TITLE PAGE

Article title:

Determining the intra- and inter-observer reliability of screening tools used in sports injury research

Authors:

Andrew Hayen, Rebecca Dennis, Caroline Finch

Keywords:

Reliability, intraclass correlation coefficient, measurement error

Abstract:

Sports injury etiological studies explore the relationships between potential injury risk factors and injury outcomes. The ability of such studies to clearly identify intrinsic risk factors for sports injury depends on the accuracy of their measurement. Measurements need to be reproducible over time and repeatable by different observers, as well as within a given individual. The importance of the reliability of pre-participation screening protocols and other clinical assessment tools has been identified in a number of published studies. However, a review of these studies indicates that a variety of statistical techniques have been used to calculate intra- and inter-observer reliability. Whilst the intra-class correlation coefficient (ICC) is the most often cited measure, a range of statistical approaches to estimating ICCs have been used. It is therefore difficult to determine which statistical method is most appropriate in the context of measuring intrinsic risk factors in sports injury research. This paper summarises a statistical method for the concurrent assessment of intra- and inter-observer reliability and presents an argument for why this approach should be adopted by sports injury researchers using screening protocols that collect continuous data.

Word count for abstract:

184

Acknowledgments:

AH was supported by the core funding provided to the NSW Injury Risk Management Research Centre by the NSW Department of Health, the NSW Roads and Traffic Authority and the NSW Motor Accidents Authority. RD was supported by an NHMRC Public Health PhD scholarship during the data collection and analysis phase, and by an NHMRC Population Health Capacity Building Grant in Injury Prevention, Trauma and Rehabilitation during the reporting and publication phase. CF was supported by an NHMRC Principal Research Fellowship. Funding for the reliability assessment example described in this paper was provided by Cricket Australia.

Determining the intra- and inter-observer reliability of screening tools used in sports injury research

The importance of reliability

Over recent years there has been an increasing call to provide a firm evidence base for sports injury prevention initiatives. As argued by Bahr and Krosshaug [1], provision of this evidence base is limited by knowledge about the etiological factors causing many sports injuries. To redress this imbalance, there needs to be considerably more effort put towards conducting studies to elucidate the intrinsic and extrinsic risk factors for sports injury.

Such studies naturally involve the measurement of potential risk factors and relating these to injury outcomes. In the prospective study ideal, measurements are made on participants in an injury-free state (eg. at the start of a playing season) and these are related to injury outcomes during the following participation period. For intrinsic risk factors, such as strength, flexibility, and balance, it is often of interest to see how these also vary over the playing season or how they differ in injured and uninjured participants at the end of the season. This necessitates taking multiple measurements.

The ability for such studies to clearly identify potential risk factors depends on the accuracy with which these measurements are made [2]. Measurements need to be reproducible over time and by different observers, as well as being repeatable within a given individual. Poor reproducibility limits the ability of researchers to reach conclusions about whether a measured variable is indeed a risk factor for injury,

because it is difficult to differentiate participants with/without the variable of interest in the presence of large random measurement error [3].

Definition of reliability and its related concepts

Validity of measurement is the degree to which a test measures what it is supposed to measure [4] and reliability refers to the consistency, or repeatability, of a measure [4, 5]. Whilst a measure can be reliable without being valid, the reverse is not true [4, 6]. Low reliability indicates that large variations in measurement will occur upon retesting so that assessment outcomes cannot be meaningfully reproduced or interpreted [7]. Whilst factors such as weight and height are typically measured with high reliability, other potential injury risk factors, such as joint range of motion (ROM), may be more prone to unreliable measurement [2]. Another consequence of unreliability is the need for an increased sample size to detect important differences between groups for the variable being measured because of the increased variability in measurement [8]. This has obvious implications for the design of prospective cohort studies and randomised controlled trials that compare control and intervention groups. In particular, this may result in an unnecessary increase in the cost, size and timing of conducting such studies.

In clinical assessments, measurement error can be introduced by the human observer (eg. a physiotherapist conducting a clinical assessment) and/or the instrument used (eg. a goniometer). Using the assessment of ROM as an example, if the goniometer has been shown to be reliable, then the reliability of the ROM measurements depends on the correct use of the goniometer by the physiotherapist. This paper deals specifically with the issue of determining the reliability of the human

observer, which is the ability of a single observer or multiple observers to produce the same measurements consistently under the same conditions with the same sample [7, 9]. Two forms of observer reliability are discussed:

- intra-observer (or within observer) reliability - the degree to which measurements taken by the same observer are consistent
- inter-observer (or between observers) reliability - the degree to which measurements taken by different observers are similar.

Related to, but not identical to reliability, is the concept of precision. Precision is defined as the spread in random measurement error that would be expected if repeated independent observations are made on an individual [3]. It is a measure of absolute error, while reliability assesses the effect of that error on the ability to differentiate between individuals [3]. Obviously, if reliability is poor, it will not be possible to have precise measurements.

Purpose of this paper

The importance of the reliability of pre-participation musculoskeletal screening protocols, fitness assessments and other clinical assessment tools has been identified in a number of published studies [10-16]. These studies, which include both inter-observer and intra-observer reliability assessments of a variety of clinical musculoskeletal tests used in sport and physical therapy, were retrieved from searches of the Medline database. The search terms used were ‘reliability’, combined with ‘screening’ or ‘test’, and at least one of ‘musculoskeletal’ or ‘fitness’ or ‘clinical’. A representative range of reliability assessments of screening tests that investigated both inter-observer and intra-observer reliability, and were published after 1996, were

selected (Table 1). As Table 1 shows, a variety of statistical techniques have been used to establish intra- and inter-observer reliability in these studies. It is now very common in the literature for the intra-class correlation (ICC) to be the measure of choice for determining reliability [23]. Whilst the ICC has been the most often cited reliability measure, a range of models and methods to calculate ICCs have been used.

Many of the ICCs given in Table 1 are based on a popular set of methods described by Shrout and Fleiss [17]. For further details regarding the methods described by Shrout and Fleiss, the reader is referred to their paper [17]. For many of the ways for calculating an ICC presented in that paper it is assumed that each observer takes only one measurement. This means that these methods cannot be applied in studies in which inter- and intra observer reliability need to be measured concurrently and so cannot be applied to inter-observer reliability studies in which observers make more than one measurement [17]. This limitation is also true for Pearson's correlation coefficient, as this cannot be used in situations where there are more than two observers or when each observer takes more than two measurements.

Unfortunately, in several of the studies shown in Table 1, more than one measurement has been taken, but the methods used to calculate the ICC adopted have assumed that only one measurement was taken. Researchers have sometimes used the mean value of each observer's repeated measurements (Table 1), but this has the effect of inflating the inter-observer reliability as an ICC calculated from the mean of multiple measurements will be higher than that based on a single measurement [5, 8, 18]. The mean of multiple measurements should only be used for the reliability measure if the usual context of application is to take multiple

measurements, which may be the case for some of the studies listed in Table 1. Some studies have just used the first of the repeated measurements taken by an observer, but this method is inefficient as it does not use all of the available information. Other studies do not state the exact method they used to calculate the ICC.

<Insert Table 1 about here>

Because a variety of statistical methods to calculate reliability have been reported in the literature, if at all, it is difficult to determine which method is most appropriate in the context of measuring intrinsic risk factors in sports injury research. The purpose of this paper, therefore, is to describe a particular statistical method (initially developed by Eliasziw and colleagues [18]) for the concurrent assessment of intra- and inter-observer reliability and to describe why this is approach should be adopted by sports injury researchers using screening protocols that collect continuous data. We also discuss some of the limitations of the ICC. It should be noted that this paper does not aim to provide a detailed critical review of all available statistical methods for assessing observer reliability, but rather to demonstrate the specific application of the method described by Eliasziw et al. in sports injury research.

Data example

This paper is aimed as an Educational Piece for researchers who conduct reliability research. Whilst this paper presents some statistical formulae, its emphasis is on providing information for application in future studies. To illustrate this, the real-world example of a reliability assessment of a musculoskeletal screening protocol used in a prospective cohort study of cricket fast bowlers is presented [19]. The reliability assessment was conducted using two observers and ten bowlers. The bowlers were

each required to attend one appointment, in which they were tested by each observer twice (in the order of Observer 1, Observer 2, Observer 1, Observer 2). There was one trial measured for each test. The tests were conducted in the same order each time, with 10-minute rest breaks between each session. The screening protocol consisted of a number of tests measuring flexibility, strength and stability. This assessment was approved by the University of New South Wales Human Ethics Review Committee.

Data from the reliability assessment of measurements of hip ROM have been extracted from the larger cricket study for the example in this paper [19]. The range of hip rotation was assessed by physiotherapists with the hip in a neutral hip position. The bowler lay in a prone position with both knees bent to 90°, chin resting on the bench, arms by sides. Internal rotation was measured first and the bowler was asked to let both ankles move away from each other as far as possible, whilst the physiotherapist ensured that pelvic motion and/or hip flexion did not occur. To determine external rotation, the bowler straightened the contralateral knee and let the ankle of the testing leg drop towards the opposite side of the body as far as possible. An assistant to the physiotherapist measured the angle formed by the line of the tibia, relative to the vertical, as determined by a spirit level goniometer [20]. The angle was recorded to the nearest degree. The data collected by Observer 1 and Observer 2 for the hip range of motion test for each of the ten bowlers for the two sessions are presented in the attached supplemental file Appendix 1.

The reliability assessment example here used a short time interval to separate the testing sessions and it must be noted that the reliability of the measurements represents their reproducibility only within this particular time frame. Test-retest

assessments within a short time interval tend to demonstrate higher reliability than those studies with longer time intervals, which may be influenced by a number of uncontrolled variables [9]. Although reliability studies with short time intervals may be appropriate for studies collecting pre-participation data, longer periods of time between assessments (eg. 1 week or 1 month) are important for clinical assessments where there is a need to evaluate patient improvements over time [9].

Statistical methodology

When conducting a reliability study, there are two main situations to consider:

1. the observers are assumed to have been drawn randomly from a larger population (*random observers*)

2. the observers are the only ones of interest (*fixed observers*).

This is an important distinction because the formulas for calculating the reliability differ slightly for these two scenarios. Our reliability assessment example with two physiotherapists is a *random observers* case, because two physiotherapists conducted the musculoskeletal screening of the fast bowlers and the results had to be generalisable to a larger population of physiotherapists. In contrast, in a clinical setting, two clinicians, for example, monitoring the progress of a patient may be the only people that will ever assess this patient. Hence, the results of a reliability assessment do not need to be applied to any other raters and the observers are fixed.

The method presented below has the distinct advantage over other methods (such as those of Shrout and Fleiss [17]) because it allows researchers to simultaneously assess inter- and intra-observer reliability.

In developing the statistical formulation below, it is important to define the terms from the outset. We assume we have *m* repeated measurements made on a sample of *n* subjects by *o* different observers, so that there are *m*×*n*×*o* measurements in total. Although we speak of observers, one can use synonymous terms, such as raters or instruments, depending on the context.

The *k*th (k=1, …, m) measurement taken by the *j*th (j= 1,…, o) observer on the *i*th (i=1, …, n) subject is denoted by Y_{ijk}. Assessing reliability is essentially a repeated measures design and we can represent each of the observations according to the following repeated measures design:

$$Y_{ijk} = \mu + S_i + O_j + (SO)_{ij} + e_{ijk},$$

where μ is the mean of all possible measurements, S_i is the effect of subject *i*, O_j is the effect of observer *j*, $(SO)_{ij}$ is the inter-observer (or across observer) random error (or hetereogeneity), and e_{ijk} is the intra-observer (or within observer) random error. We assume that S_i and e_{ijk} follow normal distributions with mean zero and variances σ_s^2 and σ_e^2 respectively. When assuming *random observers*, it is also necessary to assume that O_j and $(SO)_{ij}$ come from normal distributions with zero means and variance σ_O^2 and σ_{SO}^2 respectively. In the *fixed observer* case, the components O_j and $(SO)_{ij}$ are constrained so that $\sum_{j=1}^{o} O_j = \sum_{j=1}^{o} (SO)_{ij} = 0.$ In addition, $(SO)_{ij}$ is assumed to follow a normal distribution with mean zero and variance $(o-1)\sigma_{SO}^2 / o$ (such constraints are for technical reasons only).

Estimates of the variance components can be obtained from an analysis of variance table, such as Table 2, in which MSO is the mean square for observers, MSS is the mean square for subjects, MSSO is the mean square for subjects ×observers and MSE is the mean square for error. From these tables, the variance components can be estimated by subtraction. For example, for *random observers*, σ_{SO}^2 can be estimated as $(\mathrm{MSO} - \mathrm{MSE})/m$, which is obtained using subtraction in Table 2. The other variance components can be estimated similarly using the table. In some cases, these variance components may be calculated as a negative number, in which case they should be set to zero.

<Insert Table 2 about here>

The calculated analysis of variance for our example with cricket fast bowlers is presented in Table 3. In Table 3, we see that MSE= $\hat{\sigma}_e^2$ = 9.05. Using Table 2 and the necessary subtraction, we calculate for our example that $\hat{\sigma}_{SO}^2 = (33.46 - 9.05)/2 = 12.20$. Similarly, we obtain $\hat{\sigma}_S^2 = 48.65$ and $\hat{\sigma}_O^2 = 3.45$.

<Insert Table 3 about here>

Definition of the ICC

In this paper, the definition of the ICC is the ratio of a covariance term and a variance term, in accordance with the usual definition of correlation coefficients. The ICC ranges from zero, when all observed differences between participants are caused by measurement error, to one when the ability to distinguish participants from each other based on the variable of interest is not at all influenced by random error [3]. Therefore, an ICC equal to, or close to, one is the desired result when determining

the reliability of clinical assessment tools. As pointed out by Eliasziw et al [18], this definition is not the same as that used by other authors (eg. Fleiss [8]), who define ICCs as ratios of variance components. However, the method described here allows the simultaneous assessment of inter- and intra-observer reliability, which is not directly possible when using any of the other methods.

The case of *random observers*

The ICC for *inter-observer* reliability is: $ICC_{\text{inter}} = \text{cov}(Y_{ijk,}Y_{ilk}) / \text{var}(Y_{ijk})$, where *j* and *l* refer to different observers. This may then be estimated using the formula:

$$\hat{ICC}_{\text{inter}} = \frac{\hat{\sigma}_S^2}{\hat{\sigma}_S^2 + \hat{\sigma}_O^2 + \hat{\sigma}_{SO}^2 + \hat{\sigma}_e^2}$$

Each of the variance components may be estimated from Table 2.

For *intra-observer* reliability, the formula is $ICC_{\text{intra}} = \text{cov}(Y_{ijk,}Y_{ijl}) / \text{var}(Y_{ijk})$, where *k* and *l* refer to different measurements taken by the same observer on the same subject. This may be estimated using the formula:

$$\hat{ICC}_{\text{intra}} = \frac{\hat{\sigma}_S^2 + \hat{\sigma}_O^2 + \hat{\sigma}_{SO}^2}{\hat{\sigma}_S^2 + \hat{\sigma}_O^2 + \hat{\sigma}_{SO}^2 + \hat{\sigma}_e^2}$$

For our example, substitution of the calculated values into the formulas for *random observers* gives $\hat{ICC}_{\text{inter}} = 0.66$ and $\hat{ICC}_{\text{intra}} = 0.88$.

If we had used the mean of each observer's ratings to calculate the inter-observer reliability using the ICC(2,1) formula of Shrout and Fleiss [17], the estimated inter-observer reliability would be 0.92 (compared to our value of 0.66), which is much higher than that based on the individual observations. This is problematic because this would the reliability of the mean of two measurements, and not the reliability of individual measurements.

The case of *fixed observers*

Just as in the case above, the reliability coefficients are calculated as the ratio of a covariance and a variance term. However, we now need to use the right hand side of Table 2 to estimate the reliability coefficients, and so the formulas for the calculating the ICC is different in this case. The formulas are:

$$\hat{ICC}_{\text{inter}} = \frac{\hat{\sigma}^2_S - \hat{\sigma}^2_{SO}/o}{\hat{\sigma}^2_S + (o-1)\hat{\sigma}^2_{SO}/o + \hat{\sigma}^2_e}$$

and:

$$\hat{ICC}_{\text{intra}} = \frac{\hat{\sigma}^2_S + (o-1)\hat{\sigma}^2_{SO}/o}{\hat{\sigma}^2_S + (o-1)\hat{\sigma}^2_{SO}/o + \hat{\sigma}^2_e}$$

Once again, each of the estimates of the variance components can be estimated through the use of subtraction from Table 2.

Hypothesis tests to test whether the reliability meets a specified level

Hypothesis tests can be easily used to test whether the observed reliability meets a specified level [17, 18]. There are no universally applicable standards as to how high the ICC must be to constitute acceptable reliability, as this depends on the purpose, the use and consequences resulting from the assessment [7]. For example, an ICC of 0.6 may be considered appropriate within the context of a pre-participation screening for sports injury research. However, this may not be appropriate for a clinical assessment that will directly influence the choice of treatment for a patient. It should be noted that it is usually appropriate only to consider one sided hypothesis tests to determine whether the observed reliability coefficients meet a specified level of reliability.

The hypothesis test for the *inter-observer* ICC is as follows: the null hypothesis as $H_0 : ICC \le \lambda$ and the alternative as $H_1 : ICC > \lambda,$ where λ is a specified value between 0 and 1. The test statistic is:

$$F_{\text{inter}} = \frac{1-\lambda}{1+\left((o-1)\lambda\right)} \times \frac{MSS}{MSSO},$$

which may be compared against an F distribution with degrees of freedom (n-1) and (n-1)(o-1). Although this test statistic applies to both fixed and random observer effects, the relevant mean squares (MMS and MSSO) need to be taken from the appropriate part of Table 2.

Similarly, for the *intra-observer* reliability, a test of the hypothesis $H_0 : ICC \le \lambda$ against the alternative $H_1 : ICC > \lambda,$ where λ is between 0 and 1, has the test statistic:

$$F_{\text{intra}} = \frac{1-\lambda}{1+\left((m-1)\lambda\right)} \times \frac{MSS/o}{MSSO},$$

which may be compared with an F distribution with degrees of freedom (n-1) and n(m-1). Again, this test statistic applies to either fixed or random observer effects, but, as before, the appropriate mean squares need to be used from Table 2.

In our example of hip ROM assessment in fast bowlers, we may be interested in determining whether or not our ICC_{inter} is larger than 0.2. In this case, to test $H_0 : ICC_{\text{inter}} \leq 0.2$ versus the alternative $H_1 : ICC_{inter} > 0.2,$ substitution of the calculated values from Table 3 gives a test statistic 4.54, which is compared against an *F* distribution with degrees of freedom 9 and 9, yielding a p-value of 0.02.

Confidence intervals and sample size

Although it is possible to calculate confidence intervals for ICCs, the formulas are long and complicated, and are therefore included in the attached supplemental file Appendix 2. Application of Appendix 2 to our example, leads to a 95% CI of 0.253 to 0.896 for *inter-observer* reliability. For *intra-observer* reliability, the 95% CI is 0.539 to 0.961.

It is beyond the scope of this paper to discuss the sample sizes needed for reliability studies, though it is emphasised that this should be taken into account in their design. To obtain precise estimates of reliability coefficients, it is important to enrol an adequate number of subjects into a trial. The reader is referred to the paper by Walter and colleagues [21] for details of these calculations.

Measurement error and its relationship to reliability

Measurement error, often called the *standard error of measurement (*SEM*)*, is particularly important in clinical applications, where it used to detect real changes from those that could have occurred by chance alone. For intra-observer reliability, the formula for the SEM is:

$$\mathrm{SEM}_{\mathrm{intra}} = \hat{\sigma}_e = \sqrt{MSE}$$

The formula for SEM for inter-observer reliability is given for the case of *random observers* by:

$$\mathrm{SEM}_{\mathrm{inter,random}} = \sqrt{\hat{\sigma}_O^2 + \hat{\sigma}_{SO}^2 + \hat{\sigma}_e^2} = \sqrt{MSO + (n-1)MSSO + n(m-1)MSE/(mn)}$$

For the case of *fixed observers*, it is given by:

$$\mathrm{SEM}_{\mathrm{inter,fixed}} = \sqrt{\hat{\sigma}_{SO}^2 + \hat{\sigma}_e^2} = \sqrt{MSSO + (m-1)MSE/(m)}$$

In all cases (inter/intra and fixed/random), the ICC, the SEM and the variability of the measurements are related through the formula:

$$\mathrm{ICC} = 1 - \frac{\mathrm{SEM}^2}{\mathrm{var}(Y_{ijk})}$$

For this reason, one should be cautious about interpreting reliability coefficients from different studies, which may have been calculated on very different populations. For example, suppose that an observer measures hip external rotation with SEM of 5, so that readings are accurate to within ± 5 degrees. If the observers measure a population that is relatively variable (eg. variance of measurements = 25), then the ICC will be 0.8. However, if the same observer measures a population with a smaller

variability (eg. variance = 10), then the ICC will be 0.5. For this reason, it is important to report the variability of measurements and the standard error of measurement, as well as the ICC, so that comparisons across studies can be made.

Concluding remarks

Sports injury prevention requires a firm evidence base. An important component of this is the accuracy and reliability of measurements taken in studies of risk factors. When measurements are not reliable, it is difficult to distinguish between participants with or without risk factors because of the large measurement error.

There are many instances in which it would be advantageous to simultaneously assess inter- and intra-observer reliability. However, most of the commonly used methods of reliability assessment do not allow this. This paper presents a method, and worked example, for the valid calculation of both inter-observer and intra-observer reliability in the same study at the one time, and so has significant advantages over other approaches under these circumstances.

It is important for researchers to report the variability of the collected measurements and the standard error of measurement, as well as the ICC, so that full comparisons across studies about the reliability of measurements can be made.

References

1. Bahr, R. and Krosshaug, T., Understanding injury mechanisms: A key component of preventing injuries in sport. *British Journal of Sports Medicine*, 2005, 39(6): 324-329.
2. Bahr, R. and Holme, I., Risk factors for sports injuries - a methodological approach. *British Journal of Sports Medicine*, 2003, 37(5): 384-392.
3. Haas, M., How to evaluate intraexaminer reliability using an interexaminer reliability study design. *Journal of Manipulative and Physiological Therapeutics*, 1995, 18(1): 10-5.
4. Thomas, J.R. and Nelson, J.K., *Measuring research variables*, in *Research methods in physical activity*. 1996, Human Kinetics: Champaign, USA.
5. Hopkins, W.G., Measures of reliability in sports medicine and science. *Sports Medicine*, 2000, 30(1): 1-15.
6. Batterham, A.M. and George, K.P., Reliability in evidence-based clinical practice: A primer for allied health professionals. *Physical Therapy in Sport*, 2003, 4(3): 122-128.
7. Downing, S.M., Reliability: On the reproducibility of assessment data. *Medical Education*, 2004, 38(9): 1006-12.
8. Fleiss, J.L., *Reliability of measurement*, in *The design and analysis of clinical experiments*, Fleiss, J.L., Editor. 1986, John Wiley & Sons: New York.
9. Gajdosik, R.L. and Bohannon, R.W., Clinical measurement of range of motion. Review of goniometry emphasizing reliability and validity. *Physical Therapy*, 1987, 67(12): 1867-72.
10. Bennell, K.L., Talbot, R.C., Wajswelner, H., et al., Intra-rater and inter-rater reliability of a weight-bearing lunge measure of ankle dorsiflexion. *Australian Journal of Physiotherapy*, 1998, 44(3): 175-180.

11. Click Fenter, P., Bellew, J.W., Pitts, T.A., et al., Reliability of stabilised commercial dynamometers for measuring hip abduction strength: A pilot study. *British Journal of Sports Medicine*, 2003, 37(4): 331-334.
12. Gabbe, B.J., Bennell, K.L., Wajswelner, H., et al., Reliability of common lower extremity musculoskeletal screening tests. *Physical Therapy in Sport*, 2004, 5(2): 90-97.
13. MacDermid, J.C., Chesworth, B.M., Patterson, S., et al., Intratester and intertester reliability of goniometric measurement of passive lateral shoulder rotation. *Journal of Hand Therapy*, 1999, 12(3): 187-92.
14. Scott, D.A., Bond, E.Q., Sisto, S.A., et al., The intra- and interrater reliability of hip muscle strength assessments using a handheld versus a portable dynamometer anchoring station. *Archives of Physical Medicine and Rehabilitation*, 2004, 85(4): 598-603.
15. Shultz, S.J., Nguyen, A.D., Windley, T.C., et al., Intratester and intertester reliability of clinical measures of lower extremity anatomic characteristics: Implications for multicenter studies. *Clinical Journal of Sport Medicine*, 2006, 16(2): 155-61.
16. Tousignant, M., Boucher, N., Bourbonnais, J., et al., Intratester and intertester reliability of the Cybex electronic digital inclinometer (EDI-320) for measurement of active neck flexion and extension in healthy subjects. *Manual Therapy*, 2001, 6(4): 235-41.
17. Shrout, P.E. and Fleiss, J.L., Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 1979, 86(2): 420-428.
18. Eliasziw, M., Young, S.L., Woodbury, M.G., et al., Statistical methodology for the concurrent assessment of interrater and intrarater reliability: Using

goniometric measurements as an example. *Physical Therapy*, 1994, 74(8): 777-88.

19. Dennis, R.J. (2005). *Risk factors for repetitive microtrauma injury to adolescent and adult cricket fast bowlers*. [PhD thesis]. School of Safety Science, Faculty of Science. University of New South Wales.

20. Harvey, D., *Screening test protocols. Pre-participation screening of athletes: A program developed by the Australian Institute of Sport, Olympic Athlete Program and the Australian Sports Injury Prevention Taskforce*. 1998, Canberra, Australia: Australian Sports Commission.

21. Walter, S.D., Eliasziw, M., and Donner, A., Sample size and optimal designs for reliability studies. *Statistics in Medicine*, 1998, 17(1): 101-10.

Table 1: Overview of the methods used to determine inter-observer and intra-observer reliability in selected studies

Study	Focus of reliability assessment	Number of observers	Number of subjects	Number of sessions	How test score established (for each Observer in each Session)	Inter-observer reliability		Intra-observer reliability	
						Statistical method used *	Basis of reliability calculation	Statistical method used *	Basis of reliability calculation
Bennell et al [10]	Measure of ankle dorsiflexion	Inter-observer: 4 Intra-observer: 2	13	Inter-observer: 1 Intra-observer: 2	Mean of 3 trials	ICC (2,3)	Mean results of Observer 1, Observer 2, Observer 3 and Observer 4 compared using data from one Session	ICC (3,3)	Mean results of Session 1 and Session 2 compared within each of the 2 Observers
Click Fenter et al [11]	Dynamometers used to measure hip abduction strength	2	10	2	1 trial	ICC (2,1)	Mean results of Observer 1 and Observer 2 compared using data from both Sessions combined	ICC (2,1)	Mean results of Session 1 and Session 2 compared within each of the 2 Observers
Gabbe et al [12]	Measures of the lower extremity	2	15	2	Mean of 2 trials	ICC (2,1)	Mean results of Observer 1 and Observer 2 compared using data from Session 1 only	ICC (3,1)	Mean results of Session 1 and Session 2 compared within each of the 2 Observers
MacDermid et al [13]	Measure of shoulder rotation	2	34	2	1 trial	ICC (no further detail given)	Mean results of Observer 1 and Observer 2 compared within each Session	ICC (no further detail given)	Mean results of Session 1 and Session 2 compared within each of the 2 Observers
Scott et al [14]	Dynamometers used to measure hip muscle strength	2	15	2	Two test scores recorded: • Mean of 3 trials • Maximum value of the 3 trials	ICC calculated from 2 way mixed ANOVA	Compared in 4 ways: • Observer 1 (Session 1) compared with Observer 2 (Session 2)	ICC calculated from 1 way random effects ANOVA	Mean results of Session 1 and Session 2 compared within each of the 2 Observers

							• Observer 1 (Session 2) compared with Observer 2 (Session 1) • Observer 1 (Session 1) compared with Observer 2 (Session 1) • Observer 1 (Session 2) compared with Observer 2 (Session 2)		
Shultz et al [15]	Measures of lower extremity anatomic characteristics	4	16	2	Mean of 3 trials	ICC (2,1) calculated from repeated measures ANOVA	Mean results of Observers 1, 2, 3 and 4 compared within each Session	ICC (2,*k*) calculated from repeated measures ANOVA	Mean results of Session 1 and Session 2 compared within each of the 4 Observers
Tousignant et al [16]	Inclinometer used to measure neck flexion and extension	2	44	2	1 trial	ICC (2,1) calculated from 2 way random effects ANOVA	Mean results of Observer 1 and Observer 2 compared within each Session	ICC (1,1) calculated from 1 way random effects ANOVA	Mean results of Session 1 and Session compared within each of the 2 Observers

* Please refer to the relevant paper for the authors' definition of the statistical methods used to determine reliability

Table 2: Expected mean square values from a two-way analysis of variance

			Random observer case	**Fixed observer case**
Source of variation	**Degrees of freedom**	**Observed mean square**	**Expected mean square**	**Expected mean square**
Subjects	*n-1*	MSS	$mo\sigma_S^2 + m\sigma_{SO}^2 + \sigma_e^2$	$mo\sigma_S^2 + \sigma_e^2$
Observers	*o-1*	MSO	$mn\sigma_O^2 + m\sigma_{SO}^2 + \sigma_e^2$	$mn\sum_{j=1}^{o} O_j^2/(o-1) + m\sigma_{SO}^2 + \sigma_e^2$
Subjects × observers	*(n-1)(o-1)*	MSSO	$m\sigma_{SO}^2 + \sigma_e^2$	$m\sigma_{SO}^2 + \sigma_e^2$
Error	*no(m-1)*	MSE	σ_e^2	σ_e^2
Total	*mno-1*			

Table 3: Analysis of variance table for determining inter- and intra-observer reliability for hip external rotation in cricket fast bowlers

Source of variation	Degrees of freedom	Sum of squares	Observed mean square
Subjects	9	2052.60	228.07 (MSS)
Observers	1	102.40	102.40 (MSO)
Subjects × observers	9	301.10	33.46 (MSSO)
Error	20	181.00	9.05 (MSE)
Total	29		

Appendix 1: Data collected for the reliability assessment of the hip external rotation test used with cricket fast bowlers

Bowler	Observer 1		Observer 2	
	Session 1 (degrees)	Session 2 (degrees)	Session 1 (degrees)	Session 2 (degrees)
1	52	53	47	47
2	69	64	54	56
3	55	57	49	52
4	40	37	28	27
5	47	38	47	42
6	43	41	42	49
7	52	48	42	44
8	43	46	46	46
9	36	46	38	39
10	40	42	43	47

Appendix 2: Confidence intervals for the ICC

Confidence intervals can be constructed for inter- and intra-observer reliability. The formulas for the one-sided confidence intervals below are equivalent to the null hypotheses presented in the paper in the sense that the null hypothesis will be rejected at level α whenever the lower limit of the $100(1-\alpha)\%$ one-sided lower confidence is greater than the value used in the hypothesis test. These confidence intervals are based on earlier work in [see references 23 and 26 in the main paper].

Lower one-sided intervals

Inter-observer reliability

Random observers

An approximate $100(1-\alpha)\%$ one sided-lower limit confidence interval is given by

$$\left(\frac{n(MSS - FMSSO)}{nMSS + F\{o(MSO - MSSO) + n(o-1)MSSO + no(m-1)MSE\}}, 1\right),$$

where F is the $100(1-\alpha)$ th percentile point of the F distribution with $(n-1)$ and k_1 degrees of freedom. The denominator degrees of freedom has the following complicated expression

$$k_1 = \frac{(n-1)(o-1)[o\rho(MSO - MSSO) + n(1+(o-1)\rho)MSSO + no(m-1)\rho MSE]^2}{(n-1)(o\rho)^2 MSO^2 + [n(1+(o-1)\rho) - o\rho]^2 MSSO^2 + (n-1)(o-1)[no(m-1)]\rho^2 MSE^2}$$

where ρ is the estimated inter-observer ICC.

Fixed observers

An approximate $100(1-\alpha)\%$ one sided-lower limit confidence interval is given by

$$\left(\frac{n(MSS - FMSSO)}{nMSS + F\{n(o-1)MSSO + no(m-1)MSE\}}, 1\right),$$

where F is the $100(1-\alpha)$ th percentile point of the F distribution with $(n-1)$ and k_2 degrees of freedom, where denominator degrees of freedom k is given by

$$k_2 = \frac{(n-1)(o-1)[n(1+(o-1)\rho)MSSO + no(m-1)\rho MSE]^2}{(n(1+(o-1)\rho))^2 MSSO^2 + (n-1)(o-1)(no(m-1))\rho^2 MSE^2}$$

where ρ is the estimated inter-observer ICC.

Intra-observer reliability

The $100(1-\alpha)\%$ one sided-lower limit confidence interval is given by $\left(\frac{MSS/o - FMSE}{MSS/o + F(m-1)MSE},1\right)$,
where F denotes the $100(1-\alpha)$ th percentile point of the F distribution with $(n-1)$ and $n(m-1)$ degrees of freedom.

Two-sided intervals

Inter-observer reliability

Random observers

Two-sided confidence intervals can also be derived. An approximate $100(1-\alpha)\%$ confidence interval is given by (LL,UL)

$$LL = \frac{n(MSS - F_L MSSO)}{nMSS + F_L\{o(MSO - MSSO) + n(o-1)MSSO + no(m-1)MSE\}}$$

and

$$UL = \frac{n(F_U MSS - MSSO)}{F_U \times nMSS + \{o(MSO - MSSO) + n(o-1)MSSO + no(m-1)MSE\}},$$

where F_L denotes the $100(1-\alpha/2)$ th percentile point of the F distribution with $(n-1)$ and k_1 degrees of freedom (as given above) and F_U denotes the $100(1-\alpha/2)$ th percentile point of the F distribution with k_1 and $(n-1)$ and degrees of freedom.

Fixed observers

An approximate $100(1-\alpha)\%$ two-sided confidence interval is given by (LL,UL), where

$$LL = \frac{n(MSS - F_L MSSO)}{nMSS + F_L\{n(o-1)MSSO + no(m-1)MSE\}}$$

and

$$UL = \frac{n(F_U MSS - MSSO)}{F_U \times nMSS + \{n(o-1)MSSO + no(m-1)MSE\}}$$

where F_L denotes the $100(1-\alpha/2)$ th percentile point of the F distribution with $(n-1)$ and k_2 degrees of freedom (as given above) and F_U denotes the

$100(1-\alpha/2)$ th percentile point of the F distribution with k_2 and $(n-1)$ and degrees of freedom.

Intra-observer reliability

The $100(1-\alpha)\%$ two sided-lower limit confidence interval is given by (LL, UL), where

$$LL = \frac{MSS/o - F_L MSE}{MSS/o + F_L(m-1)MSE}$$

and

$$UL = \frac{F_U MSS/o - MSE}{F_U \times MSS/o + (m-1)MSE}$$

where F_L denotes the $100(1-\alpha/2)$ th percentile point of the F distribution with $(n-1)$ and $n(m-1)$ degrees of freedom and F_U denotes the $100(1-\alpha/2)$ th percentile point of the F distribution with $n(m-1)$ and $(n-1)$ degrees of freedom.

EXAMPLE

For the external hip rotation data (see main paper), two-sided 95% confidence intervals for inter- and intra-observer reliability can be derived using the values in Table 3.

For inter-observer reliability, the degrees of freedom are $n-1=9$ and $k_1 =$ 12.21. Using a computer package, it can be shown that $F_L = 3.436$ and $F_U = 3.860$. Substituting in values into the equations above, gives a confidence interval of 0.253 to 0.896.

Similarly, for intra-observer reliability, the degrees of freedom are $n-1=9$ and $n(m-1) = 10 \times (2-1) = 10$. Using a computer package, it can be shown that $F_L = 3.779$ and $F_U = 3.964$. Substitution of values into the formulas given above gives a confidence interval of 0.539 to 0.961