Reliability: Current Issues and Concerns

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

A T H L E T I C T R A I N E R S F I N D themselves using a variety of measurements to record their patients’ progress, including traditional paper-and-pencil, computerized, online, and physical (e.g., isokinetics, goniometry) testing tools. Establishing evidence that a treatment or intervention benefits patients forms the foundation of evidence-based practice. Evidence-based practice is the integration of research and clinical expertise in the application of therapeutic procedures. A clinician uses procedures and therapies that scientific research has demonstrated to be both reliable and valid. Evidence-based practice has become a central issue in the struggle to demonstrate that certified athletic trainers are effective clinicians.

To address the increased need for evidence-based clinical practice, a variety of tools have been developed to assess patient progress. Having an understanding of reliability and validity is important in establishing and selecting the right tool to use. These terms, however, are often misunderstood and used incorrectly. Because reliability and validity are significantly broad topics to examine, this column briefly examines the current issues with respect to reliability only.

**Background**

In order to appreciate the concept of reliability, some background information and the state of current practice are needed. According to the American Psychological Association, National Council on Measurement in Education, and American Educational Research Association, the term **reliability** refers to “the consistency of measurements when the testing procedure is repeated on a population of individuals or groups.” In the past, reliability has been thought of as a characteristic of a test, and the use of the term **reliability** in this context is easy to find. For example, look for the phrase “The reliability of the test is . . .” in a paper on reliability to see the confusion. This language and method of reporting makes reliability appear to be a single piece of evidence. Only a single reliability study is needed for a test, and the test is ready to be used by all. It is common to find many articles using one reliability study as the reference to make generalizations across groups or conditions. For example, a study reports reliability using a pain questionnaire for a group of adults. It is common practice to find that study referenced in follow-up investigations as the sole piece of evidence for the use of this questionnaire with other
groups such as children or the elderly. This practice is misleading and incorrect. It ultimately comes down to the question, Is it the reliability of the test being reported, or is it the reliability of the test scores? The answer to this question has led to calls for clarification of reliability in many disciplines including those in educational and psychological measurement.

**Debate: Is Reliability a Property of the Test or the People Tested?**

Recently, the issue of whether reliability is a property of the test or of the people tested has been argued in the literature. The dilemma stems from the language used to describe and report reliability. The statements “the test is reliable” and “the reliability of the test is .92” lie at the heart of the problem. Thompson suggests that such language is incorrect and can be detrimental if used in future research reports. Thompson explains that reliability is a characteristic of the scores or the data. The term *datametrics* has been proposed to describe reliability as a characteristic of the scores and not as a property of the instrument.

There are several reasons that reliability should be considered a property of the people tested rather than the test or instrument. Reliability is influenced by numerous factors that might include type of training, time of test administration, the scoring policy, and even the attitude of the people taking the test. The many ways reliability can be calculated and the fact that the coefficients are not interchangeable across studies support the position that reliability is not a property of the test. The bottom line is that a reliability measurement of .81 for a test does not save the world! Much more information is needed, such as the types of reliability coefficient and the demographics of the participants. Because of the many methods used to calculate reliability coefficients, the lack of description of the reliability coefficient, and all the factors that influence reliability coefficients, it must be a property of the population sample being examined. It is recommended that reliability coefficients be clearly and correctly reported in future research reports.

The idea that reliability is a characteristic of the people tested is applicable to athletic training. An examination of reliability coefficients for commonly used pain scales illustrates the issue of variability with respect to type of scale and findings (Table 1). The range of coefficients can be a result of error in many sources including the test itself, variability inherent in the characteristics of the people tested, interaction between the people and the test, and the type of reliability coefficient used.

So then, which is it, the reliability as property of the test or the test scores? We believe that reliability is a combination of the two. Reliability must be a collection of evidence incorporating the factors that influence it. Increased understanding of the type of reliability coefficient and the associated limitations is needed. A careful inspection of what is actually being measured with the test score is also essential. The design of the study will help clarify the reliability being measured.

**Considerations for Study Design**

The design of a study involving calculation of reliability coefficients is a key concept for understanding the type of reliability being reported. The larger research issue should be the focus. The types of error that can occur should be identified and addressed in the study design. Efforts should also be made to clarify the characteristic to be measured, including whether the focus of the study is to examine the reliability of the people tested, the test, or the interaction of the people and the test.

In athletic training, an example illustrating the importance of study design can be seen with balance testing. Balance is often measured for both research and clinical purposes. Following is an example of how the design and focus of the study should be altered depending on the characteristic to be measured. For this example, we want to conduct a balance assessment as an outcome for a new rehabilitation technique. There are two reliabilities that we need to address: the reliability of the force plate used to assess balance and the reliability of the balance scores (i.e., center-of-pressure velocity).

Test reliability should focus on establishing reliability coefficients for the force plate. This is the first step in our bigger research question, and the design of the study is the key. Having people balance on the force plate for multiple trials over a few days is common study design. This design, however, will not produce reliability coefficients that reflect the characteristic of the test (i.e., force plate). It is common for balance-reliability studies to report this type of reliability as the reliability of the force plate. Efforts should be made to make the force plate the focus of the study. A study design using a machine with known calibrated values...