for continuous scores: Were correlations, or the area under the receiver operating curve calculated?	Correlations or AUC calculated- Single hop is only test correlated with self-reported IKDC function at 1 year			Correlations or AUC NOT calculated
for dichotomous scores: Were sensitivity and specificity determined?	Sensitivity and specificity calculated 6m timed hop (87.7% LSI) 53;90; LR+ 5.14 LR- 0.40 Crossover hop (94.9% LSI) 88;47 LR- 0.25			Sensitivity and specificity NOT calculated

Ostenberg/Scand J Med Sci Sports/1998

STEP1: Evaluated measurement properties in the article

Check if present	Property	Location
	Reliability	Box A
	Agreement/Measurement Error	Box B
✔	Hypothesis Testing / Construct validity	Box C
	Criterion/Predictive Validity	Box D
	Responsiveness	Box E

Step 2. Determining if a study meets the standards for good methodological quality
C. Hypothesis Testing / Construct Validity n= 101

Design requirements	Excellent	Good	Fair	Poor
Was the percentage of missing items given? none	Percentage of missing items described	Percentage of missing items NOT described	NA	NA
Was there a description of how missing items were handled? none	Described how missing items were handled- all in final analysis	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
Was the sample size included in the analysis adequate?	Adequate sample size (≥100 per analysis)	Good sample size (50-99 per analysis)	Moderate sample size (30-49 per analysis)	Small sample size (<30 per analysis)
Were hypotheses regarding correlations or mean differences formulated a priori (i.e. before data collection)?	hypotheses formulated a priori	Minimal number of hypotheses formulate a priori	Hypotheses vague or not formulated but possible to deduce what was expected	Unclear what was expected
Was the expected *direction* of correlations or mean differences included in the hypotheses?	Expected direction of the correlations or differences stated-	Expected direction of the correlations or differences NOT stated –		
Was the expected absolute or relative *magnitude* of correlations or mean differences included in the	Expected magnitude of the correlations or differences stated	Expected magnitude of the correlations or differences NOT stated		

hypotheses?				
~~for convergent validity: Was an adequate description provided of the comparator instrument(s)?~~	~~Adequate description of the constructs measured by the comparator instrument(s) – same test in a healthy population~~	~~Adequate description of most of the constructs measured by the comparator instrument(s)~~	~~Poor description of the constructs measured by the comparator instrument(s)~~	~~NO description of the constructs measured by the comparator instrument(s)~~
~~for convergent validity: Were the measurement properties of the comparator instrument(s) adequately described?~~	~~Adequate measurement properties of the comparator instrument(s) in a population similar to the study population~~	~~Adequate measurement properties of the comparator instrument(s) but not sure if these apply to the study population~~	~~Some information on measurement properties (or a reference to a study on measurement properties) of the comparator instrument(s) in any study population~~	~~No information on the measurement properties of the comparator instrument(s) NOTE: First time tested in this study~~
Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study (e.g. only data presented on a comparison with an instrument that measures another construct)	Other important methodological flaws in the design or execution of the study –
Were design and statistical methods adequate for the hypotheses to be tested?	Statistical methods applied appropriate	Assumable that statistical methods were appropriate, e.g. Pearson correlations applied, but distribution of scores or mean (SD) not presented	Statistical methods applied NOT optimal-correlation coefficients would be more appropriate	Statistical methods applied NOT appropriate

Wilk/JOSPT/1994

STEP1: Evaluated measurement properties in the article

Check if present	Property	Location
	Reliability	Box A
	Agreement/Measurement Error	Box B
✔	Hypothesis Testing / Construct validity	Box C
	Criterion/Predictive Validity	Box D
	Responsiveness	Box E

Step 2. Determining if a study meets the standards for good methodological quality
C. Hypothesis Testing / Construct Validity n= 50

Design requirements	Excellent	Good	Fair	Poor
Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described	NA	NA
Was there a description of how missing items were handled?	Described how missing items were handled- all in final analysis	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
Was the sample size included in the analysis adequate?	Adequate sample size (≥100 per analysis)	Good sample size (50-99 per analysis)	Moderate sample size (30-49 per analysis)	Small sample size (<30 per analysis)
Were hypotheses regarding correlations or mean differences formulated a priori (i.e. before data collection)?	hypotheses formulated a priori	Minimal number of hypotheses formulate a priori	Hypotheses vague or not formulated but possible to deduce what was expected	Unclear what was expected
Was the expected *direction* of correlations or mean differences included in the hypotheses?	Expected direction of the correlations or differences stated-	Expected direction of the correlations or differences NOT stated –		
Was the expected absolute or relative *magnitude* of correlations or mean differences included in the hypotheses?	Expected magnitude of the correlations or differences stated	Expected magnitude of the correlations or differences NOT stated		
for convergent validity: Was an adequate description provided of the comparator instrument(s)?	Adequate description of the constructs measured by the comparator instrument(s)	Adequate description of most of the constructs measured by the comparator instrument(s)	Poor description of the constructs measured by the comparator instrument(s)	NO description of the constructs measured by the comparator instrument(s)
for convergent validity: Were the measurement properties of the comparator	Adequate measurement properties of the comparator instrument(s) in a	Adequate measurement properties of the comparator instrument(s)	Some information on measurement properties (or a reference to a study on measurement	No information on the measurement properties of the comparator

instrument(s) adequately described?	population similar to the study population-	but not sure if these apply to the study population	properties) of the comparator instrument(s) in any study population	instrument(s)
Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study (e.g. only data presented on a comparison with an instrument that measures another construct)	Other important methodological flaws in the design or execution of the study –
Were design and statistical methods adequate for the hypotheses to be tested?	Statistical methods applied appropriate	Assumable that statistical methods were appropriate, e.g. Pearson correlations applied, but distribution of scores or mean (SD) not presented	Statistical methods applied NOT optimal- correlation coefficients would be more appropriate	Statistical methods applied NOT appropriate

Augustsson/ Knee Surg Sports Traumatol Arthrosc/2004

STEP1: Evaluated measurement properties in the article

Check if present	Property	Location
	Reliability	Box A
	Agreement/Measurement Error	Box B
✔	Hypothesis Testing / Construct validity	Box C
	Criterion /Predictive Validity	Box D
	Responsiveness	Box E

Step 2. Determining if a study meets the standards for good methodological quality
C. Hypothesis Testing / Construct Validity n= 19

Design requirements	Excellent	Good	Fair	Poor
Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described	NA	NA
Was there a description of how missing items were handled?	Described how missing items were handled- all in final analysis	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	

Was the sample size included in the analysis adequate?	Adequate sample size (≥100 per analysis)	Good sample size (50-99 per analysis)	Moderate sample size (30-49 per analysis)	Small sample size (<30 per analysis)
Were hypotheses regarding correlations or mean differences formulated a priori (i.e. before data collection)?	hypotheses formulated a priori	Minimal number of hypotheses formulate a priori	Hypotheses vague or not formulated but possible to deduce what was expected	Unclear what was expected
Was the expected *direction* of correlations or mean differences included in the hypotheses?	Expected direction of the correlations or differences stated-	Expected direction of the correlations or differences NOT stated –		
Was the expected absolute or relative *magnitude* of correlations or mean differences included in the hypotheses?	Expected magnitude of the correlations or differences stated	Expected magnitude of the correlations or differences NOT stated		
~~for convergent validity: Was an adequate description provided of the comparator instrument(s)?~~	~~Adequate description of the constructs measured by the comparator instrument(s)~~	~~Adequate description of most of the constructs measured by the comparator instrument(s)~~	~~Poor description of the constructs measured by the comparator instrument(s)~~	~~NO description of the constructs measured by the comparator instrument(s)~~
~~for convergent validity: Were the measurement properties of the comparator instrument(s) adequately described?~~	~~Adequate measurement properties of the comparator instrument(s) in a population similar to the study population~~	~~Adequate measurement properties of the comparator instrument(s) but not sure if these apply to the study population~~	~~Some information on measurement properties (or a reference to a study on measurement properties) of the comparator instrument(s) in any study population~~	~~No information on the measurement properties of the comparator instrument(s)~~
Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study (e.g. only data presented on a comparison with an instrument that measures another construct)	Other important methodological flaws in the design or execution of the study –
Were design and	Statistical	Assumable that	Statistical methods	Statistical

statistical methods adequate for the hypotheses to be tested?	methods applied appropriate	statistical methods were appropriate, e.g. Pearson correlations applied, but distribution of scores or mean (SD) not presented	applied NOT optimal- correlation coefficients would be more appropriate	methods applied NOT appropriate

Jerre/ Scan J Med Sci Sports /2001

STEP1: Evaluated measurement properties in the article

Check if present	Property	Location
	Reliability	Box A
	Agreement/Measurement Error	Box B
✔	Hypothesis Testing / Construct validity	Box C
	Criterion/Predictive Validity	Box D
	Responsiveness	Box E

Step 2. Determining if a study meets the standards for good methodological quality
C. Hypothesis Testing / Construct Validity n= 275

Design requirements	Excellent	Good	Fair	Poor
Was the percentage of missing items given? none	Percentage of missing items described	Percentage of missing items NOT described	NA	NA
Was there a description of how missing items were handled?	Described how missing items were handled- all in final analysis	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
Was the sample size included in the analysis adequate?	Adequate sample size (≥100 per analysis)	Good sample size (50-99 per analysis)	Moderate sample size (30-49 per analysis)	Small sample size (<30 per analysis)
Were hypotheses regarding correlations or mean differences formulated a priori (i.e. before data collection)?	hypotheses formulated a priori	Minimal number of hypotheses formulate a priori	Hypotheses vague or not formulated but possible to deduce what was expected	Unclear what was expected
Was the expected *direction* of correlations or mean differences	Expected direction of the correlations or differences	Expected direction of the correlations or differences NOT		

included in the hypotheses?	stated-	stated –		
Was the expected absolute or relative *magnitude* of correlations or mean differences included in the hypotheses?	Expected magnitude of the correlations or differences stated	Expected magnitude of the correlations or differences NOT stated		
for convergent validity: Was an adequate description provided of the comparator instrument(s)?	Adequate description of the constructs measured by the comparator instrument(s)	Adequate description of most of the constructs measured by the comparator instrument(s)	Poor description of the constructs measured by the comparator instrument(s)	NO description of the constructs measured by the comparator instrument(s)
for convergent validity: Were the measurement properties of the comparator instrument(s) adequately described?	Adequate measurement properties of the comparator instrument(s) in a population similar to the study population-	Adequate measurement properties of the comparator instrument(s) but not sure if these apply to the study population	Some information on measurement properties (or a reference to a study on measurement properties) of the comparator instrument(s) in any study population	No information on the measurement properties of the comparator instrument(s)
Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study (e.g. only data presented on a comparison with an instrument that measures another construct)	Other important methodological flaws in the design or execution of the study –
Were design and statistical methods adequate for the hypotheses to be tested?	Statistical methods applied appropriate	Assumable that statistical methods were appropriate, e.g. Pearson correlations applied, but distribution of scores or mean (SD) not presented	Statistical methods applied NOT optimal- correlation coefficients would be more appropriate	Statistical methods applied NOT appropriate

Vandermeulen/Physother Can/1995

STEP1: Evaluated measurement properties in the article

Check if present	Property	Location
✔	Reliability	Box A

	Agreement/Measurement Error	Box B
✔	**Hypothesis Testing / Construct validity**	**Box C**
	Criterion/Predictive Validity	Box D
	Responsiveness	Box E

Step 2. Determining if a study meets the standards for good methodological quality
A. Reliability n=46

Design requirements	Excellent	Good	Fair	Poor
Was the percentage of missing subjects given? none	Percentage of missing subjects described	Percentage of missing subjects NOT described		
Was there a description of how missing subjects were handled? none	Described how missing subjects were handled	Not described but it can be deduced how missing subjects were handled	Not clear how missing subjects were handled	
Was the sample size included in the analysis adequate?	Adequate sample size (≥100)	Good sample size (50-99)	Moderate sample size (30-49)	Small sample size (<30)
Were at least two measurements available? Intra and inter intra only	At least two measurements			Only one measurement
Were the administrations independent? Intra	Independent measurements	Assumable that the measurements were independent	Doubtful whether the measurements were independent	Measurements NOT independent
Was the time interval stated? Intra	Time interval stated		Time interval NOT stated	
Were patients stable in the interim period on the construct to be measured? Intra	Patients were stable (evidence provided)	Assumable that patients were stable	Unclear if patients were stable	Patients were NOT stable
Was the time interval appropriate? Intra	Time interval appropriate		Doubtful whether time interval was appropriate	Time interval NOT appropriate
Were the test conditions similar for both measurements? e.g. type of administration, environment, instructions	Test conditions were similar (evidence provided)	Assumable that test conditions were similar	Unclear if test conditions were similar	Test conditions were NOT similar
Were there any important flaws in the design or methods of the study? Intra	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study
for continuous scores: Was an	ICC calculated	ICC calculated	Pearson or	~~No ICC or~~

intraclass correlation coefficient (ICC) calculated?	and model or formula of the ICC is described	but model or formula of the ICC not described or not optimal. Pearson or Spearman correlation coefficient calculated with evidence provided that no systematic change has occurred	Spearman correlation coefficient calculated WITHOUT evidence provided that no systematic change has occurred or WITH evidence that systematic change has occurred	~~Pearson or Spearman correlations calculated~~
~~For dichotomous/nominal/ordinal scores: Was kappa calculated?~~	~~Kappa calculated~~			~~Only percentage agreement calculated~~
~~for ordinal scores: Was a weighted kappa calculated?~~	~~Weighted Kappa calculated~~		~~Unweighted Kappa calculated~~	~~Only percentage agreement calculated – actually TEM calculated~~
~~for ordinal scores: Was the weighting scheme described? e.g. linear, quadratic~~	~~Weighting scheme described~~	~~Weighting scheme NOT described~~	~~With TEM- not sure how to answer this~~	

C. Hypothesis Testing / Construct Validity n= 46

Design requirements	Excellent	Good	Fair	Poor
Was the percentage of missing items given? none	Percentage of missing items described	Percentage of missing items NOT described	NA	NA
Was there a description of how missing items were handled?	Described how missing items were handled- all in final analysis	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
Was the sample size included in the analysis adequate?	Adequate sample size (≥100 per analysis)	Good sample size (50-99 per analysis)	Moderate sample size (30-49 per analysis)	Small sample size (<30 per analysis)
Were hypotheses regarding correlations or mean differences formulated a	hypotheses formulated a priori	Minimal number of hypotheses formulate a priori	Hypotheses vague or not formulated but possible to deduce what was expected	Unclear what was expected

priori (i.e. before data collection)?				
Was the expected *direction* of correlations or mean differences included in the hypotheses?	Expected direction of the correlations or differences stated-	Expected direction of the correlations or differences NOT stated –		
Was the expected absolute or relative *magnitude* of correlations or mean differences included in the hypotheses?	Expected magnitude of the correlations or differences stated	Expected magnitude of the correlations or differences NOT stated		
for convergent validity: Was an adequate description provided of the comparator instrument(s)?	Adequate description of the constructs measured by the comparator instrument(s)	Adequate description of most of the constructs measured by the comparator instrument(s)	Poor description of the constructs measured by the comparator instrument(s)	NO description of the constructs measured by the comparator instrument(s)
for convergent validity: Were the measurement properties of the comparator instrument(s) adequately described?	Adequate measurement properties of the comparator instrument(s) in a population similar to the study population-	Adequate measurement properties of the comparator instrument(s) but not sure if these apply to the study population	Some information on measurement properties (or a reference to a study on measurement properties) of the comparator instrument(s) in any study population	No information on the measurement properties of the comparator instrument(s)
Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study (e.g. only data presented on a comparison with an instrument that measures another construct)	Other important methodological flaws in the design or execution of the study –
Were design and statistical methods adequate for the hypotheses to be tested?	Statistical methods applied appropriate	Assumable that statistical methods were appropriate, e.g. Pearson correlations applied, but distribution of scores or mean (SD) not presented	Statistical methods applied NOT optimal- correlation coefficients would be more appropriate	Statistical methods applied NOT appropriate