Journal of Strength and Conditioning Research, 2005, 19(1), 231–240

Brief Review

QUANTIFYING TEST-RETEST RELIABILITY USING THE INTRACLASS CORRELATION COEFFICIENT AND THE *SEM*

JOSEPH P. WEIR

Applied Physiology Laboratory, Division of Physical Therapy, Des Moines University—Osteopathic Medical Center, Des Moines, Iowa 50312.

ABSTRACT. Weir, J.P. Quantifying test-retest reliability using the intraclass correlation coefficient and the *SEM*. *J. Strength Cond. Res. 19(1):231–240. 2005.*—Reliability, the consistency of a test or measurement, is frequently quantified in the movement sciences literature. A common metric is the intraclass correlation coefficient (ICC). In addition, the *SEM*, which can be calculated from the ICC, is also frequently reported in reliability studies. However, there are several versions of the ICC, and confusion exists in the movement sciences regarding which ICC to use. Further, the utility of the *SEM* is not fully appreciated. In this review, the basics of classic reliability theory are addressed in the context of choosing and interpreting an ICC. The primary distinction between ICC equations is argued to be one concerning the inclusion (equations 2,1 and 2,k) or exclusion (equations 3,1 and 3,k) of systematic error in the denominator of the ICC equation. Inferential tests of mean differences, which are performed in the process of deriving the necessary variance components for the calculation of ICC values, are useful to determine if systematic error is present. If so, the measurement schedule should be modified (removing trials where learning and/or fatigue effects are present) to remove systematic error, and ICC equations that only consider random error may be safely used. The use of ICC values is discussed in the context of estimating the effects of measurement error on sample size, statistical power, and correlation attenuation. Finally, calculation and application of the *SEM* are discussed. It is shown how the *SEM* and its variants can be used to construct confidence intervals for individual scores and to determine the minimal difference needed to be exhibited for one to be confident that a true change in performance of an individual has occurred.

KEY WORDS. reproducibility, precision, error, consistency, *SEM*, intraclass correlation coefficient

INTRODUCTION

Reliability refers to the consistency of a test or measurement. For a seemingly simple concept, the quantifying of reliability and interpretation of the resulting numbers are surprisingly unclear in the biomedical literature in general (49) and in the sport sciences literature in particular. Part of this stems from the fact that reliability can be assessed in a variety of different contexts. In the sport sciences, we are most often interested in simple test-retest reliability; this is what Fleiss (22) refers to as a simple reliability study. For example, one might be interested in the reliability of 1 repetition maximum (1RM) squat measures taken on the same athletes over different days. However, if one is interested in the ability of different testers to get the same results from the same subjects on skinfold measurements, one is now interested in the interrater reliability. The quantifying of reliability in these different situations is not necessarily the same, and the decisions regarding how to calculate reliability in these different contexts has not been adequately addressed in the sport sciences literature. In this article, I focus on test-retest reliability (but not limited in the number of retest trials). In addition, I discuss data measured on a continuous scale.

Confusion also stems from the jargon used in the context of reliability, i.e., consistency, precision, repeatability, and agreement. Intuitively, these terms describe the same concept, but in practice some are operationalized differently. Notably, reliability and agreement are not synonymous (30, 49). Further, reliability, conceptualized as consistency, consists of both absolute consistency and relative consistency (44). Absolute consistency concerns the consistency of scores of individuals, whereas relative consistency concerns the consistency of the position or rank of individuals in the group relative to others. In the fields of education and psychology, the term reliability is operationalized as relative consistency and quantified using reliability coefficients called intraclass correlation coefficients (ICCs) (49). Issues regarding quantifying ICCs and their interpretation are discussed in the first half of this article. Absolute consistency, quantified using the *SEM*, is addressed in the second half of the article. In brief, the *SEM* is an indication of the precision of a score, and its use allows one to construct confidence intervals (CIs) for scores.

Another confusing aspect of reliability calculations is that a variety of different procedures, besides ICCs and *SEM*, have been used to determine reliability. These include the Pearson *r*, the coefficient of variation, and the LOA (Bland-Altman plots). The Pearson product moment correlation coefficient (Pearson *r*) was often used in the past to quantify reliability, but the use of the Pearson *r* is typically discouraged for assessing test-retest reliability (7, 9, 29, 33, 44); however, this recommendation is not universal (43). The primary, although not exclusive, weakness of the Pearson *r* is that it cannot detect systematic error. More recently, the limits of agreement (LOA) described by Bland and Altman (10) have come into vogue in the biomedical literature (2). The LOA will not be addressed in detail herein other than to point out that the procedure was developed to examine agreement between 2 different techniques of quantifying some variable (so-called method comparison studies, e.g., one could compare testosterone concentration using 2 different bioassays), not reliability per se. The use of LOA as an index of reliability has been criticized in detail elsewhere (26, 49).

In this article, the ICC and *SEM* will be the focus.

Unfortunately, there is considerable confusion concerning both the calculation and interpretation of the ICC. Indeed, there are 6 common versions of the ICC (and others as well), and the choice of which version to use is not intuitively obvious. Similarly, the *SEM*, which is intimately related to the ICC, has useful applications that are not fully appreciated by practitioners in the movement sciences. The purposes of this article are to provide information on the choice and application of the ICC and to encourage practitioners to use the *SEM* in the interpretation of test data.

The ICC

Reliability Theory

For a group of measurements, the total variance (σ_T^2) in the data can be thought of as being due to true score variance (σ_t^2) and error variance (σ_e^2). Similarly, each observed score is composed of the true score and error (44). The theoretical true score of an individual reflects the mean of an infinite number of scores from a subject, whereas error equals the difference between the true score and the observed score (21). Sources of error include errors due to biological variability, instrumentation, error by the subject, and error by the tester. If we make a ratio of the σ_t^2 to the σ_T^2 of the observed scores, where σ_T^2 equals σ_t^2 plus σ_e^2, we have the following reliability coefficient:

$$R = \frac{\sigma_t^2}{\sigma_t^2 + \sigma_e^2}. \quad (1)$$

The closer this ratio is to 1.0, the higher the reliability and the lower the σ_e^2. Since we do not know the true score for each subject, an index of the σ_t^2 is used based on between-subjects variability, i.e., the variance due to how subjects differ from each other. In this context, reliability (relative consistency) is formally defined (5, 21, 49) as follows:

$$\text{reliability} = \frac{\text{between subjects variability}}{\text{between subjects variability} + \text{error}}. \quad (2)$$

The reliability coefficient in Equation 2 is quantified by various ICCs. So although reliability is conceptually aligned with terms such as reproducibility, repeatability, and agreement, it is defined as above. The necessary variance estimates are derived from analysis of variance (ANOVA), where appropriate mean square values are recorded from the computer printout. Specifically, the various ICCs can be calculated from mean square values derived from a within-subjects, single-factor ANOVA (i.e., a repeated-measures ANOVA).

The ICC is a relative measure of reliability (18) in that it is a ratio of variances derived from ANOVA, is unitless, and is more conceptually akin to R^2 from regression (43) than to the Pearson r. The ICC can theoretically vary between 0 and 1.0, where an ICC of 0 indicates no reliability, whereas an ICC of 1.0 indicates perfect reliability. In practice, ICCs can extend beyond the range of 0 to 1.0 (30), although with actual data this is rare. The relative nature of the ICC is reflected in the fact that the magnitude of an ICC depends on the between-subjects variability (as shown in the next section). That is, if subjects differ little from each other, ICC values are small even if trial-to-trial variability is small. If subjects differ from each other a lot, ICCs can be large even if trial-to-trial variability is large. Thus, the ICC for a test is context specific (38, 51). As noted by Streiner and Norman (49), "There is literally no such thing as the reliability of a test, unqualified; the coefficient has meaning only when applied to specific populations." Further, it is intuitive that small differences between individuals are more difficult to detect than large ones, and the ICC is reflective of this (49).

Error is typically considered as being of 2 types: systematic error (e.g., bias) and random error (2, 39). (Generalizability theory expands sources of error to include various facets of interest but is beyond the scope of this article.) Total error reflects both systematic error and random error (imprecision). Systematic error includes both constant error and bias (38). Constant error affects all scores equally, whereas bias is systematic error that affects certain scores differently than others. For physical performance measures, the distinction between constant error and bias is relatively unimportant and the focus here on systematic error is on situations that result in a unidirectional change in scores on repeated testing. In testing of physical performance, subjects may improve their test scores simply due to learning effects, e.g., performing the first test serves as practice for subsequent tests, or fatigue or soreness may result in poorer performance across trials. In contrast, random error refers to sources of error that are due to chance factors. Factors such as luck, alertness, attentiveness by the tester, and normal biological variability affect a particular score. Such errors should, in a random manner, both increase and decrease test scores on repeated testing. Thus, we can expand Equation 1 as follows:

$$R = \frac{\sigma_t^2}{\sigma_t^2 + \sigma_{se}^2 + \sigma_{re}^2}, \quad (3)$$

where σ_{se}^2 is the systematic σ_e^2 and σ_{re}^2 is the random σ_e^2.

It has been argued that systematic error is a concern of validity and not reliability (12, 43). Similarly, systematic error (e.g., learning effects, fatigue) has been suggested to be a natural phenomenon and therefore does not contribute to unreliability per se in test-retest situations (43). Thus, there is a school of thought that suggests that only random error should be assessed in reliability calculations. Under this analysis, the error term in the denominator will only reflect random error and not systematic error, increasing the size of reliability coefficients. The issue of inclusion of systematic error in the determination of reliability coefficients is addressed in a subsequent section.

The Basic Calculations

The calculation of reliability starts with the performance of a repeated-measures ANOVA. This analysis performs 2 functions. First, the inferential test of mean differences across trials is an assessment of systematic error (trend). Second, all of the subsequent calculations can be derived from the output from this ANOVA. In keeping with the nomenclature of Keppel (28), the ANOVA that is used is of a single-factor, within-subjects (repeated-measures) design. Unfortunately, the language gets a bit tortured in many sources, because the different ICC models are referred to as either 1-way or 2-way models; what is important to keep in mind is that both the 1-way and 2-way ICC models can be derived from the same single-factor, within-subjects ANOVA.

TABLE 1. Example data set.

Trial A1	Trial A2	Δ	Trial B1	Trial B2	Δ
146	140	−6	166	160	−6
148	152	+4	168	172	+4
170	152	−18	160	142	−18
90	99	+9	150	159	+9
157	145	−12	147	135	−12
156	153	−3	146	143	−3
176	167	−9	156	147	−9
205	218	+13	155	168	+13
156 ± 33	153 ± 33		156 ± 8	153 ± 13	

TABLE 2. Two-way analysis of variance summary table for data set A.*

Source	*df*	SS	Mean square	*F*	*p* value
Between subjects	7	14,689.8	2098.4 (MS_B: 1-way) (MS_S: 2-way)	36.8	
Within subjects	8	430	53.75 (MS_W)		
Trials	1	30.2	30.2 (MS_T)	0.53	0.49
Error	7	399.8	57 (MS_E)		
Total	15	15,119.8			

* MS_B = between-subjects mean square; MS_E = error mean square; MS_S = subjects mean square; MS_T = trials mean square; MS_W = within-subjects mean square; SS = sums of squares.

To illustrate the calculations, example data are presented in Table 1. ANOVA summary tables are presented in Tables 2 and 3, and the resulting ICCs are presented in Table 4. Focus on the first two columns of Table 1, which are labeled trial A1 and trial A2. As can be seen, there are 2 sets (columns) of scores, and each set has 8 scores. In this example, each of 8 subjects has provided a score in each set. Assume that each set of scores represents the subjects' scores on the 1RM squat across 2 different days (trials). A repeated-measures ANOVA is performed to primarily test whether the 2 sets of scores are significantly different from each other (i.e., do the scores systematically change between trials) and is summarized in Table 2. Equivalently, one could have used a paired *t*-test, since there were only 2 levels of trials. However, the ANOVA is applicable to situations with 2 or more trials and is consistent with the ICC literature in defining sources of variance for ICC calculations. Note that there are 3 sources of variability in Table 2: subjects, trials, and error. In a repeated-measures ANOVA such as this, it is helpful to remember that this analysis might be considered as having 2 factors: the primary factor of trials and a secondary factor called subjects (with a sample size of 1 subject per cell). The error term includes the interaction effect of trials by subjects. It is useful to keep these sources of variability in mind for 2 reasons. First, the 1-way and 2-way models of the ICC (6, 44) either collapse the variability due to trials and error together (1-way models) or keep them separate (2-way models). Note that the trials and error sources of variance, respectively, reflect the systematic and random sources of error in the σ_e^2 of the reliability coefficient. These differences are illustrated in Table 2, where the *df* and sums of squares values for error in the 1-way model (within-subjects source) are simply the sum of the respective values for trials and error in the 2-way model.

Second, unlike a between-subjects ANOVA where the "noise" due to different subjects is part of the error term, the variability due to subjects is now accounted for (due to the repeated testing) and therefore not a part of the error term. Indeed, for the calculation of the ICC, the numerator (the signal) reflects the variance due to subjects. Since the error term of the ANOVA reflects the interaction between subjects and trials, the error term is small in situations where all the subjects change similarly across test days. In situations where subjects do not change in a similar manner across test days (e.g., some subjects' scores increase, whereas others decrease), the error term is large. In the former situation, even small differences across test days, as long as they are consistent across all the subjects, can result in a statistically significant effect for trials. In this example, however, the effect for trials is not statistically significant ($p = 0.49$), indicating that there is no statistically significant systematic error in the data. It should be kept in mind, however, that the statistical power of the test of mean differences between trials is affected by sample size and random error. Small sample sizes and noisy data (i.e., high random error) will decrease power and potentially hide systematic error. Thus, an inferential test of mean differences alone is insufficient to quantify reliability. Further, evaluation of the effect for trials ought to be evaluated with a more liberal α measure, since in this case, the implications of a type 2 error are more severe than a type 1 error. In cases where systematic error is present, it may be prudent to change the measurement schedule (e.g., add trials if a learning effect is present or increase rest intervals if fatigue is present) to compensate for the bias.

Shrout and Fleiss (46) have presented 6 forms of the ICC. This system has taken hold in the physical therapy literature. However, the specific nomenclature of their system does not seem to be as prevalent in the exercise physiology, kinesiology, and sport science literature, which has instead ignored which is model used or focused on ICC terms that are centered on either 1-way or 2-way ANOVA models (6, 44). Nonetheless, the ICC models of

TABLE 3. Analysis of variance summary table for data set B.*

Source	*df*	SS	Mean square	*F*	*p* value
Between subjects	7	1330	190 (MS_B: 1-way) (MS_S: 2-way)	3.3	
Within subjects	8	430	53.75 (MS_W)		
Trials	1	30.2	30.2 (MS_T)	0.53	0.49
Error	7	399.8	57 (MS_E)		
Total	15	1760			

* MS_B = between-subjects mean square; MS_E = error mean square; MS_S = subjects mean square; MS_T = trials mean square; MS_W = within-subjects mean square; SS = sums of squares.

Shrout and Fleiss (46) overlap with the 1-way and 2-way models presented by Safrit (44) and Baumgartner (6).

Three general models of the ICC are present in the Shrout and Fleiss (46) nomenclature, which are labeled 1, 2, and 3. Each model can be calculated 1 of 2 ways. If the scores in the analysis are from single scores from each subject for each trial (or rater if assessing interrater reliability), then the ICC is given a second designation of 1. If the scores in the analysis represent the average of the k scores from each subject (i.e., the average across the trials), then the ICC is given a second designation of k. In this nomenclature then, an ICC with a model designation of 2,1 indicates an ICC calculated using model 2 with single scores. The use of these models is typically presented in the context of determining rater reliability (41). For model 1, each subject is assumed to be assessed by a different set of raters than other subjects, and these raters are assumed to be randomly sampled from the population of possible raters so that raters are a random effect. Model 2 assumes each subject was assessed by the same group of raters, and these raters were randomly sampled from the population of possible raters. In this case, raters are also considered a random effect. Model 3 assumes each subject was assessed by the same group of raters, but these particular raters are the only raters of interest, i.e., one does not wish to generalize the ICCs beyond the confines of the study. In this case, the analysis attempts to determine the reliability of the raters used by that particular study, and raters are considered a fixed effect.

The 1-way ANOVA models (6, 44) coincide with model 1,k for situations where scores are averaged and model 1,1 for single scores for a given trial (or rater). Further, ICC 1,1 coincides with the 1-way ICC model described by Bartko (3, 4), and ICC 1,k has also been termed the Spearman Brown prediction formula (4). Similarly, ICC values derived from single and averaged scores calculated using the 2-way approach (6, 44) coincide with models 3,1 and 3,k, respectively. Calculations coincident with models 2,1 and 2,k were not reported by Baumgartner (6) or Safrit (44).

More recently, McGraw and Wong (34) expanded the Shrout and Fleiss (46) system to include 2 more general forms, each also with a single score or average score version, resulting in 10 ICCs. These ICCs have now been incorporated into SPSS statistical software starting with version 8.0 (36). Fortunately, 4 of the computational formulas of Shrout and Fleiss (46) also apply to the new forms of McGraw and Wong (34), so the total number of formulas is not different.

The computational formulas for the ICC models of Shrout and Fleiss (46) and McGraw and Wong (34) are summarized in Table 5. Unfortunately, it is not intuitively obvious how the computational formulas reflect the intent of equations 1 through 3. This stems from the fact that the computational formulas reported in most sources are derived from algebraic manipulations of basic equations where mean square values from ANOVA are used to estimate the various σ^2 values reflected in equations 1 through 3. To illustrate, the manipulations for ICC 1,1 (random-effects, 1-way ANOVA model) are shown herein. First, the computational formula for ICC 1,1 is as follows:

$$\text{ICC } 1,1 = \frac{MS_B - MS_W}{MS_B + (k - 1)MS_W} \quad (4)$$

where MS_B indicates the between-subjects mean square, MS_W indicates the within-subjects mean square, and k is the number of trials (3, 46). The relevant mean square values can be found in Table 2. To relate this computational formula to equation 1, one must know that estimation of the appropriate σ^2 comes from expected mean squares from ANOVA. Specifically, for this model the expected MS_B equals σ_e^2 plus $k\sigma_s^2$, whereas the expected MS_W equals σ_e^2 (3); therefore, MS_B equals MS_W plus $k\sigma_s^2$. If from equation 1 we estimate σ_t^2 from between-subjects variance (σ_s^2), then

$$\text{ICC} = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_e^2}. \quad (5)$$

By algebraic manipulation (e.g., $\sigma_s^2 = [MS_B - MS_W]/k$) and substitution of the expected mean squares into equation 5, it can be shown that

$$\begin{aligned}\text{ICC } 1,1 &= \frac{\sigma_s^2}{\sigma_s^2 + \sigma_e^2} \\ &= \frac{\dfrac{MS_B - MS_W}{k}}{\dfrac{MS_B - MS_W}{k} + MS_W} \\ &= \frac{MS_B - MS_W}{MS_B + (k - 1)MS_W}. \end{aligned} \quad (6)$$

Similar derivations can be made for the other ICC models (3, 34, 46, 49) so that all ultimately relate to equation 1. Of note is that with the different ICC models (fixed vs. random effects, 1-way vs. 2-way ANOVA), the expected mean squares change and thus the computational formulas commonly found in the literature (30, 41) also change.

Choosing an ICC

Given the 6 ICC versions of Shrout and Fleiss (46) and the 10 versions presented by McGraw and Wong (34), the choice of ICC is perplexing, especially considering that

TABLE 4. ICC values for data sets A and B.*

ICC type	Data set A	Data set B
1,1	0.95	0.56
1,k	0.97	0.72
2,1	0.95	0.55
2,k	0.97	0.71
3,1	0.95	0.54
3,k	0.97	0.70

* ICC = intraclass correlation coefficient.

TABLE 6. Example data set with systematic error.

Trial C1	Trial C2	Δ
146	161	+14
148	162	+14
170	189	+19
90	100	+10
157	175	+18
156	171	+15
176	195	+19
205	219	+14
156 ± 33	172 ± 35	

most of the literature deals with rater reliability not test-retest reliability of physical performance measures. In a classic paper, Brozek and Alexander (11) first introduced the concept of the ICC to the movement sciences literature and detailed the implementation of an ICC for application to test-retest analysis of motor tasks. Their coefficient is equivalent to model 3,1. Thus, one might use ICC 3,1 with test-retest reliability where trials is substituted for raters. From the rater nomenclature above, if one does not wish to generalize the reliability findings but rather assert that in our hands the procedures are reliable, then ICC 3,1 seems like a logical choice. However, this ICC does not include variance associated with systematic error and is in fact closely approximated by the Pearson *r* (1, 43). Therefore, the criticism of the Pearson *r* as an index of reliability holds as well for ICCs derived from model 3. At the least, it needs to be established that the effect for trials (bias) is trivial if reporting an ICC derived from model 3. Use of effect size for the trials effect in the ANOVA would provide information in this regard. With respect to ICC 3,1, Alexander (1) notes that it "may be regarded as an estimate of the value that would have been obtained if the fluctuation [systematic error] had been avoided."

In a more general sense, there are 4 issues to be addressed in choosing an ICC: (a) 1- or 2-way model, (b) fixed- or random-effect model, (c) include or exclude systematic error in the ICC, and (d) single or mean score. With respect to choosing a 1- or 2-way model, in a 1-way model, the effect of raters or trials (replication study) is not crossed with subjects, meaning that it allows for situations where all raters do not score all subjects (48). Fleiss (22) uses the 1-way model for what he terms simple replication studies. In this model, all sources of error are lumped together into the MS_W (Tables 2 and 3). In contrast, the 2-way models allow the error to be partitioned between systematic and random error. When systematic error is small, MS_W from the 1-way model and error mean square (MS_E) from the 2-way models (reflecting random error) are similar, and the resulting ICCs are similar. This is true for both data sets A and B. When systematic error is substantial, MS_W and MS_E are disparate, as in data set C (Tables 6 and 7). Two-way models require trials or raters to be crossed with subjects (i.e., subjects provide scores for all trials or each rater rates all subjects). For test-retest situations, the design dictates that trials are crossed with subjects and therefore lend themselves to analysis by 2-way models.

Regarding fixed vs. random effects, a fixed factor is one in which all levels of the factor of interest (in this

TABLE 5. Intraclass correlation coefficient model summary table.*

Shrout and Fleiss	Computational formula	McGraw and Wong	Model
1,1	$\frac{MS_B - MS_W}{MS_B + (k-1)MS_W}$	1	1-way random
1,k	$\frac{MS_B - MS_W}{MS_B}$	k	1-way random
	Use 3,1	C,1	2-way random
	Use 3,k	C,k	2-way random
2,1	$\frac{MS_S - MS_E}{MS_S + (k-1)MS_E + \frac{k(MS_T - MS_E)}{n}}$	A,1	2-way random
2,k	$\frac{MS_S - MS_E}{MS_S + \frac{k(MS_T - MS_E)}{n}}$	A,k	2-way random
3,1	$\frac{MS_S - MS_E}{MS_S + (k-1)MS_E}$	C,1	2-way fixed
3,k	$\frac{MS_S - MS_E}{MS_S}$	C,k	2-way fixed
	Use 2,1	A,1	2-way fixed
	Use 2,k	A,k	2-way fixed

* Adapted from Shrout and Fleiss (46) and McGraw and Wong (34). Mean square abbreviations are based on the 1-way and 2-way analysis of variance illustrated in Table 2. For McGraw and Wong, A = absolute and C = consistency. MS_B = between-subjects mean square; MS_E = error mean square; MS_S = subjects mean square; MS_T = trials mean square; MS_W = within-subjects mean square.

TABLE 7. Analysis of variance summary table for data set C.*

Source	*df*	SS	Mean square	*F*	*p* value
Between subjects	7	15,925	2275 (MS_B: 1-way) (MS_S: 2-way)	482.58	
Within subjects	8	994	124.25 (MS_W)		
Trials	1	961.0	961.0 (MS_T)	203.85	***<0.0001***
Error	7	33.0	***4.71*** (MS_E)		
Total	15	16,919			

* MS_B = between-subjects mean square; MS_E = error mean square; MS_S = subjects mean square; MS_T = trials mean square; MS_W = within-subjects mean square; SS = sums of squares.

case trials) are included in the analysis and no attempt at generalization of the reliability data beyond the confines of the study is expected. Determining the reliability of a test before using it in a larger study fits this description of fixed effect. A random factor is one in which the levels of the factor in the design (trials) are but a sample of the possible levels, and the analysis will be used to generalize to other levels. For example, a study designed to evaluate the test-retest reliability of the vertical jump for use by other coaches (with similar athletes) would consider the effect of trials to be a random effect. Both Shrout and Fleiss (46) models 1 and 2 are random-effects models, whereas model 3 is a fixed-effect model. From this discussion, for the 2-way models of Shrout and Fleiss (46), the choice between model 2 and model 3 appears to hinge on a decision regarding a random- vs. fixed-effects model. However, models 2 and 3 also differ in their treatment of systematic error. As noted previously, model 3 only considers random error, whereas model 2 considers both random and systematic error. This system does not include a 2-way fixed-effects model that includes systematic error and does not offer a 2-way random-effects model that only considers random error. The expanded system of McGraw and Wong (34) includes these options. In the nomenclature of McGraw and Wong (34), the designation C refers to consistency and A refers to absolute agreement. That is, the C models consider only random error and the A models consider both random and systematic error. As noted in Table 5, no new computational formulas are required beyond those presented by Shrout and Fleiss (46). Thus, if one were to choose a 2-way random-effects model that only addressed random error, one would use equation 3,1 (or equation 3,k if the mean across k trials is the criterion score). Similarly, if one were to choose a 2-way fixed-effects model that addressed both systematic and random error, equation 2,1 would be used (or 2,k). Ultimately then, since the computational formulas do not differ between systems, the choice between using the Shrout and Fleiss (46) equations from models 2 vs. 3 hinge on decisions regarding inclusion or exclusion of systematic error in the calculations. As noted by McGraw and Wong (34), "the random-fixed effects distinction is in its effect on the interpretation, but not calculation, of an ICC."

Should systematic error be included in the ICC? First, if the effect for trials is small, the systematic differences between trials will be small, and the ICCs will be similar to each other. This is evident in both the A and B data sets (Tables 1 through 3). However, if the mean differences are large, then differences between ICCs are evident, especially between equation 3,1, which does not consider systematic error, and equations 1,1 and 2,1, which do consider systematic error. In this regard, the *F* test for trials and the ICC calculations may give contradictory results from the same data. Specifically, it can be the case that an ICC can be large (indicating good reliability), whereas the ANOVA shows a significant trials effect. An example is given in Tables 6 and 7. In this example, each score in trial C1 was altered in trial C2 so that there was a bias of +15 kg and a random component added to each score. The effect for trials was significant ($F_{1,7} = 203.85$, $p < 0.001$) and reflected a mean increase of 16 kg. For an ANOVA to be significant, the effect must be large (in this case, the mean differences between trials must be large), the noise (error term) must be small, or both. The error term is small when all subjects behave similarly across test days. When this is the case, even small mean differences can be statistically significant. In this case, the systematic differences explain a significant amount of variability in the data. Despite the rather large systematic error, the ICC values from equations 1,1; 2,1; and 3,1 were 0.896, 0.901, and 0.998, respectively. A cursory examination of just the ICC scores would suggest that the test exhibited good reliability, especially using equation 3,1, which only reflects random error. However, an approximately 10% increase in scores from trial C1 to C2 would suggest otherwise. Thus, an analysis that only focuses on the ICC without consideration of the trials effect is incomplete (31). If the effect for trials is significant, the most straightforward approach is to develop a measurement schedule that will attenuate systematic error (2, 50). For example, if learning effects are present, one might add trials until a plateau in performance occurs. Then the ICC could be calculated only on the trials in the plateau region. The identification of such a measurement schedule would be especially helpful for random-effects situations where others might be using the test being evaluated. For simplicity, all the examples here have been with only 2 levels for trials. If a trials effect is significant, however, 2 trials are insufficient to identify a plateau. The possibility of a significant trials effect should be considered in the design of the reliability study. Fortunately, the ANOVA procedures require no modification to accommodate any number of trials.

Interpreting the ICC

At one level, interpreting the ICC is fairly straightforward; it represents the proportion of variance in a set of scores that is attributable to the σ_t^2. An ICC of 0.95 means that an estimated 95% of the observed score variance is due to σ_t^2. The balance of the variance (1 − ICC = 5%) is attributable to error (51). However, how does one qualitatively evaluate the magnitude of an ICC and what can the quantity tell you? Some sources have attempted to delineate good, medium, and poor levels for the ICC, but there is certainly no consensus as to what constitutes a good ICC (45). Indeed, Charter and Feldt (15) argue that

"it is not theoretically defensible to set a universal standard for test score reliability." These interpretations are further complicated by 2 factors. First, as noted herein, the ICC varies, depending on which version of the ICC is used. Second, the magnitude of the ICC is dependent on the variability in the data (45). All other things being equal, low levels of between-subjects variability will serve to depress the ICC even if the differences between subjects' scores across test conditions are small. This is illustrated by comparing the 2 example sets of data in Table 1. Trials 1 and 2 of data sets A and B have identical mean values and identical change scores between trials 1 and 2. They differ in the variability between subjects, with greater between-subjects variability evident in data set A as shown in the larger *SD*s. In Tables 2 and 3, the ANOVA tables have identical outcomes with respect to the inferential test of the factor trials and have identical error terms (since the between-subjects variability is not part of the error term, as noted previously). Table 4 shows the ICC values calculated using the 6 different models of Shrout and Fleiss (46) on the A and B data sets. Clearly, data set B, with the lower between-subjects variability, results in smaller ICC values than data set A.

How then does one interpret an ICC? First, because of the relationship between the ICC and between-subjects variability, the heterogeneity of the subjects should be considered. A large ICC can mask poor trial-to-trial consistency when between-subjects variability is high. Conversely, a low ICC can be found even when trial-to-trial variability is low if the between-subjects variability is low. In this case, the homogeneity of the subjects means it will be difficult to differentiate between subjects even though the absolute measurement error is small. An examination of the *SEM* in conjunction with the ICC is therefore needed (32). From a practical perspective, a given test can have different reliability, at least as determined from the ICC, depending on the characteristics of the individuals included in the analysis. In the 1RM squat, combining individuals of widely different capabilities (e.g., wide receivers and defensive linemen in American football) into the same analysis increases between-subjects variability and improves the ICC, yet this may not be reflected in the expected day-to-day variation as illustrated in Tables 1 through 4. In addition, the inferential test for bias described previously needs to be considered. High between-subjects variability may result in a high ICC even if the test for bias is statistically significant.

The relationship between between-subjects variability and the magnitude of the ICC has been used as a criticism of the ICC (10, 39). This is an unfair criticism, since the ICC is used to provide information regarding inferential statistical tests not to provide an index of absolute measurement error. In essence, the ICC normalizes measurement error relative to the heterogeneity of the subjects. As an index of absolute reliability then, this is a weakness and other indices (i.e., the *SEM*) are more informative. As a relative index of reliability, the ICC behaves as intended.

What are the implications of a low ICC? First, measurement error reflected in an ICC of less than 1.0 serves to attenuate correlations (22, 38). The equation for this attenuation effect is as follows:

$$r_{xy} = \hat{r}_{xy}\sqrt{\mathrm{ICC}_x\mathrm{ICC}_y} \quad (7)$$

where r_{xy} is the observed correlation between x and y, $\hat{r}_{xy}$ is the correlation between x and y if both were measured without error (i.e., the correlations between the true scores), and ICC_x and ICC_y are the reliability coefficients for x and y, respectively. Nunnally and Bernstein (38) note that the effect of measurement error on correlation attenuation becomes minimal as ICCs increase above 0.80. In addition, reliability affects the power of statistical tests. Specifically, the lower the reliability, the greater the risk of type 2 error (14, 40). Fleiss (22) illustrates how the magnitude of an ICC can be used to adjust sample size and statistical power calculations (45). In short, low ICCs mean that more subjects are required in a study for a given effect size to be statistically significant (40). An ICC of 0.60 may be perfectly fine if the resulting effect on sample size and statistical power is within the logistical constraints of the study. If, however, an ICC of 0.60 means that, for a required level of power, more subjects must be recruited than is feasible, then 0.60 is not acceptable.

Although infrequently used in the movement sciences, the ICC of test scores can be used in the setting and interpretation of cut points for classification of individuals. Charter and Feldt (15) show how the ICC can be used to estimate the percentage of false-positive, false-negative, true-positive, and true-negative results for a clinical classification scheme. Although the details of these calculations are beyond the scope of this article, it is worthwhile to note that very high ICCs are required to classify individuals with a minimum of misclassification.

THE *SEM*

Because the general form of the ICC is a ratio of variance due to differences between subjects (the signal) to the total variability in the data (the noise), the ICC is reflective of the ability of a test to differentiate between different individuals (27, 47). It does not provide an index of the expected trial-to-trial noise in the data, which would be useful to practitioners such as strength coaches. Unlike the ICC, which is a relative measure of reliability, the *SEM* provides an absolute index of reliability. Hopkins (26) refers to this as the "typical error." The *SEM* quantifies the precision of individual scores on a test (24). The *SEM* has the same units as the measurement of interest, whereas the ICC is unitless. The interpretation of the *SEM* centers on the assessment of reliability within individual subjects (45). The direct calculation of the *SEM* involves the determination of the *SD* of a large number of scores from an individual (44). In practice, a large number of scores is not typically collected, so the *SEM* is estimated. Most references estimate the *SEM* as follows:

$$SEM = SD\sqrt{1 - \mathrm{ICC}} \quad (8)$$

where *SD* is the *SD* of the scores from all subjects (which can be determined from the ANOVA as $\sqrt{\mathrm{SS}_{\mathrm{TOTAL}}/(n - 1)}$) and ICC is the reliability coefficient. Note the similarity between the equation for the *SEM* and standard error of estimate from regression analysis. Since different forms of the ICC can result in different numbers, the choice of ICC can substantively affect the size of the *SEM*, especially if systematic error is present. However, there is an alternative way of calculating the *SEM* that avoids these uncertainties. The *SEM* can be estimated as the square root of the mean square error term from the ANOVA (20, 26, 48). Since this estimate of

the *SEM* has the advantage of being independent of the specific ICC, its use would allow for more consistency in interpreting *SEM* values from different studies. However, the mean square error terms differ when using the 1-way vs. 2-way models. In Table 2 it can be seen that using a 1-way model (22) would require the use of MS_W ($\sqrt{53.75}$ = 7.3 kg), whereas use of a 2-way model would require use of MS_E ($\sqrt{57}$ = 7.6 kg). Hopkins (26) argues that because the 1-way model combines influences of random and systematic error together, "The resulting statistic is biased high and is hard to interpret because the relative contributions of random error and changes in the mean are unknown." He therefore suggests that the error term from the 2-way model (MS_E) be used to calculate *SEM*. Note however, that in this sample, the 1-way *SEM* is smaller than the 2-way *SEM*. This is because the trials effect is small. The high bias of the 1-way model is observed when the trials effect is large (Table 7). The *SEM* calculated using the MS error from the 2-way model ($\sqrt{4.71}$ = 2.2 kg) is markedly lower than the *SEM* calculated using the 1-way model ($\sqrt{124.25}$ = 11.1 kg), since the *SEM* as defined as $\sqrt{MS_E}$ only considers random error. This is consistent with the concept of a *SE*, which defines noise symmetrically around a central value. This points to the desire of establishing a measurement schedule that is free of systematic variation.

Another difference between the ICC and *SEM* is that the *SEM* is largely independent of the population from which it was determined, i.e., the *SEM* "is considered to be a fixed characteristic of any measure, regardless of the sample of subjects under investigation" (38). Thus, the *SEM* is not affected by between-subjects variability as is the ICC. To illustrate, the MS_E for the data in Tables 2 and 3 are equal (MS_E = 57), despite large differences in between-subjects variability. The resulting *SEM* is the same for data sets A and B ($\sqrt{57}$ = 7.6 kg), but yet they have different ICC values (Table 4). The results are similar when calculating the *SEM* using equation 8, even though equation 8 uses the ICC in calculating the *SEM*, since the effects of the *SD* and the ICC tend to offset each other (38). However, the effects do not offset each other completely, and use of equation 8 results in an *SEM* estimate that is modestly affected by between-subjects variability (2).

The *SEM* is the *SE* in estimating observed scores (the scores in your data set) from true scores (38). Of course, our problem is just the opposite. We have the observed scores and would like to estimate subjects' true scores. The *SEM* has been used to define the boundaries around which we think a subject's true score lies. It is often reported (8, 17) that the 95% CI for a subject's true score can be estimated as follows:

$$T = S \pm 1.96(SEM), \quad (9)$$

where T is the subject's true score, S is the subject's observed score on the measurement, and 1.96 defines the 95% CI. However, strictly speaking this is not correct, since the *SEM* is symmetrical around the true score, not the observed score (13, 19, 24, 38), and the *SEM* reflects the *SD* of the observed scores while holding the true score constant. In lieu of equation 9, an alternate approach is to estimate the subject's true score and calculate an alternate *SE* (reflecting the *SD* of true scores while holding observed scores constant). Because of regression to the mean, obtained scores (S) are biased estimators of true scores (16, 19). Scores below the mean are biased downward, and scores above the mean are biased upward. A subject's estimated true score (T) can be calculated as follows:

$$T = \bar{X} + \mathrm{ICC}(d), \quad (10)$$

where $d = S - \bar{X}$. To illustrate, consider data set A in Table 1. With a grand mean of 154.5 and an ICC 3,1 of 0.95, an individual with an S of 120 kg would have a predicted T of 154.5 + 0.95 (120 − 154.5) = 121.8 kg. Note that because the ICC is high, the bias is small (1.8 kg). The appropriate *SE* to define the CI of the true score, which some have referred to as the standard error of estimate (13), is as follows (19, 38):

$$SEM_{TS} = SD\sqrt{\mathrm{ICC}(1 - \mathrm{ICC})}. \quad (11)$$

In this example the value is 31.74 $\sqrt{0.95}$ (1 − 0.95) = 6.92, where 31.74 equals the *SD* of the observed scores around the grand mean. The 95% CI for T is then 121.8 ± 1.96 (6.92), which defines a span of 108.2 to 135.4 kg. The entire process, which has been termed the regression-based approach (16), can be summarized as follows (24):

$$95\%\mathrm{CI\ for}\ T = \bar{X} + \mathrm{ICC}(d) \pm 1.96\, SD\sqrt{\mathrm{ICC}(1 - \mathrm{ICC})}. \quad (12)$$

If one had simply used equation 9 using S and *SEM*, the resulting interval would span 120 ± 1.96 (7.8) = 105.1 to 134.9 kg. Note that the differences between CIs is small and that the CI width from equation 9 (29.8 kg) is wider than that from equation 12 (27.2 kg). For all ICCs less than 1.0, the CI width will be narrower from equation 12 than from equation 9 (16), but the differences shrink as the ICC approaches 1.0 and as S approaches $\bar{X}$ (24).

Minimal Differences Needed to Be Considered Real

The *SEM* is an index that can be used to define the difference needed between separate measures on a subject for the difference in the measures to be considered real. For example, if the 1RM of an athlete on one day is 155 kg and at some later time is 160 kg, are you confident that the athlete really increased the 1RM by 5 kg or is this difference within what you might expect to see in repeated testing just due to the noise in the measurement? The *SEM* can be used to determine the minimum difference (MD) to be considered "real" and can be calculated as follows (8, 20, 42):

$$\mathrm{MD} = SEM \times 1.96 \times \sqrt{2}, \quad (13)$$

Once again the point is to construct a 95% CI, and the 1.96 value is simply the z score associated with a 95% CI. (One may choose a different z score instead of 1.96 if a more liberal or more conservative assessment is desired.) But where does the $\sqrt{2}$ come from?

Why can't we simply calculate the 95% CI for a subject's score as we have done above? If the score is outside that interval, then shouldn't we be 95% confident that the subject's score has really changed? Indeed, this approach has been suggested in the literature (25, 37). The key here is that we now have 2 scores from a subject. Each of these scores has a true component and an error component. That is, both scores were measured with error, and simply seeing if the second score falls outside the CI of the first score does not account for the error in the second

score. What we really want here is an index based on the variability of the difference scores. This can be quantified as the *SD* of the difference scores (*SDd*). As it turns out, when there are 2 levels of trials (as in the examples herein), the *SEM* is equal to the *SDd* divided by $\sqrt{2}$ (17, 26):

$$SEM = SDd/\sqrt{2}. \quad (14)$$

Therefore, multiplying the *SEM* by $\sqrt{2}$ solves for the *SDd* and then multiplying the *SDd* by 1.96 allows for the construction of the 95% CI. Once the MD is calculated, then any change in a subject's score, either above or below the previous score, greater than the MD is considered real. More precisely, for all people whose differences on repeated testing are at least greater than or equal to the MD, 95% of them would reflect real differences. Using data set A, the first subject has a trial A1 score of 146 kg. The *SEM* for the test is $\sqrt{57} = 7.6$ kg. From equation 13, MD = $7.6 \times 1.96 \times \sqrt{2} = 21.07$ kg. Thus, a change of at least 21.07 kg needs to occur to be confident, at the 95% level, that a change in 1RM reflects a real change and not a difference that is within what might be reasonably expected given the measurement error of the 1RM test.

However, as with defining a CI for an observed score, the process outlined herein for defining a minimal difference is not precisely accurate. As noted by Charter (13) and Dudek (19), the *SE* of prediction (*SEP*) is the correct *SE* to use in these calculations, not the *SEM*. The *SEP* is calculated as follows:

$$SEP = SD\sqrt{1 - \text{ICC}^2}. \quad (15)$$

To define a 95% CI outside which one could be confident that a retest score reflects a real change in performance, simply calculate the estimated true score (equation 10) plus or minus the *SEP*. To illustrate, consider the same data as in the example in the previous paragraph. From equation 10, we estimate the subject's true score (T) as $T = \bar{X} + \text{ICC}\,(d) = 154 + 0.95\,(146 - 154.5) \cong 146.4$ kg. The $SEP = SD \times \sqrt{(1 - \text{ICC}^2)} = 31.74 \times \sqrt{(1 - 0.95^2)} = 9.91$. The resulting 95% CI is $146.4 \pm 1.96\,(9.91)$, which defines an interval from approximately 127 to 166 kg. Therefore, any retest score outside that interval would be interpreted as reflecting a real change in performance. As given in Table 1, the retest score of 140 kg is inside the CI and would be interpreted as a change consistent with the measurement error of the test and does not reflect a real change in performance. As before, use of a different z score in place of 1.96 will allow for the construction of a more liberal or conservative CI.

OTHER CONSIDERATIONS

In this article, several considerations regarding ICC and *SEM* calculations will not be addressed in detail, but brief mention will be made here. First, assumptions of ANOVA apply to these data. The most common assumption violated is that of homoscedasticity. That is, does the size of the error correlate with the magnitude of the observed scores? If the data exhibit homoscedasticity, the answer is no. For physical performance measures, it is common that the absolute error tends to be larger for subjects who score higher (2, 26), e.g., the noise from repeated strength testing of stronger subjects is likely to be larger than the noise from weaker subjects. If the data exhibit heteroscedasticity, often a logarithmic transformation is appropriate. Second, it is important to realize that ICC and *SEM* values determined from sample data are estimates. As such, it is instructive to construct CIs for these estimates. Details of how to construct these CIs are addressed in other sources (34, 35). Third, how many subjects are required to get adequate stability for the ICC and *SEM* calculations? Unfortunately, there is no consensus in this area. The reader is referred to other studies for further discussion (16, 35, 52). Finally, reliability, as quantified by the ICC, is not synonymous with responsiveness to change (23). The MD calculation presented herein allows one to evaluate a change score after the fact. However, a small MD, in and of itself, is not a priori evidence that a given test is responsive.

PRACTICAL APPLICATIONS

For a comprehensive assessment of reliability, a 3-layered approach is recommended. First, perform a repeated-measures ANOVA and cast the summary table as a 2-way model, i.e., trials and error are separate sources of variance. Evaluate the *F* ratio for the trials effect to examine systematic error. As noted previously, it may be prudent to evaluate the effect for trials using a more liberal α measure than the traditional 0.05 level. If the effect for trials is significant (and the effect size is not trivial), it is prudent to reexamine the measurement schedule for influences of learning and fatigue. If 3 or more levels of trials were included in the analysis, a plateau in performance may be evident, and exclusion of only those levels of trials not in the plateau region in a subsequent reanalysis may be warranted. However, this exclusion of trials needs to be reported. Under these conditions, where systematic error is deemed unimportant, the ICC values will be similar and reflect random error (imprecision). However, it is suggested here that the ICC from equation 3,1 be used (Table 5), since it is most closely tied to the MS_E calculation of the *SEM*. Once the systematic error is determined to be nonsignificant or trivial, interpret the ICC and *SEM* within the analytical goals of your study (2). Specifically, researchers interested in group-level responses can use the ICC to assess correlation attenuation, statistical power, and sample size calculations. Practitioners (e.g., coaches, clinicians) can use the *SEM* (and associated *SE*s) in the interpretation of scores from individual athletes (CIs for true scores, assessing individual change). Finally, although reliability is an important aspect of measurement, a test may exhibit reliability but not be a valid test (i.e., it does not measure what it purports to measure).

REFERENCES

1. ALEXANDER, H.W. The estimation of reliability when several trials are available. *Psychometrika* 12:79–99. 1947.
2. ATKINSON, D.B., AND A.M. NEVILL. Statistical methods for assessing measurement error (reliability) in variables relevant to Sports Medicine. *Sports Med.* 26:217–238. 1998.
3. BARTKO, J.J. The intraclass reliability coefficient as a measure of reliability. *Psychol. Rep.* 19:3–11. 1966.
4. BARTKO, J.J. On various intraclass correlation coefficients. *Psychol. Bull.* 83:762–765. 1976.
5. BAUMGARTNER, T.A. Estimating reliability when all test trials are administered on the same day. *Res. Q.* 40:222–225. 1969.
6. BAUMGARTNER, T.A. Norm-referenced measurement: reliability. In: *Measurement Concepts in Physical Education and Exercise Science.* M.J. Safrit and T.M. Woods, eds. Champaign, IL: Human Kinetics, 1989. pp. 45–72.

7. BAUMGARTNER, T.A. Estimating the stability reliability of a score. *Meas. Phys. Educ. Exerc. Sci.* 4:175–178. 2000.
8. BECKERMAN, H., T.W. VOGELAAR, G.L. LANKHORST, AND A.L.M. VERBEEK. A criterion for stability of the motor function of the lower extremity in stroke patients using the Fugl-Meyer assessment scale. *Scand. J. Rehabil. Med.* 28:3–7. 1996.
9. BEDARD, M., N.J. MARTIN, P. KRUEGER, AND K. BRAZIL. Assessing reproducibility of data obtained with instruments based on continuous measurements. *Exp. Aging Res.* 26:353–365. 2000.
10. BLAND, J.M., AND D.G. ALTMAN. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1:307–310. 1986.
11. BROZEK, J., AND H. ALEXANDER. Components of variance and the consistency of repeated measurements. *Res. Q.* 18:152–166. 1947.
12. BRUTON, A., J.H. CONWAY, AND S.T. HOLGATE. Reliability: What is it and how is it measured. *Physiotherapy* 86:94–99. 2000.
13. CHARTER, R.A. Revisiting the standard error of measurement, estimate, and prediction and their application to test scores. *Percept. Mot. Skills* 82:1139–1144. 1996.
14. CHARTER, R.A. Effect of measurement error on tests of statistical significance. *J. Clin. Exp. Neuropsychol.* 19:458–462. 1997.
15. CHARTER, R.A., AND L.S. FELDT. Meaning of reliability in terms of correct and incorrect clinical decisions: The art of decision making is still alive. *J. Clin. Exp. Neuropsychol.* 23:530–537. 2001.
16. CHARTER, R.A., AND L.S. FELDT. The importance of reliability as it relates to true score CIs. *Meas. Eval. Counseling Dev.* 35: 104–112. 2002.
17. CHINN, S. Repeatability and method comparison. *Thorax* 46: 454–456. 1991.
18. CHINN, S., AND P.G. BURNEY. On measuring repeatability of data from self-administered questionnaires. *Int. J. Epidemiol.* 16:121–127. 1987.
19. DUDEK, F.J. The continuing misinterpretation of the standard error of measurement. *Psychol. Bull.* 86:335–337. 1979.
20. ELIASZIW, M., S.L. YOUNG, M.G. WOODBURY, AND K. FRYDAY-FIELD. Statistical methodology for the concurrent assessment of interrater and intrarater reliability: Using goniometric measurements as an example. *Phys. Ther.* 74:777–788. 1994.
21. FELDT, L.S., AND M.E. MCKEE. Estimation of the reliability of skill tests. *Res. Q.* 29:279–293. 1958.
22. FLEISS, J.L. *The Design and Analysis of Clinical Experiments.* New York: John Wiley and Sons, 1986.
23. GUYATT, G., S. WALTER, AND G. NORMAN. Measuring change over time: assessing the usefulness of evaluative instruments. *J. Chronic Dis.* 40:171–178. 1987.
24. HARVILL, L.M. Standard error of measurement. *Educ. Meas. Issues Pract.* 10:33–41. 1991.
25. HEBERT, R., D.J. SPIEGELHALTER, AND C. BRAYNE. Setting the minimal metrically detectable change on disability rating scales. *Arch. Phys. Med. Rehabil.* 78:1305–1308. 1997.
26. HOPKINS, W.G. Measures of reliability in sports medicine and science. *Sports Med.* 30:375–381. 2000.
27. KEATING, J., AND T. MATYAS. Unreliable inferences from reliable measurements. *Aust. Physiother.* 44:5–10. 1998.
28. KEPPEL, G. *Design and Analysis: A Researcher's Handbook* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall, 1991.
29. KROLL, W. A note on the coefficient of intraclass correlation as an estimate of reliability. *Res. Q.* 33:313–316. 1962.
30. LAHEY, M.A., R.G. DOWNEY, AND F.E. SAAL. Intraclass correlations: there's more than meets the eye. *Psychol. Bull.* 93:586–595. 1983.
31. LIBA, M. A trend test as a preliminary to reliability estimation. *Res. Q.* 38:245–248. 1962.
32. LOONEY, M.A. When is the intraclass correlation coefficient misleading? *Meas. Phys. Educ. Exerc. Sci.* 4:73–78. 2000.
33. LUDBROOK, J. Statistical techniques for comparing measures and methods of measurement: A critical review. *Clin. Exp. Pharmacol. Physiol.* 29:527–536. 2002.
34. MCGRAW, K.O., AND S.P. WONG. Forming inferences about some intraclass correlation coefficients. *Psychol. Methods* 1:30–46. 1996.
35. MORROW, J.R., AND A.W. JACKSON. How "significant" is your reliability? *Res. Q. Exerc. Sport* 64:352–355. 1993.
36. NICHOLS, C.P. Choosing an intraclass correlation coefficient. Available at: www.spss.com/tech/stat/articles/whichicc.htm. Accessed 1998.
37. NITSCHKE, J.E., J.M. MCMEEKEN, H.C. BURRY, AND T.A. MATYAS. When is a change a genuine change? A clinically meaningful interpretation of grip strength measurements in healthy and disabled women. *J. Hand Ther.* 12:25–30. 1999.
38. NUNNALLY, J.C., AND I.H. BERNSTEIN. *Psychometric Theory* (3rd ed.). New York: McGraw-Hill, 1994.
39. OLDS, T. Five errors about error. *J. Sci. Med. Sport* 5:336–340. 2002.
40. PERKINS, D.O., R.J. WYATT, AND J.J. BARTKO. Penny-wise and pound-foolish: The impact of measurement error on sample size requirements in clinical trials. *Biol. Psychiatry.* 47:762–766. 2000.
41. PORTNEY, L.G., AND M.P. WATKINS. *Foundations of Clinical Research* (2nd ed.). Upper Saddle River, NJ: Prentice Hall, 2000.
42. ROEBROECK, M.E., J. HARLAAR, AND G.J. LANKHORST. The application of generalizability theory to reliability assessment: An illustration using isometric force measurements. *Phys. Ther.* 73:386–401. 1993.
43. ROUSSON, V., T. GASSER, AND B. SEIFERT. Assessing intrarater, interrater, and test-retest reliability of continuous measurements. *Stat. Med.* 21:3431–3446. 2002.
44. SAFRIT, M.J.E. *Reliability Theory.* Washington, DC: American Alliance for Health, Physical Education, and Recreation, 1976.
45. SHROUT, P.E. Measurement reliability and agreement in psychiatry. *Stat. Methods Med. Res.* 7:301–317. 1998.
46. SHROUT, P.E., AND J.L. FLEISS. Intraclass correlations: uses in assessing rater reliability. *Psychol. Bull.* 36:420–428. 1979.
47. STRATFORD, P. Reliability: consistency or differentiating between subjects? [Letter]. *Phys. Ther.* 69:299–300. 1989.
48. STRATFORD, P.W., AND C.H. GOLDSMITH. Use of standard error as a reliability index of interest: An applied example using elbow flexor strength data. *Phys. Ther.* 77:745–750. 1997.
49. STREINER, D.L., AND G.R. NORMAN. *Measurement Scales: A Practical Guide to Their Development and Use* (2nd ed.). Oxford: Oxford University Press, 1995. pp. 104–127.
50. THOMAS, J.R., AND J.K. NELSON. *Research Methods in Physical Activity* (2nd ed.). Champaign, IL: Human Kinetics, 1990. pp. 352.
51. TRAUB, R.E., AND G.L. ROWLEY. Understanding reliability. *Educ. Meas. Issues Pract.* 10:37–45. 1991.
52. WALTER, S.D., M. ELIASZIW, AND A. DONNER. Sample size and optimal designs for reliability studies. *Stat. Med.* 17:101–110. 1998.

Acknowledgments

I am indebted to Lee Brown, Joel Cramer, Bryan Heiderscheit, Terry Housh, and Bob Oppliger for their helpful comments on drafts of the paper.

Address correspondence to Dr. Joseph P. Weir, joseph.weir@dmu.edu.