

**Appendix SI.** Definitions and rating of measurement properties

Domain	Measurement property	Definition	Measurement property rating	
Reliability	Reliability	The extent to which scores of an aerobic capacity measure are the same in patients who have not changed for repeated measurement under several conditions: e.g. comparing different scores of aerobic capacity over time (test-retest); by different persons on the same occasion (interrater); or by the same persons (i.e. raters or responders) on different occasions (intra-rater).	+	ICC or weighted kappa $\geq 0.70$
	Measurement error	The systematic and random error of a patient's aerobic capacity score that is not attributed to true changes in aerobic capacity.	-	ICC or weighted kappa $< 0.70$
Validity	Content validity	The degree to which the content of a measurement property is an adequate reflection of the construct to be measured. For a maximal exercise test, content validity is assessed by the degree to which the results of the test protocol reflect true $VO_{2\text{peak}}$ . Here we determined which percentage of patients reached pre-set criteria for achieving maximal aerobic exercise.	+	$\geq 75\%$ of patients meet $VO_{2\text{peak}}$ criteria
	Criterion validity	The degree to which the aerobic capacity score is an adequate reflection of a "gold standard". E.g. for $VO_{2\text{peak}}$ , the gold standard is GXT with respiratory gas exchange measurements on a treadmill or bicycle/arm ergometer.	-	Correlation with gold standard $\geq 0.70$
	Construct validity	The degree to which the scores of an aerobic capacity measure are consistent with hypotheses (e.g. the scores of an aerobic capacity measure discriminate between physically active and physically inactive patients).	-	Correlation with gold standard $< 0.70$
Responsiveness	Responsiveness	The ability of a measure of aerobic capacity to detect changes in aerobic capacity over time.	+	The result is in accordance with the hypothesis
			-	The result is not in accordance with the hypothesis
			+	$\geq 75\%$ of the results are in accordance with the hypotheses OR AUC $\geq 0.70$
			-	$\leq 75\%$ of the results are in accordance with the hypotheses OR AUC $\geq 0.70$

+: sufficient rating; -: insufficient rating; ICC: intraclass correlation coefficient; SDC: smallest detectable change; LoA: limits of agreement; MIC: minimal important change;  $VO_{2\text{peak}}$ : peak oxygen consumption; GXT: maximal effort graded exercise testing; AUC: area under the curve.

Ovid MEDLINE(R) AII < 1946 to June 29, 2021 > Search date: 30 June 2021

---

*(Continued)*

## **Appendix SII. (Continued) Literature search**

---

(Continued)

Ebscohost SPORTdiscus Search date: 30 June 2021

(Continued)

Web of ScienceSearch date: 30 June 2021

**Appendix SIII.** Content validity risk of bias assessment checklist

	Very good	Adequate	Doubtful	Inadequate
1. Was an appropriate protocol used for the maximal exercise test? (Start with warm-up phase, duration of test in general between 6 and 18 min, increase in workload of 5–25 watts per min, protocol well described)	Widely recognized or well justified method used	Assumable that the method was appropriate, but not described in detail	Doubtful whether the method was appropriate, or protocol not (clearly) described	Method used not appropriate
2. Were appropriate criteria used for assessing the content validity of the maximal exercise test? (Recommended criteria for maximal effort are 1) RER ≥ 1.1, 2) HR > 90% predicted max, 3) Patient exhaustion/Borg scale ≥ 9 (range 1–10), ≥ 17 (range 6–20) and 4) a plateau in VO <sub>2</sub> (25, 26)	Multiple widely recognized or well justified criteria used, a minimum of 2 criteria have to be achieved ≥ 50	Multiple widely recognized or well justified criteria used, only one criterion has to be achieved ≥ 30	Doubtful whether the criteria were appropriate < 30 or not clear	Criteria used not appropriate or not well described -
3. Was the content validity tested in an appropriate number of patients?				
4. Was skilled personnel used to conduct the test?	An experienced test supervisor conducted the test	The test supervisor had limited experience or was trained specifically for the study	Not clear if the test supervisor was trained or not trained and/or had no experience	-
5. Was an appropriate approach used to analyse the data?	A widely recognized or well justified approach was used	Assumable that the approach was appropriate, but not clearly described	Not clear what approach was used or doubtful whether the approach was appropriate	Approach not appropriate
6. Were there any other important flaws in the design or methods of the study	No other important methodological flaws	-	Other minor methodological flaws	Other important methodological flaws

**Appendix SIV.** Criterion validity risk of bias assessment checklist

	Very good	Adequate	Doubtful	Inadequate
1. Can the criterion used or employed be considered as a reasonable “gold standard”? (VO <sub>2peak</sub> : Maximal exercise test with respiratory gas analysis. Anaerobic threshold: Lactate measurement or V-slope method)	Widely recognized or well justified gold standard used	-	Doubtful whether the gold standard was appropriate, or golden standard not clearly described	Gold standard used not appropriate
2. Were correlations, or the area under the curve calculated?	Correlations or AUC calculated	-	-	Correlations or AUC NOT calculated
3. Were there any other important flaws in the design or methods of the study	No other important methodological flaws	-	Other minor methodological flaws	Other important methodological flaws

**Appendix SV.** Modified GRADE approach for grading the quality of the evidence

Quality of evidence	Lower if
High	Risk of bias
Moderate	-1 Serious
Low	-2 Very serious
Very low	-3 Extremely serious
	Inconsistency
	-1 Serious
	-2 Very serious
	Imprecision
	-1 total $n=50-100$
	-2 total $n < 50$
	Indirectness
	-1 Serious
	-2 Very serious
High	We are very confident that the true measurement property lies close to that of the pooled or summarized result of the measurement property reported in the studies
Moderate	We are moderately confident in the pooled or summarized result of the measurement property found in the studies: the true measurement property is likely to be close to the measurement property reported in the studies, but there is a possibility that it is substantially different
Low	Our confidence in the pooled or summarized result of the measurement property reported in the studies is limited: the true measurement property may be substantially different
Very Low	We have very little confidence in the pooled or summarized result of the measurement property reported in the studies: the true measurement property is likely to be substantially different