International Journal of Sports Physiology and Performance, 2009, 4, 269-277

Test Validation in Sport Physiology: Lessons Learned From Clinimetrics

Franco M. Impellizzeri and Samuele M. Marcora

We propose that physiological and performance tests used in sport science research and professional practice should be developed following a rigorous validation process, as is done in other scientific fields, such as clinimetrics, an area of research that focuses on the quality of clinical measurement and uses methods derived from psychometrics. In this commentary, we briefly review some of the attributes that must be explored when validating a test: the conceptual model, validity, reliability, and responsiveness. Examples from the sport science literature are provided.

Keywords: physiological testing, measurement, validity, reliability, responsiveness, sport

The use of laboratory and field tests is common in sport physiology and, in recent years, the number of these tests has increased exponentially. To identify the tests with the best measurement properties, we propose the same rigorous and comprehensive approach used in "Clinimetrics."[1–3] This area of research focuses on the quality of clinical measurement[4] and is based on well-established psychometric methods developed in psychology, sociology, and education.[5,6] The Scientific Advisory Committee of the Medical Outcomes Trust for Health Status and Quality of Life instruments have proposed eight attributes that warrant consideration in the evaluation of instruments measuring health status and quality of life: 1) conceptual and measurement model; 2) validity; 3) reliability; 4) responsiveness; 5) interpretability; 6) respondent and administrative burden; 7) alternative forms; 8) cultural and language adaptation.[2] Before a clinical test can be proposed for use in research and professional practice, *all* these attributes need to be adequately verified. Unfortunately, in sport physiology, many tests have been validated against only few of these attributes, for example, some forms of validity and reliability. One of the barriers for the implementation of a more comprehensive approach is the difficulty in appreciating and understanding methods originally developed for subjective measures of psychosocial constructs. The aim of this commentary is to briefly explain these attributes using, as examples, studies investigating objective measures of physical performance in soccer and endurance exercise performance.

Impellizzeri is with the Dept. of Research and Development, Schulthess Klinik, Zürich, Switzerland, and CeBiSM, Inter-University Research Center of Bioengineering and Sport Science, Rovereto, Italy. Marcora is with the School of Sport, Health, and Exercise Sciences, Bangor University, Wales, U.K.

Conceptual Model

The Scientific Advisory Committee defines the conceptual model as "a rationale for and description of the concepts and the population that a measure is intended to assess and the relationship between those concepts."[2] In terms of rationale, sport physiologists should carefully consider the purpose of a new test, for example, selection and/or longitudinal assessment. In addition, to avoid the development of redundant tests, sport physiologists should also convincingly justify why a new test is needed. In this respect, the definition of the concept that a measure is intended to assess is crucial.

As suggested by Atkinson,[7] the conceptualization of a multifactorial construct facilitates the definition of its measurable components. In Figure 1 we present the current simplified model of soccer performance (multifactorial construct), which can be measured using, for example, the final ranking in a championship. Three constructs relevant for soccer performance are i) tactical (interaction with other individuals), ii) technical (individual skills), and iii) physical performance. The relevance of these constructs is based on the assumptions that they influence soccer performance. To investigate physical performance sport physiologists commonly use the amount of high-intensity activity completed during a match. In clinimetric jargon, high-intensity activity during the match would be considered a causal indicator of the theoretical construct *physical performance*. The validity of this parameter has been supported by a study showing that professionals playing soccer in countries in the top positions of the FIFA ranking cover more distance at high-intensity than those playing in lower level championships.[8] To confirm that physical performance has a cause-effect relationship with soccer performance, it would be necessary to increase the capacity to perform high-intensity activities and determine if this manipulation improves the final ranking in a championship. This type of experimental study would be very difficult to conduct, but a model of soccer performance based on this kind of evidence would not suffer from the shortcomings of theoretical frameworks built only on correlational studies. For example, we have recently shown that high-intensity distance is not different between players of the more successful teams (ranked in the first five positions) with the players of the less successful teams (ranked in the last five positions) from the same league. The only difference was found for high-intensity activities with the ball.[9] This seems to suggest that the interaction between physical performance and technical components may be a better predictor of soccer performance. Similarly, the repeated sprint ability (RSA) test performance (which is correlated to the high-intensity activity completed during a match[10]) differentiates between professionals and amateurs, but not between top and mid-level professional players.[11] These latter findings suggest that physical performance may simply reflect the fact that professionals train more than amateur players rather than being causally related to soccer performance. Therefore, at present, the theoretical framework behind the validity of physical performance tests in soccer is not convincing, and we believe that much more research is necessary to confirm current models of soccer performance or develop more valid ones. Moreover, the complex relationships across physical, technical, and tactical performance may be more important than the individual constructs alone in determining soccer performance. If confirmed in future multivariate analyses, these complex relationships should be

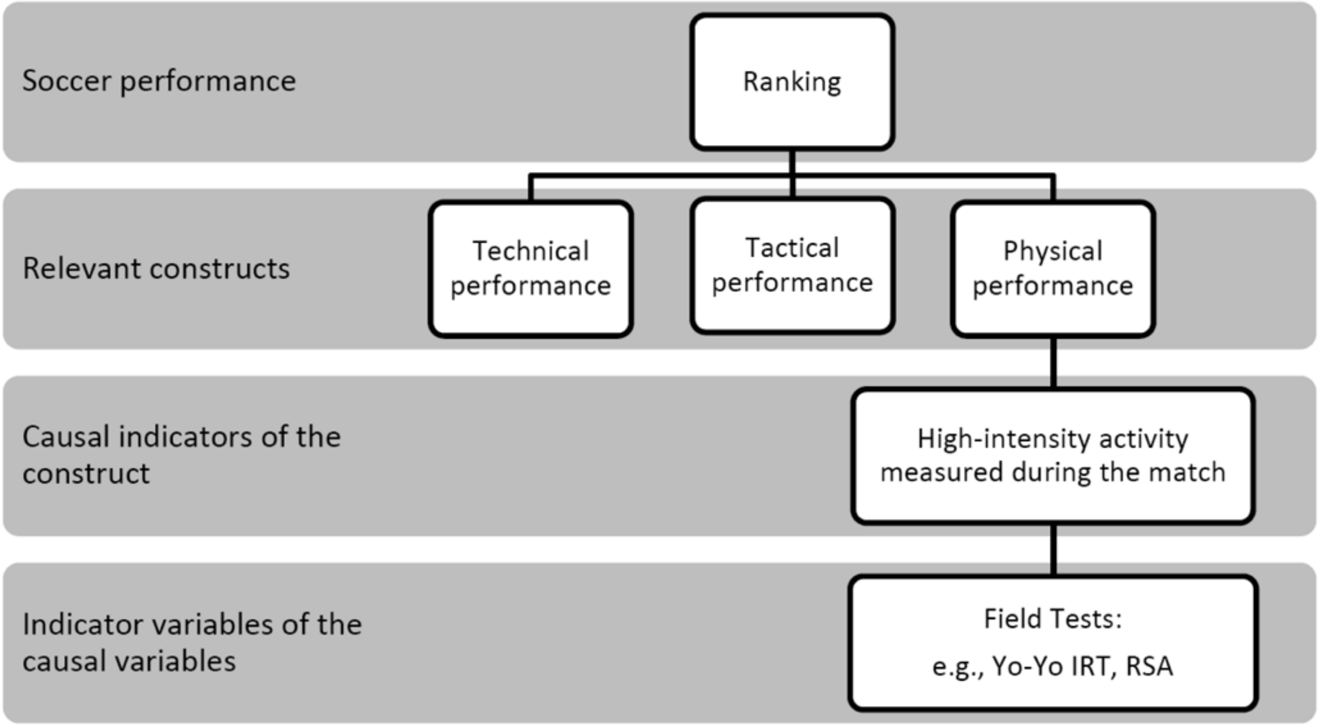

Figure 1 — Current theoretical framework of soccer performance.

taken into account when developing new tests for soccer players. Otherwise, we risk measuring variables not really relevant for performance.

Reliability, or Reproducibility

Atkinson and Nevill[12] have reported two types of reliability: absolute reliability (degree to which repeated measurements vary for individuals) and relative reliability (degree to which individuals maintain their position in a sample with repeated measurements). Other authors have used the terms *agreement* and *reliability*[13] when referring to absolute reproducibility and relative reproducibility, respectively. Terminology aside, the type of reproducibility to consider depends on the purpose of the test. When tests are used to discriminate among individuals (cross-sectional assessment), parameters of relative reliability should be used (intraclass correlation coefficient, ICC). Parameters of absolute reliability (eg, standard error of measurement, SEM) are required for evaluative tests to monitor changes over time (longitudinal assessment).[13] Although ICC and SEM are related, they provide different information becuase ICC is influenced by various factors, such as the variability among subjects and measurement error, whereas SEM, by error variation only.[13–15] Clearly, the distinction between absolute and relative reliability should be considered when reporting reliability data.

Reliability data are commonly used to decide whether a test should be employed or not. For example, in recent years, time trials have been favored over time-to-exhaustion tests as measures of endurance exercise performance because they have much lower coefficients of variation.[16] This is a mistake, as will be discussed in the responsiveness section of this article. A better use of reliability data is to calculate the minimal detectable change (ie, the minimal individual change that can be interpreted as real with an acceptable probability level). These are extremely important data when testing individual athletes. In clinimetrics, it has been proposed that the minimal detectable change should be lower that the smallest worthwhile change.[1] For sport performance, a method integrating the minimal important change, the smallest detectable change, and the test reliability has been proposed and described by Hopkins,[17] and we have applied it in a recent study.[11] For example, the mean time during an RSA test we developed for soccer players has a SEM of about 0.8% to 1.0%.[11] Such values would be commonly interpreted as showing very good reliability. However, since the smallest worthwhile change for this test was 0.5%, the ability to detect small but worthwhile changes in individual RSA performance is poor, thus limiting its use as an evaluative test in professional practice. In other words, the noise (eg, SEM) should be lower than the smallest worthwhile signal. As aforementioned, in sport science, sound methods have been developed to determine the probability that individual changes in a test are greater than the smallest worthwhile changes given the test reliability.[11,18] Unfortunately, despite the availability on the Web of practical spreadsheets (www.sportsci.org), these methods are not widely applied in sport physiology yet.

Validity

Validity is the degree to which the test measures what it purports to measure.[2] Several studies are required to build up a body of evidence to support the validity of a test. Such evidence can be based on the inherent characteristics of the test (content and logical validity), its relation to a criterion (predictive, concurrent, or postdictive validity), or a construct (convergent and divergent validity, known-groups difference technique).[5] *Construct validity* is often used as an overarching term encompassing all types of validity.[19] This validation is theory dependent, and, therefore, the conceptual model is critical to operationally define the constructs of interest and their measurable indicators. The selection of the most appropriate method for validating a test also depends on its purpose (discriminative or evaluative) and its application (research or routine practice).

If the aim of the test is to select athletes, it should be able to discriminate individuals of different competitive levels. In sport physiology, this type of validity is commonly established by testing differences between groups of players of different competitive levels and/or playing positions. In clinimetrics, alternative methods such as the receiver operator characteristics (ROC) curve are gaining popularity and can be used to validate the discriminant ability (and the responsiveness) of physiological and performance tests.[20,21] For example, we found that professional players have better RSA test performance compared with amateur players.[11] Using the same data, we have calculated the area under the ROC curve, which is 0.89 (CI 95%, 0.813 to 0.940, $P < .0001$; Figure 2). Values above 0.70 are commonly considered to indicate good discriminant ability.[20,21] The area under the ROC curve represents the probability of correctly discriminating professional from amateur players using the RSA mean time. The test score able to distinguish between these competitive levels is 7.37 s. This cut-off value gives a "true-positive rate" (sensitivity) of 88% and a "false-positive rate" (1 − specificity) of 76%. Therefore, this type of statistical analysis suggests that the RSA test has excellent discriminant ability if its purpose were to differentiate between professional and amateur soccer players. However, is it practically useful to make this differentiation? This example shows once more how crucial it is to have a sound theoretical framework for assessing the validity of a test, regardless of how sophisticated or novel the statistical analysis.

Unfortunately, the cross-sectional methods described above are often used to validate tests used in sport physiology to assess changes over time. However, discriminant ability is not sufficient or even relevant to evaluative tests. These tests should be validated against a criterion (ie, a "gold standard") or an indicator of the construct of interest. A correlation larger than 0.70 between the new test and the reference measure is conventionally used as benchmark for construct or criterion validity.[1] However, benchmarks should not be always taken too strictly and a correlation of 0.65 instead of 0.70 cannot be interpreted as evidence against construct or criterion validity. Furthermore, the confidence interval of these relations should be also taken into consideration. To understand if a certain value is acceptable or not, it is important to understand the kind of validity we are examining. For example, while correlations higher than 0.70 can be acceptable for providing convergent evidence of construct validity, they are certainly not appropriate for predictive validity, such as when we want to use a field test to estimate the actual maximal

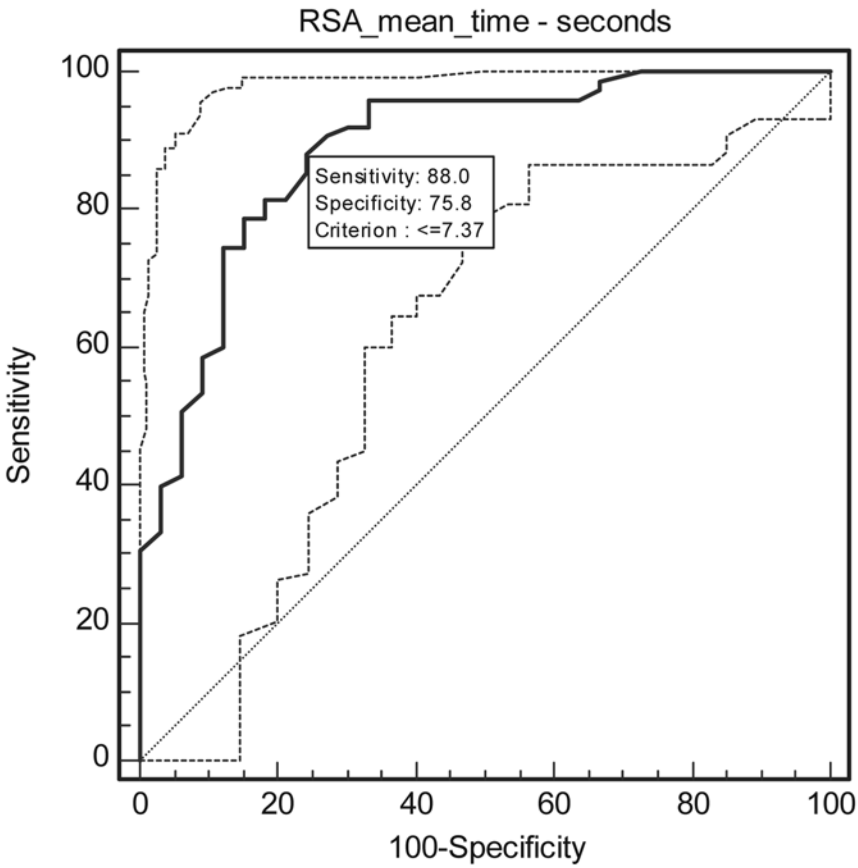

Figure 2 — Receiver operating characteristics curve for the mean time taken to complete six sprints interspersed by 20 s of recovery (RSA test), showing its ability to discriminate professional from amateur players.

oxygen uptake and even more so if we want to estimate this at an individual level. However, as we stated earlier, several studies are required to build up a body of evidence for or against the validity of a test. Indeed, after cross-sectional evidence of construct validity has been provided, another extremely important step is to assess the longitudinal validity of the test. Such a property, also termed *external responsiveness*, is the ability of a test to measure changes in the reference measure.[22] Changes in a test displaying external responsiveness would correlate with changes in the criterion or indicator of the construct of interest. For example, in soccer referees, the high correlation found between improvement in the yo-yo test and improvement in high-intensity activity during a match ($r = .77$) proves the longitudinal validity of the yo-yo test in this population.[23] To the best of our knowledge, this is the only example of longitudinal validation of a physical performance test in soccer.

Responsiveness

In clinimetrics, responsiveness is considered the most essential property of an evaluative instrument.[24] Responsiveness is classified as external (see the previous section on longitudinal validity) and internal. The latter is also called *sensitivity to change* and refers to the ability of a measure to change over a particular time frame.[22,25] There are different methods that can be used to calculate internal responsiveness, most of which are variations of the same methods.[24] In short, these methods are i) Cohen's effect size, ii) the standardized response mean, and iii) Guyatt's responsiveness index.[22] To the best of our knowledge, only one study has formally evaluated internal responsiveness in sport physiology.[26] In this study, the internal responsiveness of a time trial and a time-to-exhaustion test of endurance performance was compared in a group of competitive cyclists breathing hypoxic or hyperoxic gas mixtures. A variation of standardized response mean was used to assess the sensitivity to change in endurance performance compared with normoxia. In spite of suggestions that time trials would have better "signal-to-noise ratio" because of lower coefficient of variation (ie, noise),[16] the internal responsiveness of the time-to-exhaustion test was as good as that of the time trial. This finding was expected because the change in performance caused by hypoxia and hyperoxia (ie, signal) was much larger for the time-to-exhaustion test. These results clearly demonstrate that, when choosing an evaluative test for an experimental study, internal responsiveness should be considered rather than reliability parameters alone.

Other characteristics should be evaluated when developing or selecting a test. In clinimetrics, respondent and administrative burden is often taken into consideration because this may have an impact on test acceptability, adherence, and motivation.[2] For example, contrary to time trials, the aforementioned time-to-exhaustion test does not require familiarization trials.[26] Similarly, soccer match simulations can be useful for research purposes but less practical for the routine evaluation of the players**.** Interpretability is another important test characteristic. This is achieved by comparing individual or group results to normative data and the minimal important and detectable changes. The reader is referred to recent articles for suggestions on how to derive the probability that a change in a test is real and worthwhile.[11,18]

In conclusion, we believe that the application of more rigorous methods for the development and validation of physiological and performance tests would improve the quality of sport science research and professional practice. A similar approach has been also recently advocated for the quality of intervention studies in sport science.[27,28] As often happens in statistics, for each of the eight attributes, there is still debate about the best analytical approach. For example, in clinimetrics and psychometrics, the best methods to examine responsiveness have yet to be established.[22,24,25] However, statistical debates should not dissuade sport physiologists from applying more rigorous approaches for test validation, such as in clinimetrics and psychometrics. We hope that future investigations will assess all the relevant test attributes presented here, rather than leading to further test proliferation.

References

1. Terwee CB, Bot SD, de Boer MR, et al. Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol*. 2007;60:34–42.
2. Scientific Advisor Committee. Medical Outcomes Trust. Assessing health status and quality-of-life instruments: attributes and review criteria. *Qual Life Res*. 2002;11:193–205.
3. Pollard B, Johnston M, Dixon D. Theoretical framework and methodological development of common subjective health outcome measures in osteoarthritis: a critical review. *Health Qual Life Outcomes*. 2007;5:14.
4. de Vet HC, Terwee CB, Bouter LM. Current challenges in clinimetrics. *J Clin Epidemiol*. 2003;56:1137–1141.
5. Ary D, Cheser Jacobs L, Razavieh A, Sorensen C. *Introduction to Research in Education*. Belmont, CA: Wadsworth; 2006.
6. Coolican H. *Research Methods and Statistics in Psychology*. London: Hodder & Stoughton; 2004.
7. Atkinson G. Sport performance: variable or construct? *J Sports Sci*. 2002;20:291–292.
8. Mohr M, Krustrup P, Bangsbo J. Match performance of high-standard soccer players with special reference to development of fatigue. *J Sports Sci*. 2003;21:519–528.
9. Rampinini E, Impellizzeri FM, Castagna C, Coutts AJ, Wisløff U. Technical performance during soccer matches of the Italian Serie A league: effect of fatigue and competitive level. *J Sci Med Sport*. 2009;12(1):227-233.
10. Rampinini E, Bishop D, Marcora SM, Ferrari Bravo D, Sassi R, Impellizzeri FM. Validity of simple field tests as indicators of match-related physical performance in top-level professional soccer players. *Int J Sports Med*. 2007;28:228–235.
11. Impellizzeri FM, Rampinini E, Castagna C, et al. Validity of a Repeated-Sprint Test for Football. *Int J Sports Med*. 2008;29:899–905.
12. Atkinson G, Nevill AM. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Med*. 1998;26:217–238.
13. de Vet HC, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. *J Clin Epidemiol*. 2006;59:1033–1039.
14. Stratford PW, Goldsmith CH. Use of the standard error as a reliability index of interest: an applied example using elbow flexors strength data. *Phys Ther*. 1997;77:745–750.
15. Weir JP. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *J Strength Cond Res*. 2005;19:231–240.
16. Jeukendrup A, Saris WH, Brouns F, Kester AD. A new validated endurance performance test. *Med Sci Sports Exerc*. 1996;28:266–270.
17. Hopkins WG, Hawley JA, Burke LM. Design and analysis of research on sport performance enhancement. *Med Sci Sports Exerc*. 1999;31:472–485.
18. Batterham AM, Hopkins WG. Making meaningful inferences about magnitudes. *Int J Sports Physiol Perform*. 2006;1:50–57.
19. Westen D, Rosenthal R. Quantifying construct validity: two simple measures. *J Pers Soc Psychol*. 2003;84:608–618.
20. Deyo RA, Centor RM. Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test performance. *J Chronic Dis*. 1986;39:897–906.
21. Mannion AF, Elfering A, Staerkle R, et al. Outcome assessment in low back pain: how low can you go? *Eur Spine J*. 2005;14:1014–1026.
22. Husted JA, Cook RJ, Farewell VT, Gladman DD. Methods for assessing responsiveness: a critical review and recommendations. *J Clin Epidemiol*. 2000;53:459–468.

23. Bangsbo J, Iaia FM, Krustrup P. The yo-yo intermittent recovery test : a useful tool for evaluation of physical performance in intermittent sports. *Sports Med*. 2008;38:37–51.
24. Terwee CB, Dekker FW, Wiersinga WM, Prummel MF, Bossuyt PM. On assessing responsiveness of health-related quality of life instruments: guidelines for instrument evaluation. *Qual Life Res*. 2003;12:349–362.
25. Epstein RS. Responsiveness in quality-of-life assessment: nomenclature, determinants, and clinical applications. *Med Care*. 2000;38:II91–II94.
26. Amann M, Hopkins WG, Marcora SM. Similar sensitivity of time to exhaustion and time-trial time to changes in endurance. *Med Sci Sports Exerc*. 2008;40:574–578.
27. Atkinson G, Batterham A, Drust B. Is it time for sports performance researchers to adopt a clinical-type research framework? *Int J Sports Med*. 2008;29:703–705.
28. Bishop D. An applied research model for the sport sciences. *Sports Med*. 2008;38:253–263.