

# Reliability: Current Issues and Concerns

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

**A**THLETIC TRAINERS FIND 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.<sup>1</sup> 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.”<sup>2(p25)</sup> In the past, reliability has been thought of as a characteristic of the test, and its use, as such, 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 the reliability of the test being reported or is the reliability of the test scores being reported? The answer to this question has led to debate calling for clarification of reliability in many disciplines including those in educational and psychological measurement.

### Debate: "Is Reliability a Property of the Test or the Person?"

Recently, the issue of whether reliability is a property of the test or of the person 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<sup>3</sup> suggests that the language is incorrect and can be detrimental if reported as such 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.<sup>4</sup>

There are several reasons that reliability should be considered a property of the person 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 it is not a property of the test. The bottom line is in the fact 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 person or population 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 person is applicable to athletic training. An examination of reliability coefficients for commonly used pain scales illustrates the issue of variability with respect to type and findings (Table 1). The range of coefficients can be a result of error in many sources including the test itself, inherent variability in the characteristic of the people studied, 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

TABLE 1. RELIABILITY OF PAIN MEASURES

Authors (Ref. #)	Sample	Protocol	Reliability
Grossman et al. <sup>5</sup>	71 cancer patients	VAS, VRS	Test-retest: VAS $r = .97$ , VRS $r = .94$ .
Gaston-Johansson <sup>6</sup>	279 patients; rheumatoid arthritis = 60 patients	POM, VAS, VRS	Test-retest: VAS .88, VRS .84.
DeLoach et al. <sup>7</sup>	60 post-op patients	VAS, NRS	Repeatability = 17.6, 23, 13.5 mm; accuracy of measure is $\pm 20$ mm.
Bergh <sup>8</sup>	167 elderly	VAS, NRS, GRS	Test-retest $r = .75$ to $.83$ .
Carey <sup>9</sup>	267 patients	VAS, NRS, Faces	alpha = .88.
Denegar and Perrin <sup>10</sup>	30 patients with DOMS	GRS	No published evidence: Relies on VAS established reliability because of similarity in scales.
Scott and Huskisson <sup>11</sup>	Series of 10 experiments	VAS, GRS, VRS	Only horizontal VAS with verbal descriptors along the scale showed good reliability with frequency distribution.

Note. VAS = visual analogue scale; VRS = verbal rating scale; POM = Pain-O-Meter; NRS = numeric rating scale; GRS = graphic rating scale; DOMS = delayed-onset muscle soreness.

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 the study where reliability coefficients are reported is a key concept in better understanding the type of reliability being reported. The larger research issue should be the focus. The types and areas where error can occur should be identified and included 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 person, the test, or the interaction of the person and the test.

In athletic training, an example illustrating the importance of the 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 intended 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 is needed to establish the reliability of the test. The characteristic of the force plate is its consistency in measuring a value. It is critical that the value be standardized and known to establish the reliability of the force plate, and these data are often made available by the manufacturer.

Establishing the reliability of the test scores should then be the second step in the research process. After the reliability of the force plate has been established, the reliability of the test scores of the participants can be addressed. Designing this type of study is vital in establishing the evidence of clinical validity. The design of this type of study lends itself to the use of generalizability theory.<sup>12</sup>

## Generalizability Theory

Generalizability theory (G-theory) takes the intraclass correlation coefficient farther by examining the error variances through ANOVA procedures.<sup>13</sup> G-theory can be used to determine how many measurement occasions are needed or to help design the measurement procedure that allows for dependable scores. G-theory was developed in the 1960s by Cronbach et al.<sup>14</sup> and was later applied to physical education settings through the work of Atwater et al.<sup>15</sup> It was additionally refined in concept and through the production of relevant application software by Brennan<sup>12</sup> and Crick.<sup>16</sup> The first chapter discussing G-theory appeared in a textbook in 1989.<sup>13</sup>

G-theory measurement procedure refers to what and how the data are collected from the design. All scores contain some degree of error caused by sources of error in the measurement procedure or design. Using traditional methods such as the intraclass correlation coefficient, the error component of the reliability analysis was not well defined or singular. In G-theory, various possible components of the error variance can be defined and examined for their role. G-theory allows researchers to specify the sources of variation in the study design, provide guidelines for deciding whether the identified sources of variation are errors of measurement or contribute to understanding the construct being measured, and provide information in designing an efficient measurement procedure to reduce the error and achieve the desired reliability. For more information on G-theory, we refer the reader to Shavelson and Webb,<sup>17</sup> Brennan,<sup>18</sup> and Morrow.<sup>13</sup>

## Conclusions

*Reliability* is a term that is routinely used in research and clinical practice and sometimes stated in simplistic and incorrect ways. One study or piece of evidence

does not sufficiently address reliability. There must be a collection of reliability evidence. The phrase “the reliability of the test is . . .” is not accurate in describing the type of reliability measured. Reliability of the test and the characteristic of the examinees are both needed to establish reliability. The test’s or measuring device’s reliability must be established before the reliability of the examinees is measured. The design and inclusion of possible areas where error can occur in the model are needed to address both instrument error and variability in the characteristic of the examinees. The use of G-theory holds promise in addressing the designs of measurement procedures that yield reliable data for both researchers and clinicians. ■

## References

1. Sackett DL, Straus SE, Richardson WS, et al. *Evidence-Based Medicine: How to Practice and Teach EBM*. 2nd ed. London, UK: Churchill Livingstone; 2000.
2. American Psychological Association, National Council on Measurement in Education, American Educational Research Association. *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association; 1999.
3. Thompson B. Guidelines for authors. *Educ Psychol Meas*. 1994;54:837-847.
4. Sawilowsky SS. Psychometrics versus datametrics: comment on Vachon-Haase’s “reliability generalization” method and some EPM editorial policies. *Educ Psychol Meas*. 2000;60:157-173.
5. Grossman SA, Sheidler VR, McGuire DB, Geer C, Santor D, Piantadosi S. A comparison of the Hopkins Pain Rating Instrument with standard visual analogue and verbal descriptor scales in patients with cancer pain. *J Pain Symptom Manage*. 1992;7:196-203.
6. Gaston-Johansson F, Gustafsson M. Rheumatoid arthritis: determination of pain characteristics and comparison of RAI and VAS in its measurement. *Pain*. 1990;41:35-40.
7. DeLoach L, Higgins M, Caplan A, Stiff J. The visual analog scale in the immediate postoperative period: intrasubject variability and correlation with a numeric scale. *Anesth Analg*. 1998;86:102-106.
8. Bergh I, Sjoström B, Oden A, Steen B. An application of pain rating scales in geriatric patients. *Aging (Milano)*. 2000;12:380-387.
9. Carey SJ, Turpin C, Smith J, Whatley J, Haddox D. Improving pain management in an acute care setting. The Crawford Long Hospital of Emory University experience. *Orthop Nurs*. 1997;16:29-36.
10. Denegar C, Perrin D. Effects of transcutaneous electrical nerve stimulation, cold, and a combination treatment on pain, decreased range of motion, and strength loss associated with delayed onset muscle soreness. *J Athl Train*. 1992;27:200-206.
11. Scott J, Huskisson EC. Graphic representation of pain. *Pain*. 1976;2(2):175-184.
12. Brennan RL. Some applications of generalizability theory to the dependability of domain-referenced tests. ACT Technical Bulletin no. 32. *Annual Meeting of the National Council on Measurement in Education*. San Francisco, Calif, April 1979:79. ■
13. Morrow JR. Generalizability theory. In: Safrit M, Wood T, eds. *Measurement Concepts in Physical Education and Exercise Science*. Champaign, Ill: Human Kinetics; 1989:73-96.
14. Cronbach L, Gleser G, Nanda H, Rajaratnam N. The dependability of behavioral measurements. In: *Theory of Generalizability for Scores and Profiles*. New York, NY: Wiley; 1972:■.
15. Atwater AE, Safrit MJ, Baumgartner TA, West C. Reliability theory. In: *AAHPER Publications*. Washington, DC: American Alliance for Health, Physical Education and Recreation; 1976:46.
16. Crick J, Brennan R. GENOVA: A generalized analysis of variance system computer program. 2nd ed. Dorchester: University of Massachusetts at Boston, Computer Facilities; 1984.
17. Shavelson RJ, Webb NM. Generalizability theory: a primer. In: *Measurement Methods for the Social Sciences Series*. Vol 1. Newbury Park Calif.: Sage Publications; 1991:xiii, 137.
18. Brennan RL. An NCME instructional module on generalizability theory. *Educ Meas: Issues Pract*. 1992;11:27-34.

---

**Brian Ragan** 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** is an assistant professor in the Department of Health, Physical Education, Recreation, and Safety at the Middle Tennessee State University.