

# Measurement Issues in Concussion Testing

Brian G. Ragan, PhD, ATC • University of Northern Iowa  
Minsoo Kang, PhD • Middle Tennessee State University

**M**EASUREMENTS must be valid and reliable to make critical clinical decisions that are based on test scores. Recently, the usefulness of neuropsychological (NP) tests in the management of sports-related concussions has been questioned.<sup>1</sup> Randolph et al.<sup>1</sup> concluded that no NP tests had established evidence in key areas to support their clinical application. The authors admitted that there is a theoretical rationale for the use of NP tests in concussion management, but they did not provide reasons that the data derived from them do not support clinical use. The key criteria used to determine usefulness were components validating existing NP tests. Some of these tests were designed for other purposes, and we are forcing them to fit a sports medicine application.

An NP test in sports must function as a measure that can identify immediate changes in cognitive function and then provide an indication of changing status throughout the recovery period. The overall poor performance of NP tests, including paper-and-pencil tests and some computer tests, may be influenced by more fundamental issues relating to the design of the test. Many existing NP tests have been constructed for populations other than athletes with sports-related concussion and are administered in an unintended manner (serial testing). The purpose of this report is to provide a critical review of design assumptions and related psychometric issues that may contribute to the poor performance of NP tests. It is also important to note that these measurement issues may be relevant to other methods used in the assessment of concussion, such as mental status ratings, symptom scales, and postural stability measures.

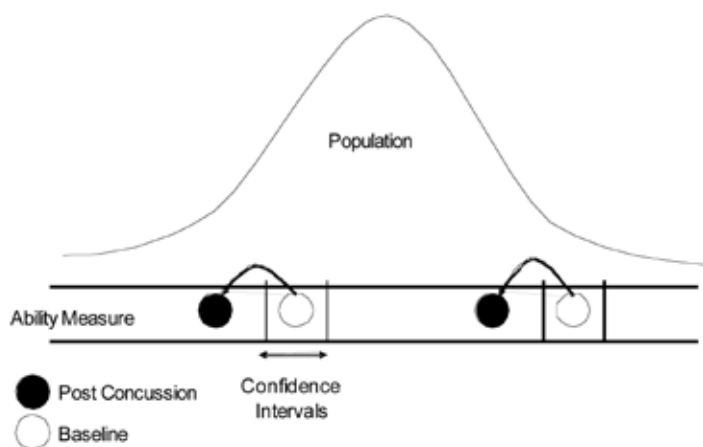
## Design Assumptions and Psychometric Problems

To fully understand the critical psychometric problems of concussion tests, several major measurement concepts must be addressed, including reference standard and interpretations, ceiling effects, and other psychometric properties.

### Reference Standard and Interpretations

There are two kinds of test practice and score interpretations employed in concussion testing. The most popular practice is the individual-centered standard, which is based on the assumption that the latent ability is normally distributed and that the test has the ability to discriminate well at all ability levels. Thus, an individual-centered standard is very much like a norm-referenced standard, which involves comparison of a score to a normative value that is based on a large sample. The difference in concussion testing is that the score is not compared with a normative value, but rather the preconcussion score of the individual. In Figure 1, an individual-centered standard for interpretation is used to identify change.

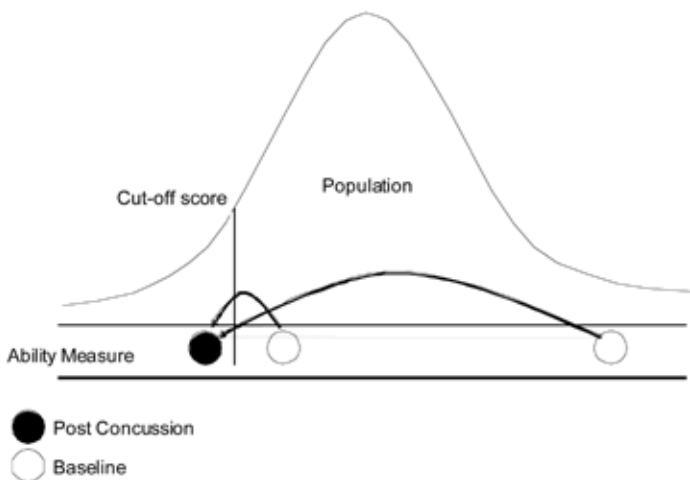
In the example, two individuals of different ability levels were baseline tested, and confidence intervals were calculated for the observed performance scores. Both experienced a concussion, and the postconcussion test is compared to the baseline reference point to assess decline in cognitive function. This is done by comparing the postconcussion test score to the baseline (preconcussion) score and the confidence interval. If the postconcussion score falls outside the confidence interval limit, then change is confirmed. The important



**Figure 1** Depiction of a normal distribution of a population's ability and the use of an individual-centered standard for postconcussion test score interpretation (2 individuals with different baseline test scores and different postconcussion test scores).

issue with an individual-centered standard is its degree of accuracy for identification of improved status, on which a decision about fitness for return to sports may be based.

The second practice in concussion testing is the criterion-referenced standard, most recently proposed by McCrea et al.,<sup>2</sup> which does not require baseline testing. This method uses a population cut-off score to identify group membership. A criterion-referenced standard does not involve comparison of an individual's performance to that of others who have also been tested. Rather, it compares the individual's performance to a well-defined performance standard for interpretation of the score's meaning.<sup>3</sup> A criterion-referenced standard may be used to identify the existence of psychopathology or to define an acceptable skill level in a specified psychological domain.<sup>4</sup> The development of a criterion-referenced standard for concussion assessment is based on the assumption that any concussion will produce NP dysfunction beyond a certain threshold of injury severity. The authors who proposed this method later changed their advocated approach to the individual-centered standard.<sup>5,6</sup> It is important to note that this change appears to have been the result of the NP test's poor performance and not the result of a systematic investigation. McCrea et al.<sup>5</sup> reported that the 25 point threshold accurately identified mild traumatic brain injury (MTBI) on the Standardized Assessment of Concussion. Currently, most NP tests have not addressed this conceptual dilemma for score interpretation.



**Figure 2** Depiction of the use of a criterion-referenced standard (cut-off) for postconcussion test score interpretation (2 individuals with different baseline test scores, but the same postconcussion test score).

Figure 2 further illustrates the use of a criterion-referenced standard. Two people of different baseline ability levels (indicated by the white circles) experience a concussion. Both individuals on the postconcussion testing had similar scores. The important aspect of the criterion-referenced standard is that the baseline ability level does not influence the interpretation of the postconcussion test results. The dysfunction resulting from the concussion causes the injured athletes to score below a level that identifies an appropriate cut-off score. This cut-off score could then be used without baseline tests to indicate dysfunction. Both individuals in this example scored below the cut-off score (vertical line), indicative of cognitive dysfunction.

Criterion-referenced score interpretation for clinical decision-making does not lack evidence of clinical benefit. Perhaps the most widely used scale for quantification of injury severity is the Glasgow Coma Scale (GCS), which measures level of consciousness and uses criterion-referenced interpretation.<sup>7</sup> Most emergency medicine professionals regularly use this scale to categorize a patient's level of consciousness. The GCS is not a NP test, but it does use a criterion-referenced approach and multiple cut-off scores for the classification of traumatic brain injuries as mild, moderate, and severe. GCS scores between 13 and 15 are considered to represent a range of consciousness states that includes both a normal state and the state associated with mild traumatic brain injury. Thus, the GCS is not sufficiently sensitive to differentiate between patients experiencing mild cognitive impairment from

MTBI and those who do not have an impairment. This limitation has led to the development of NP tests, which are intended to have sufficient sensitivity to detect small changes in cognitive function.

Surprisingly, the type of reference standard that should be used to interpret the extent of cognitive dysfunction associated with the NP test score of an athlete has sustained a concussion has not received very much consideration. The use of either an individual-centered standard or a criterion-referenced standard varies among the existing NP tests. This fundamental concept dictates the manner in which the test is constructed and the manner in which the resulting score is interpreted.

### Test Assumptions for Interpretation Standards Not Satisfied

The individual-centered standard and the serial testing protocol used today present unique psychometric demands that must be considered during scale development and score evaluation. The Concussion in Sports Group<sup>8</sup> has identified the individual-centered standard as the fundamental basis of NP testing, which requires baseline preconcussion testing and serial postconcussion follow-up testing. This approach is based on the assumption of normally distributed ability levels among the members of the population, a distribution characterized by a bell-shaped curve. Reaction time measurements support the individual-centered standard for assessment of concussion effect, because baseline scores are normally distributed.<sup>9</sup> The test must accurately measure all ability levels, which is directly related to test-retest reliability.

There appears to be a significant discrepancy between the published results of NP tests and the basis for the recommendation to utilize an individual-centered standard. It seems that most NP test developers selected items that should be used for construction of a criterion-referenced test. The nature of the questions selected are basic and the degrees of item difficulty do not represent a wide range of ability levels. The expectation of a postconcussion decrease in performance may have influenced the development strategy. This problem may impose severe limitations on the accuracy of preconcussion baseline scores. Lacking a valid baseline measure, no valid interpretation of postconcussion measures can be derived from the individual-centered standard.

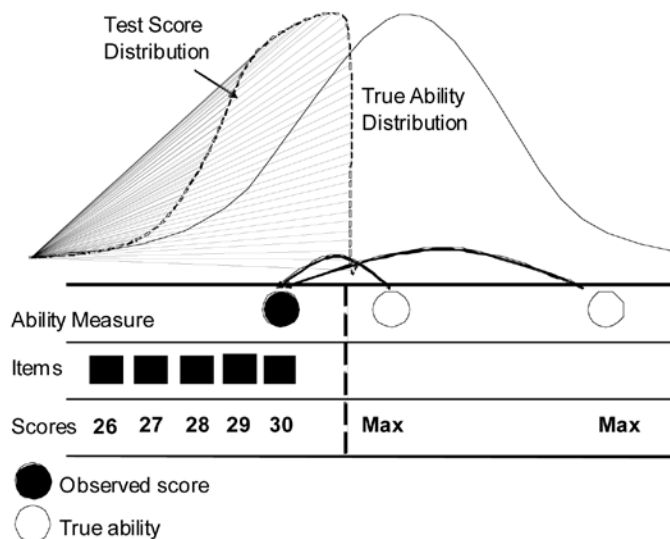
Although some tests are designed for strict score interpretation with either an individual-centered standard or a criterion-referenced standard, it is possible

to successfully construct a test that can be interpreted using either standard. Through prospective research and systematic clinical observations, an individual-centered standard for score interpretation can facilitate development of a criterion-referenced interpretation.<sup>4</sup> This can be achieved by accurately measuring all ability levels among members of the population. As the understanding of the range of ability levels increases, characteristics associated with various score levels will support development of a criterion-referenced score interpretation. The potential use of both interpretation standards is needed for assessment of cognitive changes attributable to MTBI. The test must identify a dysfunctional state and then measure the amount of change in the dysfunctional state periodically throughout the recovery period. Relative performance of the individual in relation to the ability continuum and accuracy in classification of cognitive change must both be addressed for optimal benefit to be derived from NP tests. A major limitation of NP tests may be attributable to ceiling effects.

### Ceiling Effects

It is common for some sections of popular NP tests to present significant ceiling effects.<sup>10</sup> A ceiling effect is problematic, because it indicates that the test cannot accurately measure every individual's true baseline ability (i.e., the upper-level baseline ability of some individuals is underestimated). A ceiling effect presents a major problem for use of the individual-centered standard and serial testing protocol used by some NP tests. How can underestimated baseline values be used to accurately quantify dysfunction? An example of this problem presented by a ceiling effect is illustrated in Figure 3. In this example, two individuals having different latent ability levels receive the same score because there are no items that discriminate between them. If both experience a concussion, it is possible the test score will not accurately represent the extent of cognitive impairment in one or both individuals. However, the test will be able to identify a person with an ability level that is lower than the ability level of those who receive the ceiling score.

This discrepancy between the true ability distribution and the observed score distribution is a major measurement concern. This problem may relate to the possible connection between vulnerability to concussion and learning disabilities.<sup>11</sup> It is unlikely that latent ability cognitive function has a naturally skewed distribution; the likely cause of a skewed distribution

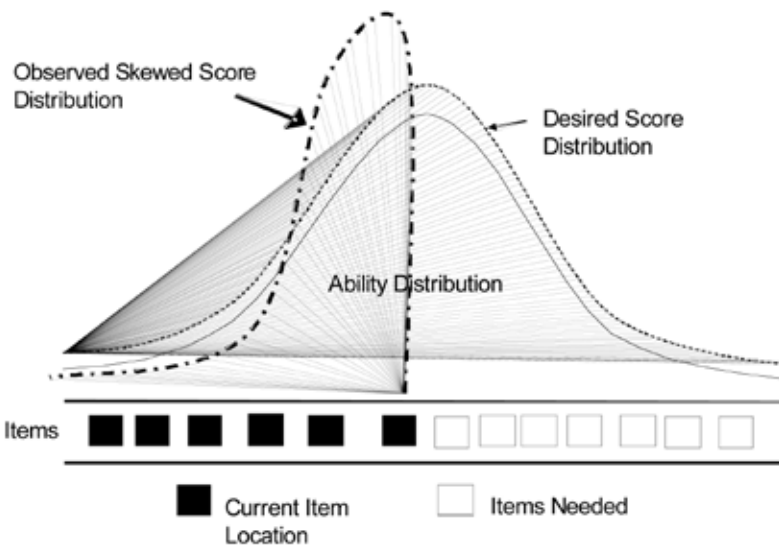


**Figure 3** Depiction of a test score ceiling effect (2 individuals with different ability levels, but the same baseline test score).

of cognitive function scores a poorly designed scale. Further research will help to guide development of an NP test that appropriately utilizes an individual-centered standard for score interpretation.

A ceiling effect may be due to a discrepancy between the score distribution and the ability distribution. If abilities in the various dimensions of cognitive function are normally distributed, but the observed NP test scores are not, the test was poorly constructed from a psychometric perspective. A test that does not contain any difficult items violates the assumption on which the test scores are interpreted. A solution to the ceiling effect is illustrated in Figure 4. In this example, a greater number of difficult items are needed to avoid a ceiling effect and to more accurately measure cognitive ability at baseline testing. The test must include items that represent the entire continuum of cognitive ability for the score distribution to match the ability distribution.

NP testing is considered a vital tool for quantification of an individual athlete's cognitive function; however, there are many problems that preclude heavy reliance on NP testing as a diagnostic tool for MTBI. It is estimated that as many as 50% of subtle brain injuries are not detected by NP testing.<sup>12</sup> Because many NP tests have a low ceiling, the baseline preinjury scores of individuals with high-level cognitive ability may fall within a normal mid-level range. If this is the case, the NP test will not possess sufficient sensitivity to identify a postconcussion decrease in function.<sup>10,12</sup>

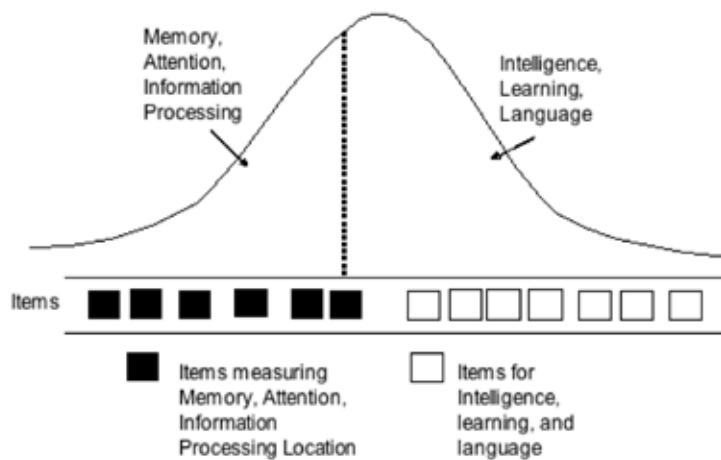


**Figure 4** Depiction of a discrepancy between a true ability distribution and the score distribution of test that has a ceiling effect (test items represented by black squares do not adequately assess the entire continuum of ability levels; white squares represent items that assess high-level abilities, which would improve test sensitivity if added).

A low ceiling may account for the failure of NP test to detect evidence of relative mild brain injuries.

The validity of an NP test score is heavily dependent upon the clinical knowledge and expertise guiding the test designers' selection of items to represent various cognitive functions. Relatively easy items that assess memory, attention, and information processing represent basic cognitive functions that are commonly affected by MTBI. More difficult items may be needed to assess high levels of intelligence, learning, and language. Lezak<sup>13</sup> has suggested that memory and learning are central to all cognitive functioning, but the different dimensions of cognitive functioning may not be easily distinguishable from one another. The possible overlap in the dimensions may be important to understand to greater degree for construction of a better NP test.

Figure 5 illustrates the idea that some dimensions of cognitive function might be easier to perform than others, which may be responsible for an NP test ceiling effect. Assuming the existence of a strong relationship between the various dimensions, a more sensitive unidimensional representation of cognitive function might be achieved through the addition of items that assess more difficult aspects of cognitive function, such as learning and language. This approach is commonly used to assess learning in educational settings. For example, a math test may be designed to assess different math abilities, such as addition, multiplication, and algebra, yet all of the test items of varying levels of difficulty collectively estimate the math abil-



**Figure 5** Depiction of the theoretical contributions of component cognitive abilities of differing levels of complexity (assessed by various test items) to a unidimensional representation of cognitive function (test score).

ity of the examinee. This approach may also work for quantification of preconcussion and postconcussion cognitive ability.

### Other Psychometric Concerns

In addition to concerns about NP test design and score interpretation, there are other issues that must be considered to ensure that valid and reliable concussion assessment measures are obtained. Every NP test item should be subjected to extensive psychometric analysis for determination of its difficulty level and discrimination capability. The test administration duration and the possible existence of practice effects should also be examined. Current NP test designs and the manner in which NP test scores are utilized raise concerns about the psychometric properties of the tests. These issues are exceedingly complex. Advanced knowledge of instrument design methodology, including Rasch analyses and computer adaptive testing, is essential for development of a psychometrically sound and sufficiently sensitive NP test.

### Conclusions

NP tests developed by content experts, even those designed by respected professionals who have had significant clinical experience, must have their psychometric properties thoroughly evaluated. The relatively poor performance of currently available NP tests may

be attributable to test design inadequacies and utilization of inappropriate test score interpretation standards. NP test items must adequately measure both baseline preconcussion cognitive ability and follow-up postconcussion cognitive ability for the test scores to be utilized in clinical decision making. Improved NP test design and development of better score interpretation standards could significantly promote the safety and well-being of athletes who sustain a concussion. ■

### References

1. Randolph C, McCrea M, Barr WB. Is neuropsychological testing useful in the management of sport-related concussion? *J Athl Train.* 2005;40(3):139-154.
2. McCrea M, Kelly JP, Randolph C, Kluge J, Bartolic E, Finn G, Baxter B. Standardized assessment of concussion: onsite mental status evaluation of the athlete. *J Head Trauma Rehabil.* 1998; 13:27-35.
3. Safrit MJ. Criterion-referenced measurement: validity. In: Safrit MJ, Wood TM, Eds. *Measurement Concepts in Physical Education and Exercise Science* 1<sup>st</sup> ed. Champaign, IL: Human Kinetics; 1989:119-136.
4. American Educational Research Association, American Psychological Association, & the National Council on Measurement in Education. *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association; 1999:37-48.
5. McCrea M, Kelly JP, Randolph C. *Standardized Assessment of Concussion (SAC): manual for administration, scoring and interpretation.* Waukesha, WI: CNS Inc; 1996.
6. Barr WB, McCrea M. Sensitivity and specificity of standardized neurocognitive testing immediately following sports concussion. *J Int Neuropsychol Soc.* 2001;7:693-702.
7. Teasdale G, Jennett B. Assessment of coma and impaired consciousness: a practical scale. *Lancet.* 1974;2:81-84.
8. Concussion in Sports Group. Summary and agreement statement of first international conference on concussion in sports. *Br J Sports Med.* 2002; 36:6-10.
9. Erlanger DM, Saliba E, Barth J, Almquist J, Webright W, Freeman J. Monitoring resolution of postconcussion symptoms in athletes: preliminary results of a web-based neuropsychological test protocol. *J Athl Train.* 2001;36:280-287.
10. Echemendia RJ, Julian LJ. Mild traumatic brain injury in sports: neuropsychology's contribution to a developing field. *Neuropsychol Rev.* 2001;11:69-88.
11. Segalowitz SJ, Brown D. Mild head injury as a source of developmental disabilities. *J Learn Disabil.* 1991; 24:551-559.
12. Posthuma A, Wild U. Use of neuropsychological testing in mild traumatic head injuries. *Cogn Rehabil.* 1988; March/April:22-24.
13. Lezak, MD. (1995). *Neuropsychological Assessment.* New York: Oxford University Press.

**Brian G. Ragan, PhD, ATC** is an assistant professor in the Division of Athletic Training in the School of Health, Physical Education, & Leisure Studies at the University of Northern Iowa.

**Minsoo Kang, PhD** is an assistant professor in the Department of Health and Human Performance at Middle Tennessee State University.