Journal of Athletic Training 2018;53(1):98–103
doi: 10.4085/1062-6050-503-16

www.natajournals.org

Statistical Primer for Athletic Trainers: The Essentials of Understanding Measures of Reliability and Minimal Important Change

Bryan L. Riemann, PhD, ATC, FNATA*; Monica R. Lininger, PhD, LAT, ATC†

*Department of Health Sciences, Georgia Southern University, Savannah; †Athletic Training Education Program, Northern Arizona University, Flagstaff

Objective: To describe the concepts of measurement reliability and minimal important change.

Background: All measurements have some magnitude of error. Because clinical practice involves measurement, clinicians need to understand measurement reliability. The reliability of an instrument is integral in determining if a change in patient status is meaningful.

Description: Measurement reliability is the extent to which a test result is consistent and free of error. Three perspectives of reliability—relative reliability, systematic bias, and absolute reliability—are often reported. However, absolute reliability statistics, such as the minimal detectable difference, are most relevant to clinicians because they provide an expected error estimate. The minimal important difference is the smallest change in a treatment outcome that the patient would identify as important.

Recommendations: Clinicians should use absolute reliability characteristics, preferably the minimal detectable difference, to determine the extent of error around a patient's measurement. The minimal detectable difference, coupled with an appropriately estimated minimal important difference, can assist the practitioner in identifying clinically meaningful changes in patients.

Key Words: minimal detectable difference, reporting statistical findings, outcomes

Previously, we[1] made the case that clinicians need to consider the clinical meaningfulness of research results that attain statistical significance, as statistical significance alone does not guarantee that an intervention has a sufficient effect in decreasing injury risk or restoring function after injury. Confidence intervals (CIs) and effect sizes were subsequently introduced as tools to assist in determining the clinical meaningfulness of a research result.[2] In this paper, we focus on measurement reliability. *Measurement reliability* can be considered the consistency or stability of a measurement. In practice, all measurements have some magnitude of error, whether it is attributable to the instrument, clinician, or patient, and therefore the values obtained may fluctuate with serial evaluations.

For several reasons, clinicians must have a rudimentary understanding of measurement reliability. First, when selecting tools and methods for evaluating a patient, the reliability of the tool or method should be heavily weighted. Second, as we clinicians provide patient care, we need to understand how to interpret changes in a patient's score regarding whether the changes exceed expected measurement error and whether the patient has achieved a clinically meaningful change. It is beyond the scope of this paper to offer a comprehensive and inclusive review of all reliability statistics; instead, we will focus on the reliability-related concepts and terms used most frequently with continuous measures in the *Journal of Athletic Training*. We will also address the concept of the minimal clinically important difference (MCID). Finally, we will provide a detailed example that applies all of the aforementioned concepts.

RELIABILITY AND SOURCES OF MEASUREMENT ERROR

Many terms are used to describe measurement reliability, including *agreement*, *repeatability*, *precision*, *consistency*, and *minimal detectable change*. Again, *reliability* refers to the extent to which a test or instrument provides a measure that is free of error over repeated trials. The repeated measurements can be taken by the same clinician (intratester reliability) or by different clinicians (intertester reliability); in other circumstances, the repeated measurements may be taken within a single session (intrasession reliability) or between sessions (intersession reliability). As described in our discussion[1] of using sample statistics to estimate population parameters, it is rarely possible to know the true value being measured. We use the value obtained as an estimate of the true value with the understanding that the reliability statistics associated with the test or instrument provide an estimate about the margin of error that surrounds the obtained value. Although this seems simple, the many different approaches, coupled with the variety of interpretations, make understanding the reliability of a test or instrument a challenge not only for clinicians but for researchers as well. Often, reliability is determined by a simple test-retest scenario in which participants are tested and then retested some time later; critical to this approach is the assumption that the underlying characteristic being measured is not changing.

Total measurement error can be considered the sum of 2 error sources: systematic changes (bias) and random error. *Systematic error* refers to the general trend for scores to

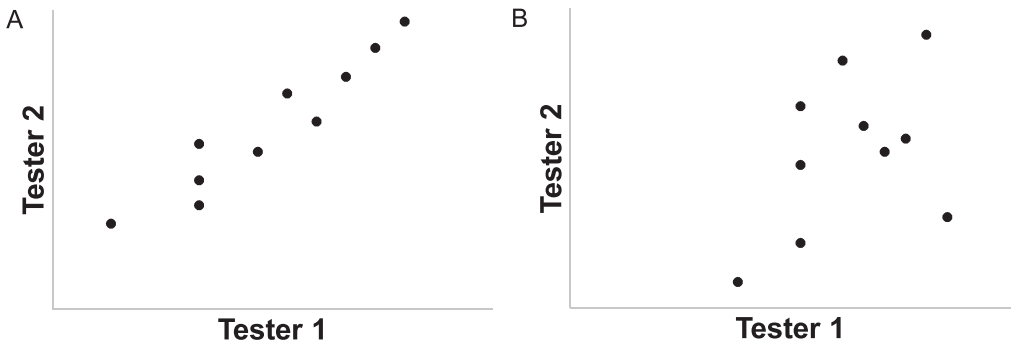

Figure. Scatterplots for 2 scenarios of relative reliability between 2 testers. A, Good relative reliability (intraclass correlation coefficient = 0.95). B, Poor relative reliability (intraclass correlation coefficient = 0.36).

change in a particular direction, either increasing or decreasing, between repeated measurements.[3] Systematic increases are typically demonstrated when learning or familiarity with the test occurs, either by the patient or the examiner as he or she becomes more competent. Systematic decreases often occur when recovery time is insufficient between repeated trials and fatigue results. *Random errors* are unpredictable changes in the obtained measurement that arise because of inherent biological, mechanical, or protocol variations.[3] Examples of random errors include changes in motoneuron-pool excitability during strength measurements (biological), small fluctuations in treadmill belt speed during a running analysis (mechanical), and subtle changes in the specific cues given to participants completing an agility task (protocol). Although protocol variations can be reduced with standardized methods and sufficient practice, random fluctuations from biological and mechanical sources are more difficult to minimize.

To illustrate these concepts, we will continue with the ankle-dorsiflexion active range-of-motion (AROM) example introduced previously.[1] To determine the reliability of the AROM, we could ask a group of participants to perform 3 successive trials of dorsiflexion AROM while our tester assesses each angular displacement with a standard goniometer. Systematic bias could arise from increased stretch tolerance prompted by repeatedly stretching the posterior ankle structures, resulting in the participant's moving through a slightly greater AROM during the third trial compared with the first.[4] Random error could occur from the examiner's viewing the alignment of the goniometer arm segment landmarks from slightly different angles during each trial or the participant's having a slightly different focus when moving to end range. These 2 sources of error are not all-inclusive but rather illustrate both error components.

STATISTICAL APPROACHES FOR EVALUATION AND DESCRIBING MEASUREMENT RELIABILITY

The variety of statistical approaches used to assess and describe reliability make reading and interpreting these measures a daunting task. Similarly to previous authors,[3,5] we advocate considering the various measures in 3 categories: relative reliability (ie, test-retest correlation), systematic bias (ie, change in means), and absolute reliability (ie, repeated-measurement variability). Each category provides a different perspective on reliability and therefore it is typical to see multiple methods reported in a research paper. An in-depth examination of each statistical method is beyond the scope of this paper. However, once a basic understanding of each category is gained, clinicians will be better able to interpret reliability statistics. We will focus on the most common methods in each category that are reported in the *Journal of Athletic Training*.

Methods of Determining Relative Reliability

Relative reliability statistics are the most frequently reported. They describe the consistency of the rank or position of individuals within a group across repeated assessments. They are based on the correlation or relationship between sets of scores and can be illustrated by a scatterplot (Figure). The most popular statistic for relative reliability is the intraclass correlation coefficient (ICC). If we are considering interrater reliability, the ICC indicates whether both raters score the same individuals as low or high relative to the other participants. Although the idea behind the ICC is straightforward, interpreting the clinical meaningfulness is a bit more difficult. The maximal value for an ICC is 1, which indicates perfect relative reliability. The minimal value for an ICC is conceptually zero, which indicates no relative reliability. The difficulty is in deciding if the ICC is close enough to 1 for the test to be considered reliable. No universal standard for an acceptable coefficient is currently available, but most experts consider values above 0.75 as supporting good reliability.[6]

One challenge with the ICC is the different terminologies used to describe the variations (Table); however, we will focus on the nomenclature used by Shrout and Fleiss.[7] The reader should understand the major variations of the ICC because the model reported can profoundly influence the resulting coefficient. The ICC variations are denoted by 2 numbers. The first number indicates the model (1, 2, or 3). If the ICC is meant to describe the reliability specific to the current study with no intent to generalize the reliability estimates beyond that, then model 3 (sometimes referred to as a *mixed model*) is appropriate. This value is used to show that a group of raters used in a study are consistent with one another. In contrast, if the ICC is intended to be generalized beyond the current study, such as to other testers with similar experience, then model 2 (*random model*) is appropriate. In model 1 (which is seldom used), the participants are assessed by different, randomly selected raters. The second number reflects the form of the ICC. The value indicates whether the reliability was based on a single measurement or an average of all measurements (some-

Table. Intraclass Correlation Coefficient Models Used in Athletic Training

Example	Shrout and Fleiss[7]	SPSS Model[a]
A random set of raters evaluate participants.	1,1	1-way random
A random set of raters evaluate participants, and scores are averaged.	1,k	1-way random
The same raters evaluate all participants, and the results will be generalized to another group.	2,1	2-way random
The same raters evaluate all participants, and the results (averaged) will be generalized to another group.	2,k	2-way random
The same raters evaluate all participants, but the results will be used only for the current sample.	3,1	2-way mixed
The same raters evaluate all participants, but the results (averaged) will be used only for the current sample.	3,k	2-way mixed

[a] https://www.ibm.com/support/knowledgecenter/SSLVMB_21.0.0/com.ibm.spss.statistics.help/alg_reliability_intraclass.htm.

times generically denoted with a k) recorded by the raters (1 for a single measurement, 2 for the average of 2 measurements, etc). Generally, if the same scores are subjected to the different models and forms, the ICC will be higher with a mixed average-score model (3,k).

An additional aspect of interpreting a reported ICC is the variability of scores among participants. If scores vary little among participants, a weak ICC can occur even if there is very little within-subject (trial to trial, session to session, or tester to tester) variability. Similarly, if scores vary widely among participants, a strong ICC can occur even with large within-subject variability. Thus, the ICC must be interpreted in the context of the scores achieved and the population studied. This can be accomplished by examining the mean and standard deviation of the scores and the characteristics of the participants. Ideally, the scores and participants should represent the intended application of the test. For example, if a clinician is considering which lower extremity functional performance test (such as the single-legged–hop test variations) to apply in high school athletes, the reliability estimates used to make the decision should be from high school athletes rather than a mix of athletes and nonathletes or ages before or after high school. For a more extensive discussion of the ICC, we recommend that readers review the work of Weir.[8]

Methods of Determining Systematic Bias

Again, *systematic bias* refers to the general trend of scores to increase or decrease between repeated applications of the test. When 2 sets of scores have been collected, systematic bias can be detected with a paired *t* test. When more than 2 sets of scores have been collected, sequential paired *t* tests[9] or a repeated-measures analysis of variance can be used. A statistically significant comparison indicates the presence of systematic bias. However, a lack of statistical significance should not be interpreted as evidence that the measurement is reliable because large random score fluctuations can still exist. Significant systematic bias may challenge the validity (accuracy) of the measurement. If repeated measurements will be used, such as in serial testing of a patient's progress, clinicians need to take into account the systematic bias (eg, stretch tolerance) when interpreting score changes across the repeated applications.

Methods of Determining Absolute Reliability

Absolute error is probably the most pertinent form of reliability for clinicians. In contrast to *relative reliability*, which indicates the consistency of a participant's rank between raters or tests (test-retest scenario), *absolute reliability* measures the precision of a score by estimating the expected error, either from the true score or from test-retest fluctuations. Although several approaches to absolute reliability exist, the result is generally expressed either in the original measurement units or as a proportion or percentage of the measurement values.

Most often, the standard error of measurement (SEM) is reported in conjunction with the ICC as an indicator of absolute reliability. This likely occurs because the most commonly applied SEM formula uses the ICC. One limitation to using the ICC in the computation is that the SEM is influenced by the same factors as the ICC (eg, model, variability of scores). In a test-retest context, the SEM covers only 52% of the differences between tests ($\sqrt{2}$ × standard deviation of the differences), not the 68% of the true score error typically attributed to the SEM for a single measurement. For this reason, the SEM has been criticized as being too small for most applications.[10] Readers seeking a more in-depth discussion of the SEM, as well as a method for computing the SEM independent of the ICC, should consult the work of Weir.[8]

The *minimal detectable difference* (MDD), also frequently called the minimal detectable change or smallest detectable change, can be considered an extension of the SEM. The MDD is interpreted as the boundaries of measurement error. With repeated testing, any change outside these boundaries can be considered a true change in the entity being assessed. The MDD is often computed based on 90% ($MDD_{90\%}$) or 95% ($MDD_{95\%}$) confidence by multiplying the SEM by 2.33 ($\sqrt{2} \times 1.65$) or 2.77 ($\sqrt{2} \times 1.96$), respectively. The result is a fairly conservative estimate of measurement error. When the MDD is used to interpret measured changes in a patient, changes larger than the MDD are very likely to be real (ie, beyond measurement error). A large MDD may lead to the conclusion that the patient has experienced no real change when, in fact, he or she has experienced a clinically meaningful change. Further discussion on this topic appears in the next section. Additionally, intervention study results can be enhanced by including the proportion of participants whose scores exceeded the MDD as a supplement to reporting statistically significant group differences.

The *coefficient of variation*, in which error is expressed as a percentage of the sample's mean score ([standard deviation/mean] × 100), is another approach used to estimate measurement error. As a unitless estimate of measurement error, the coefficient of variation facilitates comparisons across various measures and study reports. It cannot be used when the mean of a variable is zero (ie, the scale includes negative values) or close to zero. The most useful application of the coefficient of variation is when the magnitude of measurement error is directly related to the

magnitude of scores (heteroscedasticity). For example, when assessing isokinetic strength in pushing and pulling, stronger individuals demonstrated more measurement error than weaker individuals.[11] Although heteroscedasticity has not received much attention in the *Journal of Athletic Training*, it is likely that many reported measures exhibited this characteristic,[12] which is an important consideration for future reliability studies in the *Journal*. For a more extensive discussion of heteroscedasticity, readers are advised to review the works by Atkinson and Nevill[3] and Bland and Altman.[13]

MINIMAL IMPORTANT DIFFERENCE

When interpreting changes in a patient, a clinician must consider not only statistical significance and the reliability of the measure but also whether the change is meaningful. The definition of a meaningful change depends on one's perspective.[14] A patient may have 1 perspective, while the attending clinician or payer (eg, health insurer) may have other perspectives. Furthermore, a change in an outcome measure may be important for 1 patient group (ie, injury type, severity, athletic level) but less important for another. In an attempt to provide an estimation of whether a change is meaningful to a patient group, Jaeschke et al[15] defined the *minimal clinically important difference* (MCID) as "the smallest difference in score in the domain of interest which patients perceive as beneficial and which would mandate, in the absence of troublesome side effects and excessive cost, a change in the patient's management."(p408) Several authors[16,17] subsequently recommended the term *MCID* be changed to *minimal important difference* (MID) to emphasize the patient's perspective as paramount. Since then, many other terms have been used in the literature to refer to the same broad concept, such as *minimal worthwhile effect* and *subjectively significant difference*, each with slight variations in their exact definitions and derivations.[18]

Regardless of the term used (we will use *MID*), the concept can be simplified as the threshold of change beyond the MDD that the patient perceives as meaningful or worthwhile, such that he or she would elect to repeat the intervention if given the choice.[19] Most authors who provided MIDs studied patient-reported outcomes and health-related quality-of-life measures, although an increasing number of investigators are reporting the MID for other commonly used measures relevant to athletic training practice. Establishing the MID for a measure takes 2 general approaches. The first approach, which is anchor based, relies on an external criterion (anchor) to indicate that change has occurred. Most often, patient perceptions of their improvement are used as anchors, such as global assessment ratings, which ask patients to rate their change along a continuum of *better*, *unchanged*, or *worse*.[19] The second approach, which is distribution based, relies on the statistical characteristics of the data, such as the measurement precision, statistical significance of the change, and variability of scores.[14] Distribution-based approaches include effect sizes, measures of absolute reliability (ie, SEM, MDD), and paired *t* tests. The major criticism of a distribution-based approach is it does not take into account the patient's perspective of important changes and, therefore, the MID estimate often differs from the result of an anchor-based approach.[20,21] Various authors[14,18,19,22,23] have weighed in on the strengths and weaknesses of both MID approaches, as well as the specific methods of each approach, but to date, universal consensus is lacking.

As previously mentioned, the MID is typically larger than the MDD. However, for instruments such as the commonly used Shortened Disabilities of the Arm, Shoulder and Hand Questionnaire (QuickDASH)[24] and American Shoulder and Elbow Surgeons Standardized Shoulder Assessment Form,[25] which assess upper extremity disability, the MID is smaller than the minimal detectable change. Although 1 perspective[26] on this circumstance is to question the utility of the instrument, others[25] have suggested that the MID is more important because it is related to the patient's perception of meaningful change. Regardless of the perspective taken, this discrepancy highlights the potential incongruence between distribution (MDD)- and anchor-based approaches to establishing clinical meaningfulness.[27]

In addition to the computation challenges, clinicians must be aware that MID values for a measure have limited generalizability, meaning that they may not transfer to different patient populations (injury type, severity, sex, athletic level, etc) and are highly dependent on the method used to establish a reported MID.[14,19,28] Until universal agreement and recommendations are available to address these challenges, the MID should be used prudently in clinical settings and research,[28] particularly when a patient differs from the populations and settings from which the MID was derived. Regardless, clinicians should prioritize treatment interventions that demonstrate improvements exceeding the MID when making treatment decisions with a patient. Furthermore, as with the MDD, researchers investigating treatment interventions should report the proportion of patients who demonstrated change that met or exceeded the MID relative to all patients in the trial (ie, responders). This information can assist the clinician in estimating the likelihood that his or her patient will respond favorably to a similar treatment intervention.[29] When possible, clinicians should choose an outcome measure that has a reported MID derived from a population similar to that of the patient. In this manner, they will be able to better monitor patient improvements relative to a threshold indicating meaningful change.

PRECISION OF RELIABILITY ESTIMATES

In our previous discussion[2] of CIs, we described the width (precision) of a CI as an important factor that influences the clinical utility of the reported interval. Furthermore, sample size was noted to have a potent influence on the precision of a CI. In a similar manner, the MDD and MID estimates of reliability are also point estimates for the population based on a study sample. Consequentially, CIs are frequently provided for these measures. When assessing the clinical meaningfulness of reliability estimates, in addition to interpreting the reported reliability estimates, clinicians need to consider the sample size from which the reliability estimates were derived, as well as the precision of the CIs around the reported reliability estimates.

SUMMARY EXAMPLE

To illustrate and summarize the concepts in this paper, we will consider 3 patients seeking treatment for shoulder pain. The 3 patients are women in their early 30s, employed as office workers, and recreational tennis players on the weekends. Evaluation of their shoulders suggests shoulder impingement without shoulder instability. As part of the initial evaluation and re-evaluation after 3 weeks of treatment, the patients completed the QuickDASH.[30] The QuickDASH contains 11 items to be completed by the patient. The scores can range from 0 to 100, and higher scores indicate more disability. Between the initial and follow-up evaluations, the first patient demonstrated a 4-point decrease, the second patient demonstrated a 15-point decrease, and the third patient demonstrated a 9-point decrease.

To interpret these scores with respect to measurement error and meaningful change founded upon an MID, we rely on a report of the QuickDASH's psychometric properties by Mintken et al.[24] The rationale for using their report is the similarities among our patients, demographics, conditions, severity of baseline symptoms, and treatment time before reevaluation. Mintken et al[24] reported the relative reliability for the QuickDASH using an ICC (2,1) across 14 days as 0.90. This result represents strong reliability, and the ICC model used (random, single trial) suggests that the scores can be readily generalized by other clinicians administering the QuickDASH to similar patients across a 14-day interval. The reported SEM was 4.8 points, which corresponds to an 11.2-point $MDD_{90\%}$ (4.8×2.33). Thus, for the clinician to be certain the observed patient improvement exceeds measurement error, the patient would need to demonstrate a 12-point score reduction. Using an anchor-based approach (global rating of change), Mintken et al[24] reported an 8-point MID.

With this information, we can now interpret the changes in our 3 patients. Our first patient's change is below both the MDD and the MID, suggesting that it is neither beyond measurement error nor clinically meaningful. Our second patient's change exceeds both measurement error and clinical meaningfulness, which strongly suggests that our treatment has been successful in improving her perception of her disability. Interpretation of the third patient's change score is more tenuous. The 9-point improvement suggests clinical meaningfulness in her perception of disability change; however, the change did not exceed the measurement error. As previously discussed, the QuickDASH demonstrates an MID smaller than the MDD. In this circumstance, we can appreciate the patient's perception of improvement as evidence of attaining some treatment success but would likely choose to continue treatment to be certain that the improvement is real and not simply measurement error.

RECOMMENDATIONS

Assessing a patient's injury or illness status or monitoring change over the course of a treatment intervention requires that clinical tests and measures be selected based on satisfactory reliability and validity, along with sufficient precision and responsiveness to identify changes in a patient's status when a true change has occurred. Over the past decade, reliability reporting has increased for many clinical tests and measures; we recommend that this trend continue. We advocate for reliability reporting to include relative reliability, systematic bias, and absolute reliability. Clinicians should use absolute reliability characteristics, preferably the MDD, to determine the extent of error around a patient's measurement. The MDD, coupled with an appropriately selected MID estimate, can assist the practitioner in triangulating the clinically meaningful changes in patients undergoing treatment. Clinicians should consider adopting treatment interventions that have produced changes that exceeded the MID for large proportions of patients. To facilitate use of the tests and measures that have adequate reliability and MID, as well as treatments shown to produce meaningful changes, we strongly advise that these characteristics be fully incorporated, not only into research reports but also into injury assessment and rehabilitation resources such as textbooks.

REFERENCES

1. Riemann BL, Lininger M. Statistical primer for athletic trainers: the difference between statistical and clinical meaningfulness. *J Athl Train*. 2015;50(12):1223–1225.
2. Lininger M, Riemann BL. Statistical primer for athletic trainers: using confidence intervals and effect sizes to evaluate clinical meaningfulness. *J Athl Train*. 2016;51(12):1045–1048.
3. Atkinson G, Nevill AM. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Med*. 1998;26(4):217–238.
4. Magnusson SP, Aagard P, Simonsen E, Bojsen-Moller F. A biomechanical evaluation of cyclic and static stretch in human skeletal muscle. *Int J Sports Med*. 1998;19(5):310–316.
5. Hopkins WG, Hawley JA, Burke LM. Design and analysis of research on sport performance enhancement. *Med Sci Sports Exerc*. 1999;31(3):472–485.
6. Portney L, Watkins M. *Foundations of Clinical Research: Applications to Practice*. 3rd ed. Upper Saddle River, NJ: Pearson Prentice Hall; 2009.
7. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*. 1979;86(2):420–428.
8. Weir JP. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *J Strength Cond Res*. 2005; 19(1):231–240.
9. Hopkins WG. Measures of reliability in sports medicine and science. *Sports Med*. 2000;30(1):1–15.
10. Atkinson G, Nevill A. Typical error versus limits of agreement. *Sports Med*. 2000;30(5):375–381.
11. Riemann BL, Davis SE, Huet K, Davies GJ. Intersession reliablity of upper extremity isokinetic push-pull testing. *Int J Sports Phys Ther*. 2016;11(1):85–93.
12. Nevill AM, Atkinson G. Assessing agreement between measurements recorded on a ratio scale in sports medicine and sports science. *Br J Sports Med*. 1997;31(4):314–318.
13. Bland JM, Altman DG. Measurement error proportional to the mean. *BMJ*. 1996;313(7049):106.
14. Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. *J Clin Epidemiol*. 2003;56(5):395–407.
15. Jaeschke R, Singer J, Guyatt GH. Measurement of health status: ascertaining the minimal clinically important difference. *Control Clin Trials*. 1989;10(4):407–415.
16. Juniper EF, Guyatt GH, Willan A, Griffith LE. Determining a minimal important change in a disease-specific Quality of Life Questionnaire. *J Clin Epidemiol*. 1994;47(1):81–87.

17. Schunemann HJ, Puhan M, Goldstein R, Jaeschke R, Guyatt GH. Measurement properties and interpretability of the Chronic Respiratory Disease Questionnaire (CRQ). *COPD*. 2005;2(1):81–89.
18. King MT. A point of minimal important difference (MID): a critique of terminology and methods. *Expert Rev Pharmacoecon Outcomes Res*. 2011;11(2):171–184.
19. Copay AG, Subach BR, Glassman SD, Polly DW Jr, Schuler TC. Understanding the minimum clinically important difference: a review of concepts and methods. *Spine J*. 2007;7(5):541–546.
20. McGlothlin AE, Lewis RJ. Minimal clinically important difference: defining what really matters to patients. *JAMA*. 2014;312(13):1342–1343.
21. Turner D, Schunemann HJ, Griffith LE, et al. The minimal detectable change cannot reliably replace the minimal important difference. *J Clin Epidemiol*. 2010;63(1):28–36.
22. de Vet HC, Terluin B, Knol DL, et al. Three ways to quantify uncertainty in individually applied "minimally important change" values. *J Clin Epidemiol*. 2010;63(1):37–45.
23. Sloan JA. Assessing the minimally clinically significant difference: scientific considerations, challenges and solutions. *COPD*. 2005;2(1):57–62.
24. Mintken PE, Glynn P, Cleland JA. Psychometric properties of the shortened disabilities of the Arm, Shoulder, and Hand Questionnaire (QuickDASH) and Numeric Pain Rating Scale in patients with shoulder pain. *J Shoulder Elbow Surg*. 2009;18(6):920–926.
25. Michener LA, McClure PW, Sennett BJ. American Shoulder and Elbow Surgeons Standardized Shoulder Assessment Form, patient self-report section: reliability, validity, and responsiveness. *J Shoulder Elbow Surg*. 2002;11(6):587–594.
26. Hagg O, Fritzell P, Nordwall A. The clinical importance of changes in outcome scores after treatment for chronic low back pain. *Eur Spine J*. 2003;12(1):12–20.
27. Wright A, Hannon J, Hegedus EJ, Kavchak AE. Clinimetrics corner: a closer look at the minimal clinically important difference (MCID). *J Man Manip Ther*. 2012;20(3):160–166.
28. Terwee CB, Roorda LD, Dekker J, et al. Mind the MIC: large variation among populations and methods. *J Clin Epidemiol*. 2010;63(5):524–534.
29. Farrar JT, Portenoy RK, Berlin JA, Kinman JL, Strom BL. Defining the clinically important difference in pain outcome measures. *Pain*. 2000;88(3):287–294.
30. Beaton DE, Wright JG, Katz JN. Development of the QuickDASH: comparison of three item-reduction approaches. *J Bone Joint Surg Am*. 2005;87(5):1038–1046.

Address correspondence to Bryan L. Riemann, PhD, ATC, FNATA, Department of Health Sciences, Georgia Southern University, 154 University Hall, 11935 Abercorn Street, Savannah, GA 31419. Address e-mail to bryan.riemann@armstrong.edu.