REVIEW ARTICLE

Sports Med 1998 Oct; 26 (4): 217-238
0112-1642/98/0010-0217/$11.00/0

Statistical Methods For Assessing Measurement Error (Reliability) in Variables Relevant to Sports Medicine

Greg Atkinson and *Alan M. Nevill*

Research Institute for Sport and Exercise Sciences, Liverpool John Moores University, Liverpool, England

Contents

Abstract

Minimal measurement error (reliability) during the collection of interval- and ratio-type data is critically important to sports medicine research. The main components of measurement error are systematic bias (e.g. general learning or fatigue effects on the tests) and random error due to biological or mechanical variation. Both error components should be meaningfully quantified for the sports physician to relate the described error to judgements regarding 'analytical goals' (the requirements of the measurement tool for effective practical use) rather than the statistical significance of any reliability indicators.

Methods based on correlation coefficients and regression provide an indication of 'relative reliability'. Since these methods are highly influenced by the range of measured values, researchers should be cautious in: (i) concluding accept-

able relative reliability even if a correlation is above 0.9; (ii) extrapolating the results of a test-retest correlation to a new sample of individuals involved in an experiment; and (iii) comparing test-retest correlations between different reliability studies.

Methods used to describe 'absolute reliability' include the standard error of measurements (SEM), coefficient of variation (CV) and limits of agreement (LOA). These statistics are more appropriate for comparing reliability between different measurement tools in different studies. They can be used in multiple retest studies from ANOVA procedures, help predict the magnitude of a 'real' change in individual athletes and be employed to estimate statistical power for a repeated-measures experiment.

These methods vary considerably in the way they are calculated and their use also assumes the presence (CV) or absence (SEM) of heteroscedasticity. Most methods of calculating SEM and CV represent approximately 68% of the error that is actually present in the repeated measurements for the 'average' individual in the sample. LOA represent the test-retest differences for 95% of a population. The associated Bland-Altman plot shows the measurement error schematically and helps to identify the presence of heteroscedasticity. If there is evidence of heteroscedasticity or non-normality, one should logarithmically transform the data and quote the bias and random error as ratios. This allows simple comparisons of reliability across different measurement tools.

It is recommended that sports clinicians and researchers should cite and interpret a number of statistical methods for assessing reliability. We encourage the inclusion of the LOA method, especially the exploration of heteroscedasticity that is inherent in this analysis. We also stress the importance of relating the results of any reliability statistic to 'analytical goals' in sports medicine.

It is extremely important to ensure that the measurements made as part of research or athlete support work in sports medicine are adequately reliable and valid. The sport medic's dependence on adequate measurements was recently mentioned in reviews on the sports medicine subdisciplines of biomechanics, physiology and psychology research.[1-3] This multidisciplinary nature of sports medicine means that a variety of different types of data are collected by researchers. Nevertheless, the most common measurements in sports medicine are continuous and on an interval or ratio scale. For example, body temperature measured in degrees Celsius or whole body flexibility measured in centimetres above or below the position of the feet when elevated above ground are not theoretically bounded by zero and are therefore considered to be interval data.[4] On the other hand, it is impossible to obtain values of muscle strength or body mass, for example, that are lower than zero. Such variables would be measured on a ratio scale.[5] Both types of data are considered continuous, since the values may not merely be whole numbers, but can be expressed as any number of decimal points depending on the accuracy of the measurement tool.[6]

Mainstream clinical tools may hold sufficient reliability to detect the often large differences in interval or ratio measurements that exist between healthy and diseased patients. Formulae are now available for clinicians to calculate, from continuous measurements, the probability of this concept of 'discordant classification' amongst patients.[7] Nevertheless, laboratory measures of human performance may need to be sensitive enough to distinguish between the smaller differences that exist between elite and subelite athletes (the ability to detect changes in performance, which may be very small, but still meaningful to athletic performance). For sports medicine support work, it is desirable

that a measurement tool be reliable enough to be used on individual athletes. For example, a clinician may need to know whether an improvement in strength following an injury-rehabilitation programme is real or merely due to measurement error. Researchers in sports medicine may need to know the influence of measurement error on statistical power and sample size estimation for experiments. A full discussion of this latter issue is beyond the scope of this review but interested readers are directed towards Bates et al.[8] and Dufek et al.[9] who recently outlined the importance of data reliability on statistical power (the ability to detect real differences between conditions or groups).

The issue of which statistical test to employ for the quantification of 'good' measurement has been recently raised in the newsletter of the British Association of Sport and Exercise Sciences[10] and in an Editorial of the *Journal of Sports Science*,[11] as well as in other sources related to subdisciplines of sport and exercise science.[12-15] Atkinson[10] and Nevill[11] promoted the use of '95% limits of agreement'[16] to supplement any analyses that are performed in measurement studies. This generated much discussion between sports scientists through personal communication with respect to the choice of statistics for assessing the adequacy of measurements. This review is an attempt to communicate these discussions formally.

1. Definition of Terms

Studies concerning measurement issues cover all the sports medicine subdisciplines. The most common topics involve the assessment of the reliability and validity of a particular measurement tool. Validity is, generally, the ability of the measurement tool to reflect what it is designed to measure. This concept is not covered in great detail in the present review (apart from a special type of validity called 'method comparison', which is mentioned in the discussion) mainly because of the different interpretations and methods of assessing validity amongst researchers. Detailed discussions of validity issues can be found in the book edited by Safrit and Wood.[17]

Reliability can be defined as the consistency of measurements, or of an individual's performance, on a test; or 'the absence of measurement error'.[17] Realistically, some amount of error is always present with continuous measurements. Therefore, reliability could be considered as the amount of measurement error that has been deemed acceptable for the effective practical use of a measurement tool. Logically, it is reliability that should be tested for first in a new measurement tool, since it will never be valid if it is not adequately consistent in whatever value it indicates from repeated measurements. Terms that have been used interchangeably with 'reliability', in the literature, are 'repeatability', 'reproducibility', 'consistency', 'agreement', 'concordance' and 'stability'.

Baumgarter[18] identified 2 types of reliability: relative and absolute. Relative reliability is the degree to which individuals maintain their position in a sample with repeated measurements. This type of reliability is usually assessed with some type of correlation coefficient. Absolute reliability is the degree to which repeated measurements vary for individuals. This type of reliability is expressed either in the actual units of measurement or as a proportion of the measured values (dimensionless ratio).

Baumgarter[18] also defined reliability in terms of the source of measurement error. For example, internal consistency reliability is the variability between repeated trials within a day. Researchers should be careful in the interpretation of this type of reliability, since the results might be influenced by systematic bias due to circadian variation in performance.[19] Stability reliability was defined as the day-to-day variability in measurements. This is the most common type of reliability analysis, although it is stressed that exercise performance tests may need more than one day between repeated measurements to allow for bias due to inadequate recovery. Objectivity is the degree to which different observers agree on the measurements and is sometimes referred to as rater reliability.[20] This type of reliability assessment is relevant to meas-

urements that might be administered by different clinicians over time.

These different definitions of reliability have little impact on the present review, since they have all been analysed with similar statistical methods in the sports medicine literature. Nevertheless, a researcher may be interested in examining the relative influence of these different types of reliability within the same study. Generalisability theory (the partitioning of measurement error due to different sources) is appropriate for this type of analysis. This review considers one of the base statistics [the standard error of measurement (SEM)] for measurement error that happens to be used in generalisability theory, but does not cover the actual concept itself. Interested readers are directed to Morrow[21] for a fuller discussion and Roebroeck et al.[22] for an example of the use of the theory in a sports medicine application.

1.1 Systematic Bias and Random Error

Irrespective of the type of reliability that is assessed (internal consistency, stability, objectivity), there are 2 components of variability associated with each assessment of measurement error. These are systematic bias and random error. The sum total of these components of variation is known as total error.[23]

Systematic bias refers to a general trend for measurements to be different in a particular direction (either positive or negative) between repeated tests. There might be a trend for a retest to be higher than a prior test due to a learning effect being present. For example, Coldwells et al.[24] found a bias due to learning effects for the measurement of back strength using a portable dynamometer. Bias may also be due to there being insufficient recovery between tests. In this case, a retest would show a 'worse' score than a prior test. It may be that, after a large number of repeated tests, systematic bias due to training effects (if the test is physically challenging) or transient increases in motivation becomes apparent. For example, Hickey et al.[25] found that the final test of some 16km cycling time trial performances was significantly better than the 3 prior tests measured on a week-to-week basis. Such bias should be investigated if the test is to administered many times as part of an experiment and should be controlled by attempting to maximise motivation on all the tests with individuals who are already well trained.

The other component of variability between repeated tests is the degree of random error. Large amounts of random differences could arise due to inherent biological or mechanical variation, or inconsistencies in the measurement protocol, e.g. not controlling posture in a consistent way during measurements of muscle strength.[24] Whilst such obvious sources of error as protocol variation can be controlled, the random error component is still usually larger than that due to bias. Unfortunately, the researcher can do relatively little to reduce random error once the measurement tool has been purchased, especially if it is due wholly to inherent mechanical (instrument) variation. An important issue here, therefore, is that the researcher could compare magnitudes of random error between different pieces of equipment that measure the same variable so that the 'best' measurement tool is purchased. This denotes that, whatever the choice of a statistic of measurement error, researchers investigating the reliability of a measurement tool should also be consistent in this choice (or provide a number of statistical analyses for global comparison amongst future researchers).

1.2 Heteroscedasticity and Homoscedasticity

One issue that is rarely mentioned in sport and exercise reliability studies is how the measurement error relates to the magnitude of the measured variable. When the amount of random error increases as the measured values increase, the data are said to be heteroscedastic. Heteroscedastic data can also show departures from a normal distribution (i.e. positive skewness).[6] When there is no relation between the error and the size of the measured value, the data are described as homoscedastic. Such characteristics of the data influence how the described error is eventually expressed and analysed.[26,27]

Homoscedastic errors can be expressed in the actual units of measurement but heteroscedastic data should be measured on a ratio scale (although this can be interpreted back into the units of measurement by multiplying and dividing a particular measured value by the error ratio). With homoscedastic errors, providing they are also normally distributed, the raw data can be analysed with conventional parametric analyses, but heteroscedastic data should be transformed logarithmically before analysis or investigated with an analysis based on ranks.

There could be practical research implications of the presence of heteroscedastic errors in measurements. Heteroscedasticity means that the individuals who score the highest values on a particular test also show the greatest amount of measurement error (in the units of measurement). It is also likely that these high-scoring individuals show the smallest changes (in the units of measurement) in response to a certain experimental intervention.[28] Therefore, in line with the discussions on measurement error and statistical power referred to in the introduction, it may be that the detection of small but meaningful changes in sports medicine–related variables measured on a ratio scale is particularly difficult with individuals who score highly on those particular variables.

2. Can a Measurement Tool be Significantly Reliable?

The statistical philosophy for assessing agreement between measurements can be considered to be different from that surrounding the testing of research hypotheses.[29,30] Indeed, the identification of, and adherence to, a single statistical method (or the citation of several different methods in a paper on reliability) could be considered as more important for measurement issues than it is for testing hypotheses. There are several different statistical methods that can help examine a particular hypothesis. For example, in a multifactorial experiment involving comparisons of changes over time between different treatments, one may employ analysis of summary statistics[31] or multifactorial analysis of variance (ANOVA) models[32] to test hypotheses. The consideration of measurement error is a different concept, since one may not necessarily be concerned with hypothesis testing, but the correct, meaningful and consistent quantification of variability between different methods or repeated tests. Coupled with this, the researcher would need to arrive at the final decision as to whether a measurement tool is reliable or not (whether the measurement error is acceptable for practical use).

3. Analytical Goals

The above concept for reliability assessment basically entails the researcher relating measurement error to 'analytical goals' rather than the significance of hypothesis tests. The consideration of analytical goals is routine in laboratory medicine[33,34] but seems to have been neglected in sport and exercise science.

One way of arriving at the acceptance of a certain degree of measurement error (attaining an analytical goal), as already mentioned, is estimating the implications of the measurement error on sample size estimation for experiments or on individuals' differences/changes. The present authors were able to locate only 3 published reliability studies relevant to sports science/medicine which have calculated the influence of the described measurement error on sample size estimation for future research.[35-37] Hopkins[38] provides methods, based on test-retest correlations, in which researchers might perform this extrapolation of measurement error to sample size estimation. Sample size can also be estimated from absolute reliability statistics such as the standard deviation (SD) of test-retest differences.[6,39]

Researchers in sport science have, at least, recognised that an analytical goal might not necessarily be the same as the acceptance of significance on a hypothesis test.[29] In the present review, by considering each reliability statistic in turn, we aim to highlight how an 'acceptable' level of measurement error might still be falsely accepted, when statistical criteria that are still not based on any well-defined analytical goals are employed (e.g.

correlations >0.9, sample mean coefficients of variation <10%). Such criteria are in common use in the sport and exercise sciences.

4. Statistical Methods for Assessing Reliability in Sports Medicine

Many statistical tests have been proposed in the sport science literature for the appraisal of measurement issues. This is illustrated in table I which cites the different methods used in the 'measurement' studies presented at the 1996 conference of the American College of Sports Medicine. It is stressed that some of these studies were 'method comparison' (validity) studies, although the majority investigated reliability issues. It can be seen that the most common methods involve the use of hypothesis tests (paired t-tests, ANOVA) and/or correlation coefficients (Pearson's, intraclass correlation). Other methods cited in the literature involve regression analysis, coefficient of variation (CV) or various methods that calculate 'percentage variation'. A little-quoted method in studies relevant to sport science is the 'limits of agreement' technique outlined by Bland and Altman in 1983[16,41] and refined in later years.[42-44] In the following sections of this review, each statistical method for assessing reliability will be considered using, where possible, real data relevant to sports science and medicine.

Table I. The various statistical methods used in repeatability and validity studies presented at the 43rd meeting of the American College of Sports Medicine[40]a

Type of analysis	Number of studies
Hypothesis test for bias (i.e paired t-test, ANOVA)	16
Pearson's correlation coefficient (r)	17
ICC	3
Hypothesis test and Pearson's correlation coefficient (r)	11
Hypothesis test and ICC	9
CV	4
Absolute error	7
Regression	3
Total	70[b]

a Validity studies as well as reliability investigations were included in this literature search. The critique of the statistical analyses in the present review may not necessarily apply to validity examination.

b 5.6% of the total number of studies presented, 1256.

ANOVA = analysis of variance; **CV** = coefficient of variation; **ICC** = intraclass correlation.

4.1 Paired t-Test for Detection of Systematic Bias

This test would be used to compare the means of a test and retest i.e. it tests whether there is any statistically significant bias between the tests. Although this is useful, it should not of course be employed on its own as an assessment of reliability, since the t-statistic provides no indication of random variation between tests. Altman[30] and Bland and Altman[42] stressed caution in the interpretation of a paired t-test to assess reliability, since the detection of a significant difference is actually dependent on the amount of random variation between tests.

Specifically, because of the nature of the formula employed to calculate the t-value, significant systematic bias will be less likely to be detected if it is accompanied by large amounts of random error between tests. For example, a paired t-test was performed on the data presented in table II to assess repeatability of the 'Fitech' step test for predicting maximal oxygen uptake ($\dot{V}O_{2max}$). The mean systematic bias between week 1 and week 2 of 1.5 ml/kg/min was not statistically significant (t_{29} = 1.22, p = 0.234), a finding that has been used on its own by some researchers (table I) to conclude that a tool has acceptable measurement error. However, if one examines the data from individual participants, it can be seen that there are differences between the 2 weeks of up to ±16 ml/kg/min (participant 23 recorded 61 ml/kg/min in the first test but only 45 ml/kg/min in the retest).

The possible compromising effect of large amounts of random error on the results of the paired t-test is further illustrated by applying it to the hypothetical data in table III. With these data, a test-retest t-value of zero would be obtained (p = 0.99), which could be interpreted as excellent reliability, even though there are very large random differences in the individual cases. With the use of a t-test

Table II. Test-retest data for the Fitech step test predicting maximal oxygen consumption.[a] The data have been ranked to show that a high correlation may not necessarily mean that individuals maintain their positions in a sample following repeated measurements (adequate relative reliability).

Individual	Test 1 (ml/kg/min)	Test 1 (ranks)	Test 2 (ml/kg/min)	Test 2 (ranks)	Difference (ml/kg/min)	Absolute difference in ranks
1	31	2.0	27	1.0	–4	1.0
2	33	3.0	35	3.0	+2	0
3	42	9.0	47	13.5	+5	4.5
4	40	6.0	44	8.0	+4	2
5	63	28.0	63	28.0	0	0
6	28	1.0	31	2.0	+3	2
7	43	12.5	54	23.5	+11	11
8	44	15.0	54	23.5	+10	8.5
9	68	29.0	68	30.0	0	1
10	47	18.0	58	25.5	+11	7.5
11	47	18.0	48	16.0	+1	2
12	40	6.0	43	5.5	+3	0.5
13	43	12.5	45	11.0	+2	1.5
14	47	18.0	52	20.0	+5	2
15	58	24.5	48	16.0	+10	8.5
16	61	26.5	61	27.0	0	0.5
17	45	16.0	52	20.0	+7	4
18	43	12.5	44	8.0	+1	4.5
19	58	24.5	48	16.0	–10	8.5
20	40	6.0	44	8.0	+4	2
21	48	20.5	47	13.5	–1	7
22	42	9.0	52	20.0	+10	11
23	61	26.5	45	11.0	–16	15.5
24	48	20.5	43	5.5	–5	15
25	43	12.5	52	20.0	+11	7.5
26	50	22.0	52	20.0	+2	2
27	39	4.0	40	4.0	+1	0
28	52	23.0	58	25.5	+6	2.5
29	42	9.0	45	11.0	+3	2
30	77	30.0	67	29.0	–10	1
Mean (SD)						
	47.4 (10.9)		48.9 (9.4)		+1.5 (6.6)	

a Data obtained in a laboratory practical at Liverpool John Moores University. t = 1.22 (p = 0.234); r = 0.80 (p < 0.001); ICC = 0.88; r_c = 0.78; sample CV = 7.6%; limits of agreement = –1.5 ± 12.9 ml/kg/min (0.97 ×/÷ 0.29 as a ratio).

CV = coefficient of variation; **ICC** = intraclass correlation; **r** = Pearson's product-moment correlation; $\mathbf{r_c}$ = concordance correlation; **SD** = standard deviation; **t** = test statistic from t-test.

per se, very unreliable (relatively high random error) measurements would be concluded as very reliable (relatively small bias)! It should be noted that the correlation between test and retest may not, in all data sets, be a good indicator of the amount of absolute random error present, which is the basis of the denominator in the paired t-test equation (see the discussion of correlation methods below).

The use of a t-test may still be recommended in a measurement study that investigates a simple test and retest, since it will detect large systematic bias (relative to the random error), and the terms in the formula for the t-value can be used in the calculation of measures of random error (e.g. limits of agreement). Nevertheless, the researcher may need to supplement this analysis with the consideration

Table III. Hypothetical data from a validity study comparing a test and a retest of spinal flexibility. The sole use of a t-test on these data would provide a t-value = 0 (p = 0.99), which may lead some researchers to conclude good reliability when large random variation is evident

Test 1 (degrees)	Test 2 (degrees)	Difference (degrees)
1	10	+9
10	1	–9
2	20	+18
20	2	–18
3	30	+27
30	3	–27
4	40	+36
40	4	–36
Mean (SD)		
13.8 (14.7)	13.8 (14.7)	0 (26.3)

SD = standard deviation.

of an analytical goal. For example, the bias of 1.5 ml/kg/min for the data in table I represents about 3% of the grand mean $\dot{V}O_{2max}$ of the sample. This seems small in relation to the amount of random error in these data (see section 8). Besides, a good experiment would be designed to control for any such bias (i.e. control groups/conditions). Nevertheless, it could be that the bias (probably due to familiarisation in this case) is reduced if more retests are conducted and examined for reliability. This denotes the use of ANOVA procedures.

4.2 Analysis of Variation for Detection of Systematic Bias

ANOVA with repeated measures (preferably with a correction for 'sphericity')[32] has been used for comparing more than one retest with a test.[32,45] With appropriate *a priori* or *post hoc* multiple comparisons (e.g. Tukey tests), it can be used to assess systematic bias between tests. However, the sole use of ANOVA is associated with exactly the same drawback as the paired t-test in that the detection of systematic bias is affected by large random (residual) variation. Again it should be noted that a correlation coefficient (intraclass in the case of ANOVA) may not be as sensitive an indicator of this random error as an examination of the residual mean squared error itself in the ANOVA results table (the calculation of an F-value for differences between tests in a repeated measures ANOVA involves variance due to tests and residual error. The variance due to individuals is involved in the calculation of a intraclass correlation but is 'partitioned out' of a repeated measures ANOVA hypothesis test; see section 8).

As with the t-test, ANOVA is useful for detecting large systematic errors and the mean squared error term from ANOVA can be used in the calculation of indicators of absolute reliability.[39,46] An important point in the use of a hypothesis test to assess agreement, whether it be either a paired t-test or ANOVA, is that if significant (or large enough to be important) systematic bias is detected, a researcher would need to adapt the measurement protocol to remove the learning or fatigue effect on the test (e.g. include more familiarisation trials or increase the time between repeated measurements, respectively). It is preferable that the method should then be reassessed for reliability.[18] An intuitive researcher may suspect that a test would show some bias because of familiarisation. It follows, therefore, that a reliability study may be best planned to have multiple retests. The researcher would then not need to go 'back to the drawing board' but merely examine when the bias between tests is considered negligible. The number of tests performed before this decision is made would be suggested as familiarisation sessions to a future researcher. This concept is discussed in greater detail by Baumgarter.[18]

4.3 Pearson's Correlation Coefficient

The Pearson's correlation coefficient has been the most common technique for assessing reliability. The idea is that if a high (>0.8) and statistically significant correlation coefficient is obtained, the equipment is deemed to be sufficiently reliable.[47] Baumgarter[18] pointed out that correlation methods actually indicate the degree of relative reliability. This is definitely conceptually useful, since a researcher could, in theory, tell how consistently the measurement tool distinguishes between individuals in a particular population. However, Bland

and Altman[42] and Sale[48] considered the use of the correlation coefficient as being inappropriate, since, among other criticisms, it cannot, on its own, assess systematic bias and it depends greatly on the range of values in the sample.[49] The latter note of caution in the use of test-retest correlation coefficients is the most important. For example, we have already seen that there is substantial random variation among the individual data in table II, but if correlation was used to examine this, it would be concluded that the test has good repeatability (test-retest correlation of $r = 0.80$, $p < 0.001$). Note that the sample in table II is very varied in maximal oxygen consumption (28 to 77 ml/kg/min).

In table IV, the same data from table II have been manipulated to decrease the interindividual variation while retaining exactly the same level of absolute reliability [indicated by the differences column and the standard deviation (SD) of these differences]. When Pearson's r is calculated for these data, it drops to a nonsignificant 0.27 ($p > 0.05$). This phenomenon suggests that researchers should be extremely cautious in the 2 common procedures of: (i) extrapolating test-retest correlations, which have been deemed acceptable to a new and possibly more homogeneous sample of individuals (e.g. elite athletes); and (ii) comparing test-retest r-values between different reliability studies (e.g. Perrin[50]). To overcome these difficulties, there are methods for correcting the correlation coefficient for interindividual variability.[51] Conceptually, this correction procedure would be similar to the use of an indicator of absolute reliability; these statistics are relatively unaffected by population heterogeneity (see section 8).

5. Correlation and Relative Reliability

Despite the above notes of caution when comparing correlation results, it could be argued that a high correlation coefficient reflects adequate relative reliability for use of the measurement tool in the particular population that has been investigated. This seems sensible, since the more homogeneous a population is, the less the measurement error would need to be in order to detect differences

Table IV. The same data as in table II but manipulated to give a less heterogeneous sample (indicated by the test and retest sample standard deviations (SDs) being approximately half those in table II). The data has exactly the same degree of agreement (indicated by column of differences) between the test and retest as the data presented in table II[a]

Test 1 (ml/kg/min)	Test 2 (ml/kg/min)	Difference (ml/kg/min)
41	37	–4
43	45	2
42	47	5
40	44	4
43	43	0
48	51	3
43	54	11
44	54	10
48	48	0
47	58	11
47	48	1
40	43	3
43	45	2
47	52	5
58	48	–10
41	41	0
45	52	7
43	44	1
58	48	–10
40	44	4
48	47	–1
42	52	10
61	45	–16
48	43	–5
43	52	9
50	52	2
39	40	1
52	58	6
42	45	3
57	47	–10
Mean (SD)		
46.1 (5.9)	47.6 (5.1)	1.5 (6.6)

a $t = 1.22$ ($p = 0.234$), $r = 0.27$ ($p > 0.05$), ICC = 0.43, $r_c = 0.28$, sample CV = 7.6%, limits of agreement = $–1.5 \pm 12.9$ ml/kg/min (0.97 ×/÷ 1.29 as a ratio). Note that the results of the correlation methods are very different from those calculated on the data from table II.

CV = coefficient of variation; **ICC** = intraclass correlation; **r** = Pearson's product-moment correlation; $\mathbf{r_c}$ = concordance correlation; **t** = test statistic from t-test.

between individuals within that population. Using our examples, the correlation coefficients suggest that relative reliability is worse for the data in table IV than those in table II, since the former data is more homogeneous and, therefore, it is more difficult to detect differences between individuals for that given degree of absolute measurement error.

The use of correlation to assess this population-specific relative reliability is quite informative but, unfortunately, the ability of a high correlation coefficient to reflect an adequate consistency of group positions in any one sample is also questionable with certain data sets. For example, a researcher may have the 'analytical goal' that the $\dot{V}O_{2max}$ test (table II) can be used as a performance test to consistently rank athletes in a group. The researcher may follow convention and deem that this analytical goal has been accomplished, since a highly significant test-retest correlation of 0.80 ($p < 0.001$) was obtained (in fact, it would be extremely difficult not to obtain a significant correlation in a reliability study with the sort of sample that is commonly used in studies on measurement issues: males and females, individuals of varying age with a wide range of performance abilities).

If one now examines, in table II, the actual rankings of the sample based on the 2 tests using the measurement tool, it can be seen that only 3 individuals maintained their positions in the group following the retest. Although the maintenance of the exact same rank of individuals in a sample may be a rather strict analytical goal for a measurement tool in sports medicine (although this has not been investigated), it should be noted that 4 individuals in this highly correlated data-set actually moved more than 10 positions following the retest compared with the original test. In this respect, a correlation coefficient based on ranks (e.g. Spearman's) may be more informative for the quantification and judgement of 'relative reliability'. This would have the added benefit of making no assumptions on the shape of the data distribution and being less affected by outliers in the data.[52]

A rank correlation or a correlation on the logged test-retest data is rarely used in reliability studies. This is surprising given the high likelihood that heteroscedasticity is present in data recorded on the ratio scale.[53] The presence of such a characteristic in the described error would mean that a conventional correlation analysis on the raw data is not really appropriate.[26] Taking this point further, a reliability study employing both conventional correlation analysis on the raw, untransformed data (which assumes no evidence of heteroscedasticity) and the CV statistic (which does assume heteroscedasticity is present) can be criticised somewhat for mixing statistical 'apples and oranges'.

5.1 Implications of Poor Interpretation of Test-Retest Correlations

The above disparity between the results of correlation analysis and the perceived reliability may mean that there may be measurement tools in sports medicine that have been concluded as reliable on the basis of the correlation coefficient, but they will not, in practical use, realise certain analytical goals. For example, the majority of tools and protocols for the measurement of isokinetic muscle strength have been tested for reliability with correlation methods applied to heterogeneous data. Most of these correlations are above 0.8.[50] Only recently, with the emergence of more appropriate analysis techniques, is it emerging that the repeatability of these measurements is relatively poor at faster isokinetic speeds.[54] Nevill and Atkinson[53] examined the reliability of 23 common measurement tools in sport and exercise science research. The use of an absolute measure of reliability (ratio limits of agreement) showed that there were considerable differences in reliability between measurement tools.

There are several other pieces of evidence which support the lack of sensitivity of correlation for assessing even relative reliability; Bailey et al.[55] and Sarmandal et al.[56] assessed the reliability of several clinical measures. Test-retest correlations ranged from 0.89 to 0.98, but when a measure of absolute reliability (limits of agreement) was related to the interindividual variation, the usefulness of the measurement tools was questionable. Atkin-

son et al.[57] examined if the measurement error of several performance tests was influenced by the time of day that the measurements were obtained. Test-retest correlations were consistently very high at all times of day. Only when an absolute indicator of reliability was examined did it become apparent that random measurement error seemed to be higher when data was collected at night. Ottenbacher and Tomchek[15] also showed that the correlation coefficient is not sensitive enough to detect inadequate method comparison based on inter-individual differences in a sample. It was found in a data simulation study that a between-method correlation only dropped from 0.99 to 0.98, even though absolute reliability was altered to a degree whereby it would affect the drawing of conclusions from the measurements. The statistical implications of this study would apply equally to the assessment of measurement error and relative reliability.

It is clear that the concept of 'relative reliability' is useful and correlation analysis does provide some indication of this. Interestingly, in clinical chemistry, a statistical criterion for 'relative reliability' is not a high correlation coefficient, but the related measure of absolute reliability expressed as a certain proportion of the interindividual variance.[33,34] Bailey et al.[55] and Sarmandal et al.[56] adopted a similar approach when they related the limits of agreement between 2 observers to population percentile (or qualitative categories) charts. Taking this stance, the ultimate analytical goal for relative reliability would be that the measurement error is less than the difference between individual differences or analytical goal-related population centiles. It is recommended that statisticians working in sport and exercise sciences tackle the problem of defining an acceptable degree of relative reliability for practical use of a measurement tool together with an investigation of the statistic that is most sensitive for the assessment of relative reliability. We suggest the employment of analytical simulations applied to reliability data sets[15] in order to realise these aims.

It is possible to relate test-retest correlations to analytical goals regarding adequate sample sizes for experiments.[38,58] Interestingly, for the estimation of sample sizes in repeated-measures experiments, the correlation would be converted, mathematically, to an absolute reliability statistic. Bland[39] showed how the SD of the differences or residual error (measures of absolute reliability) could be obtained from a test-retest correlation coefficient to estimate sample size. It is residual error, not the correlation coefficient, that is the denominator in 'repeated measures' hypothesis tests and is therefore used in this type of statistical power estimation.

6. Intraclass Correlation

Intraclass correlation (ICC) methods have become a popular choice of statistics in reliability studies, not least because they are the advised methods in the 2 textbooks on research methodology in sports science.[32,45] The most common methods of ICC are based on the terms used in the calculation of the F-value from repeated measures ANOVA.[18] The main advantages of this statistic over Pearson's correlation are maintained to be that the ICC is univariate rather than bivariate and it can be used when more than one retest is being compared with a test.[18] The ICC can also be calculated in such a way that it is sensitive to the presence of systematic bias in the data (there is an argument, discussed in section 7, against the sole citation of such an indicator of 'total error' which combines both bias and random variation into a single coefficient). In fact, there are at least 6 ways of calculating an ICC, all giving different results.[15,59] Eliasziw et al.[60] discussed the choice of an appropriate ICC. The most important implication of this, as Krebs[61] stressed, is that researchers must detail exactly how this choice is made and how an ICC has been calculated in a reliability study.

Whatever the type of ICC that is calculated, it is suggested that, like Pearson's r, an ICC close to 1 indicates 'excellent' reliability. Various categories of agreement based on the ICC, ranging from 'questionable' (0.7 to 0.8) to 'high' (>0.9), are

provided by Vincent.[32] The present authors were unable to locate any reference in the sport and exercise science literature relating these ICC 'cut-off' points to any analytical goals for research. A more informative approach would be to calculate confidence intervals for a given ICC as detailed by Morrow and Jackson.[29]

The calculated ICC[45] of 0.88 for the data in table II would suggest 'good' reliability of the measurements. This is especially true when it has already been seen that there are quite large test-retest differences in some individuals and the relative reliability, by examining the stability of the sample ranks, might not be sufficient for some analytical goals. When the ICC is calculated on the less heterogeneous data in table IV (same degree of agreement as in data from table II), it drops to a very poor 0.43. Therefore, it is apparent that the ICC is prone to exactly the same constraints as Pearson's r, in that it includes the variance term for individuals and is therefore affected by sample heterogeneity to such a degree that a high correlation may still mean unacceptable measurement error for some analytical goals.[62,63]

Myrer et al.[64] highlighted with a practical example the difficulties in interpreting ICCs. Ottenbacher and Tomcheck[15] showed in data simulations that an ICC never dropped below 0.94. This occurred despite marked changes in the absolute agreement between 2 methods of measurement and whilst the sampling characteristics were controlled. Quan and Shih[65] maintained that the ICC should really only be employed when a fixed population of individuals can be well defined. We support the citation of the ICC in any reliability study but believe it should not be employed as the sole statistic and more work is needed to define acceptable ICCs based on the realisation of definite analytical goals.

7. Other Methods Based on Correlation

In an effort to rectify a perceived problem with Pearson's correlation (that it is not sensitive to disagreement between methods/tests due to systematic bias), Lin[66] introduced the 'concordance correlation coefficient' (r_c), which is the correlation between the 2 readings that fall on the 45 degree line through the origin (the line of identity on a scatterplot). Nickerson[67] maintained that this statistic is exactly the same as one type of ICC that is already used by researchers. First, this method is again sensitive to sample heterogeneity.[68] The r_c for the heterogeneous data in table II is 0.78 compared with 0.28 for the less heterogeneous (but same level of agreement) data in table IV. Second, although it may seem convenient to have a single measure of agreement (one that is sensitive to both bias and random error), it may be inconvenient in practical terms when this 'total error' is cited on its own, so the reader of the reliability study is left wondering whether the measurement protocol needs adapting to correct for bias or is associated with high amounts of random variation.[68] This possibility of 'over-generalising' the error, which may constrain the practical solutions to this error, also applies to both the type of ICC which includes the between trials mean-squared-error term as well as the mean squared residual term in its calculation[32,45] and the limits of agreement method if bias and random error are not cited separately (see section 8.3).

7.1 Regression Analysis

This is another common method of analysis in agreement studies but, like hypothesis tests and correlation methods, it may be misleading in some reliability assessments.[42,66] Conceptually, one is not dealing with a predictor and a response variable, which is the philosophy behind regression. In addition, sample heterogeneity is, again, a possible problem for extrapolation of the reliability analysis; the R^2 and regression analysis for the data in table II are 0.64 and $F = 49.01$ ($p < 0.0001$), respectively, thus, indicating 'good' reliability. For the more homogeneous but equally agreeable data (in terms of absolute reliability) in table IV, the R^2 and regression analysis are 0.08 and $F = 2.54$ ($p > 0.10$), respectively, indicating very poor reliability.

For systematic bias, the null hypothesis that the intercept of the regression line equals zero would

be tested. As with the t-test, a wide scatter of individual differences may lead to a false acceptance of this hypothesis (the conclusion that bias is not significant, even though it may be large enough to be important).

8. Statistical Measures of Absolute Reliability

The most common methods of analysing absolute reliability are the SEM and the CV. A little-used statistic in sport and exercise sciences, which could be considered to measure absolute reliability, is the limits of agreement method. One aspect that these statistics have in common is that they are unaffected by the range of measurements. Therefore, they all theoretically provide an indication of the variability in repeated tests for specific individuals, irrespective of where the individuals rank in a particular sample. The general advantage of these statistics over indicators of relative reliability is that it is easier, both to extrapolate the results of absolute reliability studies to new individuals and to compare reliability between different measurement tools. As discussed in sections 8.1 to 8.3, these 3 statistics do seem to differ in the way absolute reliability is expressed. They also make different assumptions regarding the presence of heteroscedasticity (a positive relationship between the degree of measurement error and the magnitude of the measured value).

8.1 Standard Error of Measurement

One indicator of absolute reliability is the 'standard error of measurement'.[45,60,69] The most common way of calculating this statistic that is cited in the sports science literature is by means of the following equation:[18,45]

$$SEM = SD\sqrt{1 - ICC}$$

where SEM = 'standard error of measurement', SD = the sample standard deviation and ICC = the calculated intraclass correlation coefficient. The use of SD in the equation, in effect, partially 'cancels out' the interindividual variation that was used to in the calculation of the ICC. Nevertheless, the statistic (calculated this way) is still affected by sample heterogeneity (3.5 ml/kg/min for the data in table II versus 2.8 ml/kg/min for the data with the same SD of differences in table IV).

Stratford and Goldsmith[69] and Eliasziw et al.[60] stated that SEM can be calculated from the square root of the mean square error term in a repeated measures ANOVA. This statistic would be totally unaffected by the range of measured values. To add to the confusion over the method of calculation, Bland and Altman[43] called this statistic 'the intra-individual SD'. In addition to the differences in the terminology, this latter calculation also seems to give a slightly different result (4.7 ml/kg/min for the data in table II and table IV) from that obtained with the above equation for SEM based on the ICC. The cause of this seems to lie in the type of ICC that is employed (random error or random error + bias). For the above calculations, we employed the ICC without the bias error according to the methods of Thomas and Nelson.[45]

The statistic is expressed in the actual units of measurement, which is useful since the smaller the SEM the more reliable the measurements. The SEM is also used as a 'summary statistic' in generalisability theory to investigate different sources of variation in test scores.[22] Useful methods have also been formulated to compare SEMs between measurement tools.[69]

The question of 'how does one know if a particular SEM statistic indicates adequate reliability?' seems to be unanswered in the literature. Baumgarter[18] showed how an SEM could be used to ascertain whether the difference in measurements between 2 individuals is real or due to measurement error. It was stated that 'confidence bands' based on the SEM are formed around the individual scores. If these bands do not overlap, it was maintained that the difference between the measurements is real. However, researchers should be extremely cautious in following this advice, since the SEM covers about 68% of the variability and not, as Thomas and Nelson[45] discussed, 95%, which is the conventional criterion used in confidence interval comparisons. Eliasziw et al.[60] also discussed

the use of SEM to differentiate between real changes and those due to measurement error and suggested $1.96\sqrt{2} \times \text{SEM}$ which, interestingly, approximates to the limits of agreement statistic (see section 8.3).

Besides the lack of clarity over an acceptable SEM, the use of this statistic is associated with several assumptions. First, it is assumed that there is a 'population' of measurements for each individual (the SEM actually approximates to the mean SD for repeated measurements in individuals), and that this population is normally distributed and that there are no carry-over effects between the repeated tests. Payne[70] discussed these assumptions in more detail. The use of the SEM also denotes that heteroscedasticity is not present in the data, so that it is appropriate only if the data are purely interval in nature. Therefore, if for example an SEM of 3.5 ml/kg/min is calculated, it is assumed that this amount of absolute error is the same for individuals recording high values in the sample as those scoring low values. Nevill and Atkinson[53] have shown that this homoscedasticity is uncommon in ratio variables relevant to sports medicine. Practically, for researchers who are examining a subsample of individuals who score highly on certain tests, the use of SEM may mislead them into thinking that the measurement error is only a small percentage of these scores (the measurement error has been underestimated relative to the particular sample that is examined). This denotes that, if heteroscedasticity is present in data, the use of a ratio statistic (e.g. CV) may be more useful to the researchers.

8.2 Coefficient of Variation

The CV is common in biochemistry studies where it is cited as a measure of the reliability of a particular assay.[71] It is somewhat easier to perform multiple repeated tests in this field than it is in studies on human performance. There are various methods of calculating CV, but the simplest way is with data from repeated measurements on a single case, where the SD of the data is divided by the mean and multiplied by 100.[48] An extension of this on a sample of individuals is to calculate the mean CV from individual CVs. The use of a dimensionless statistic like the CV has great appeal, since the reliability of different measurement tools can be compared.[72] However, as discussed in detail by Allison[73] and Yao and Sayre,[74] there may be certain limitations in the use of CV.

Researchers should be aware that the assumption of normality for an assumed 'population' of repeated tests applies to CV in the same way as with the SEM. Detwiler et al.[75] discussed the difficulty of examining these assumptions for CV with a small number of repeated measures. Unlike SEM, CV methods apply to data in which the degree of agreement between tests does depend on the magnitude of the measured values. In other words, the use of CV assumes that the largest test-retest variation occurs in the individuals scoring the highest values on the test.[42] Although this characteristic is probably very common with sports science data on a ratio scale (see section 8.3),[53] it is best if heteroscedasticity is actually explored and quantified before assuming it is present. This exploration is not very common amongst sport science researchers carrying out reliability studies. Besides, there are reliability data sets which definitely should not be described by CV. For example, a CV would be meaningless for data that can show negative values (not bounded by zero), since the use of the CV denotes that the measurement error approximates zero for measured values that are close to zero. This would not be so if zero values were midway on a measurement scale (e.g. whole body flexibility measures).

Another cautionary note on the use of CV centres around its practical meaning to researchers performing experiments. Some scientists seem to have chosen, quite arbitrarily, an analytical goal of the CV being 10% or below.[76] This does not mean that all variability between tests is always less than 10% of the mean. A CV of 10% obtained on an individual actually means that, assuming the data are normally distributed, 68% of the differences between tests lie within 10% of the mean of the data.[71] Therefore, as with the SEM statistic, the variability

is not described for 32% of the individual differences. For example, if a test-retest CV of 10% was obtained with a test of maximal oxygen consumption and the grand sample mean of the 2 tests was 50 ml/kg/min, the CV of 10% might be considered an indicator of acceptable agreement. Realistically, there could be test-retest differences of greater than 10 ml/kg/min (20% of mean) in some individuals.

The criticism of CV that it is very rarely applied to an analytical goal applies in particular to the common situation in which means are calculated from a sample of individual CVs. The true variation between tests may be underestimated for some new individuals in this case. For example, the sample mean CV for the data in table II is 7.6%, which could be used to indicate very good reliability. This is unrealistic given that over a third of the sample shows individual differences that can be calculated to be greater than 13% of the respective means.

Sarmandal et al.[56] and Bailey et al.[55] also showed with practical examples how mean CVs of 1.6 to 4% did not reflect adequate reliability for some clinical measurements. It is probably more informative if the sample SD of the repeated tests is multiplied by 1.96 before being expressed as a CV for each individual,[77] as this would cover 95% of the repeated measurements. It is stressed, however, that if a sample mean CV is then calculated, this may still not reflect the repeated test error for all individuals, but only the 'average individual' (50% of the individuals in the sample). For this reason, Quan and Shih[65] termed this statistic the 'naive estimator' of CV and suggested that it should not be used. These researchers and others[38,44] described more appropriate CV calculations based on the mean square error term (from ANOVA) of logarithmically transformed data. This is an important part of the last statistical method that is to be discussed; the limits of agreement technique.

8.3 Bland and Altman's 95% Limits of Agreement

Altman and Bland[41] recognised several of the above limitations with these different forms of analysis and introduced the method of 'limits of agreement', an indicator of absolute reliability like SEM and CV. The main difference between these statistics seems to be that the limits of agreement assume a population of individual test-retest differences. Chatburn[23] termed this type of statistic an error interval. SEM and CV, as discussed above, involve an assumed population of repeated measurements around a 'true value' for each individual. Chatburn[23] called this concept a tolerance interval. Although there are differences here on the statistical philosophy, the present review is more concerned with the practical use of these statistics.

The first step in the limits of agreement analysis is to present and explore the test-retest data with a Bland-Altman plot, which is the individual subject differences between the tests plotted against the respective individual means (it is a mistake to plot the differences against the scores obtained for just one of the tests).[78] An example of a Bland-Altman plot using the data in table II is provided in figure 1. Using this plot rather than the conventional test-retest scattergram, a rough indication of systematic bias and random error is provided by examining the direction and magnitude of the scatter around the zero line, respectively. It is also important to observe whether there is any heteroscedasticity in the data (whether the differences depend on the

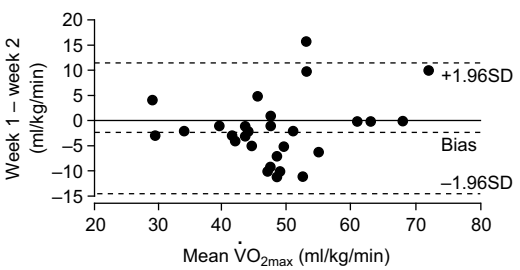

Fig. 1. A Bland-Altman plot for the data presented in table II. The differences between the tests/methods are plotted against each individual's mean for the 2 tests. The bias line and random error lines forming the 95% limits of agreement are also presented on the plot. Visual inspection of the data suggests that the differences are greater with the highest maximal oxygen uptake ($\dot{V}O_{2max}$) values. A similar plot can be formed from the results of analysis of variance (ANOVA) by plotting the residuals against the actual scores. **SD** = standard deviation.

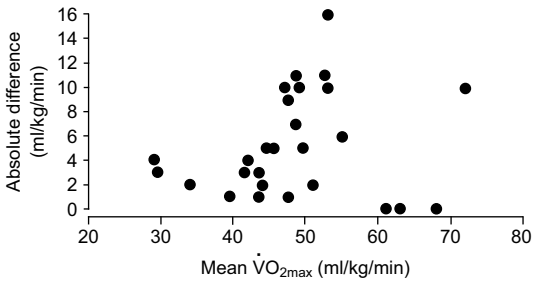

Fig. 2. A plot of the absolute differences between the tests/methods and the individual means for the examination of heteroscedasticity in the data presented in table II (r = 0.18, p = 0.345). This correlation is decreased to 0.01 when the data are logarithmically transformed. Therefore, there is evidence that the limits of agreement would be best expressed as ratios (the absolute measurement error is greater for the individuals who score highly on the test). **SD** = standard deviation.

magnitude of the mean). Heteroscedasticity can be examined formally by plotting the absolute differences against the individual means (fig. 2) and calculating the correlation coefficient (correlation is appropriate here, since the alternative hypothesis is that there is a relationship present). If heteroscedasticity is suspected, the analysis is more complicated (see below).

If the heteroscedasticity correlation is close to zero and the differences are normally distributed, one may proceed to calculate the limits of agreement as follows. First, the SD of the differences between test 1 and test 2 is calculated. The SD of the differences of the data in table II is 6.6 ml/kg/min. This is then multiplied by 1.96 to obtain the 95% random error component of 12.9 ml/kg/min (the 95th percentile is the way reliability data should be presented according to the British Standards Institute).[79] If there is no significant systematic bias (identified by a paired t-test) then there is a rationale for expressing the limits of agreement as ± this value. However, one of the discussed drawbacks of the t-test was that significant bias would not be detected if it is accompanied by large random variation. One could quote the random error with the bias to form the limits of agreement, even if it is not statistically significant. For the data in table II, since there is a slight bias of –1.5 ml/kg/min, the limits of agreement are –14.4 to +11.4 ml/kg/min. Expressed this way, the limits of agreement are actually a measure of 'total error' (bias and random error together). It is probably more informative to researchers reading abstracts of reliability studies if the bias and random error components are cited separately, e.g. –1.5 ± 12.9 ml/kg/min.

It was stated earlier that CV methods should be used only if the variability depends on the magnitude of the mean values (heteroscedasticity). If, from the positive correlation between the absolute differences and the individual means, there is heteroscedasticity in the data, then Bland and Altman[16] recommend the logarithmic (natural) transformation of the data before the calculation of limits of agreement. The final step would be to antilog the data. Bland and Altman[16] provide a worked example for this.

In the examination of heteroscedasticity, Nevill and Atkinson[53] found that, if the correlation between absolute differences and individual means is positive but not necessarily significant in a set of data, it is usually beneficial to take logarithmic values when calculating the limits of agreement. For example, there is very slight heteroscedasticity present in the data from table II (fig. 2, r = 0.18, p = 0.345). If logs are taken, this correlation is reduced to 0.01. Having taken logs of the measurements from both weeks, the mean ± 95% limits of agreement is calculated to be –0.0356 ± 0.257. Taking antilogs of these values the mean bias on the ratio scale is 0.97 and the random error component is now ×/÷ 1.29. Therefore, 95% of the ratios should lie between 0.97 ×/÷ 1.29. If the sample of differences is not normally distributed, which has also been observed with some measurements relevant to sports medicine,[53] the data would, again, benefit from logarithmic transformation. The data in table II are actually not normally distributed (Anderson-Darling test), but after log transformation they follow normality. It follows that the previous tests for bias (paired t-test, ANOVA) should have, strictly speaking been performed on the log transformed data. Nevertheless, this does not de-

tract, in any way, from the points that were made regarding the use of these tests to detect bias in reliability studies.

9. Limits of Agreement and Analytical Goals

The next step is the interpretation of the limits of agreement. Some researchers[80] have concluded acceptable measurement error by observing that only a few of the test-retest differences fall outside the 95% limits of agreement that were calculated from those same differences. This is not how the limits should be interpreted. Rather, it can be said that for a new individual from the studied population, it would be expected (an approximate 95% probability) that the difference between any 2 tests should lie within the limits of agreement. Therefore, in the case of the Fitech test, we expect the differences between the test and retest of an individual from the particular population to lie between –14.4 and +11.5 ml/kg/min. Since there was evidence that heteroscedasticity was present in the Fitech data (the heteroscedasticity correlation reduced following logarithmic transformation of the data), the limits are best represented by ratios.

From the ratio limits of agreement calculated above (0.97 ×/÷ 1.29), it can be said that for any individual from the population, assuming the bias that is present (3%) is negligible, any 2 tests will differ due to measurement error by no more than 29% either in a positive or negative direction (the error is actually slightly greater in the positive than the negative direction with true ratio data that are heteroscedastic). It should be noted, as Bland[39] observed, that this value is very similar to the value of 27% calculated in an arguably simpler manner from $100 \times (1.96 \times \text{SD diff/grand mean})$ on the data prior to logging, where 'SD diff' represents standard deviation of the differences between test and retest and 'grand mean' represents (mean of test 1 + mean of test 2)/2.

As discussed earlier, it is the task of the researcher to judge, using analytical goals, whether the limits of agreement are narrow enough for the test to be of practical use. The comparison of reliability between different measurement tools using limits of agreement is, at present difficult, since there have been so few studies employing limits of agreement for sports science measurements. With respect to the Fitech test data, the limits of agreement for reliability are very similar to those published for the similar-in-principle Astrand-Rhyming test of predicted maximal oxygen consumption.[81]

We would conclude that these tests are probably not reliable enough to monitor the small changes in maximal oxygen consumption that result from increasing the training of an already athletic person.[82] However, these predictive tests may detect large differences in maximal oxygen consumption, for example, after an initially sedentary person performs a conditioning programme.[83] One could arrive at a more conclusive decision of adequate (or inadequate) reliability by using analytical goals based on sample sizes for future experimental uses. The SD of the differences (or the mean squared residual in the case of ANOVA) can be used to estimate sample sizes for repeated measures experiments.[6] It would be clear, even without such calculations, that the greater the random error component of the limits of agreement, the more individuals would be needed in an experiment for a given hypothesised experimental change. Alternatively, the greater the random error indicated by the limits of agreement, the larger the minimal detectable change would be for a given sample size in an experiment. Zar[6] also provides calculations for this issue of estimating minimal detectable changes for measurement tools. One cannot judge the magnitude of a correlation coefficient *per se* as simply as this, since there is an 'added factor' of interindividual variability in this statistic.

We have 3 comments on the use of limits of agreement in sports science and medicine:

(1) Only recently[39] has the limits of agreement method been applied to multiple retests using an ANOVA approach. This is preferable for the in-depth investigation of bias and also because the examination of heteroscedasticity is enhanced (the degrees of freedom are increased). The random

error component of the 95% limits of agreement is calculated from

$$1.96\sqrt{2 \times MSE}$$

where MSE is the mean squared error term from a repeated measures ANOVA. Recently, Bland and Altman[43,44] accepted that measurement error may be expressed in relation to a 'population' of repeated tests in individuals, which is the basis of SEM and CV. They calculated this from $\sqrt{MSE}$, which equates to one method of calculating the SEM.[69] They did stress, however, the need to multiply this value by 1.96 in order to represent the difference between measured and the 'true' value for 95% of observations. For the example data in table II, the $\sqrt{MSE}$ is 4.7 ml/kg/min so the '95% SEM' is $1.96 \times 4.7 = \pm 9.2$ ml/kg/min. For logged data, one would antilog the $\sqrt{MSE}$ from ANOVA and express this CV to the power of 1.96 to cover 95% of observations. This would be $1.097^{1.96} = \times/\div 1.20$ for the data in table II when expressed as a ratio.

Hopkins[38] cites a very similar statistic to Bland and Altman's 68% ratio CV of 1.097 (9.7%), although it is calculated in a slightly different way and always expressed as a percentage (±9.3% for our example data). Note that both these methods of calculating CV (from ANOVA) give slightly higher values than the 'naive estimator' of the mean value of 7.6% calculated from individual CVs. This agrees with the observations of Quan and Shih.[65] Note also that expressing a CV as ± percent rather than as ×/÷ ratio may be misleading since a characteristic of ratio data is that the range of error will always be slightly less, below a given measured value compared with the error above a measured value. The calculation of ± CV implies, erroneously with true ratio data, that the error is of equal magnitude either side of a particular measured value.

(2) Because the calculated limits of agreement are meant to be extrapolated to a given population, it is recommended that a large sample size ($n > 40$) is examined in any measurement study.[30] Bland and Altman[16] also advise the calculation of the standard errors of the limits of agreement to show how precise they are in relation to the whole population. From these, confidence intervals can be calculated, which may allow statistical meta-analysis for comparison of limits of agreement between different studies.

(3) The reliability examples cited in Bland and Altman's work[16,39] appear not to consider that bias can occur in repeated measurements.[84] Only the method comparison (validity) examples incorporate the bias estimation in the limits of agreement. This might be because the clinician is dealing frequently with biological assays, which are not affected by learning or fatigue of testing. Since these effects are likely to influence measurements of human performance, it is recommended that the bias between repeated trials is always reported (separately from the random error component) by the sports scientist.

10. Discussion

This review has attempted to evaluate the most common statistical methods for evaluating reliability. In view of the importance of minimal measurement error to sports science research and, although one book on the subject has been published,[17] it is surprising how neglected discussions on measurement issues are in sports science and medicine. An important point is that correlation methods (including ICC) should be interpreted with caution in such studies. This is a difficult notion to promote given the popularity of judging a high correlation as indicating adequate reliability. An implication of the poor interpretation of correlation analyses is that equipment used routinely in the sport and exercise sciences may have been erroneously concluded as being sufficiently reliable (realising certain analytical goals for sports science use). It would be sensible for researchers to reappraise the results of test-retest correlations and supplement this with the application of absolute indicators of reliability. Ideally, a database should exist providing information on the reliability of every measurement tool used routinely in sports medicine. This has been attempted, using correlation, with isokinetic

muscle strength measurements.[50] At present, the limits of agreement method has been applied most, amongst sport science–relevant variables, to the reliability and validity of adipose tissue measurements.[85-89]

The present review has attempted to highlight that some reliability statistics are cited in the sports science literature without adequate investigation of underlying assumptions. The important assumption regarding the relationship between error and the magnitude of the measured value is rarely explored by reliability researchers. It may be that with some measurements, the variability decreases instead of increases as the measured values increase (negative heteroscedasticity). In this case, the data might need to be transformed differently before application of an absolute indicator of reliability. Statisticians are currently working on such problems.[42] It is imperative that the sports physician keeps abreast of the correct statistical solutions to these issues. One practical recommendation is that future reliability studies include an examination of how the measurement error relates to the magnitude of the measured variables, irrespective of which type of absolute reliability statistic is employed (SEM, CV, limits of agreement). The simplest way to do this is by plotting the calculated residuals from ANOVA against the fitted values and observing if the classic 'funnelling' of heteroscedasticity is evident.

One issue which Bland and Altman consistently discuss in their work on measurement issues is that of 'method comparison'.[16,42] They maintain that the disadvantages of many statistics used in reliability studies also apply to studies investigating whether different methods can be used interchangeably or whether a method agrees with a gold standard measurement tool. They propose that the use of limits of agreement is also more appropriate in these situations, which happen to be very common in sports science as part of validity examinations.[90] Obviously, such a use of limits of agreement would be alien to the sports scientist who may be accustomed to hypothesis tests, and regression and correlation methods as part of this type of validity analysis. An important issue such as this should warrant further discussion amongst sports science researchers.

To conclude, it seems ironic that the many statistics designed to assess agreement seem so inconsistent in their quantification of measurement error and their interpretation amongst researchers for deciding whether a measurement tool can be reliably employed in future research work. In brief, there are difficulties with relative reliability statistics both in their interpretation and extrapolation of results to future research. There are also many different methods of calculating the reliability statistic, ICC. Moreover, the expression of absolute reliability statistics differs to such an extent that one statistic (SEM) can be calculated in a way ($SD\sqrt{1-ICC}$) that makes it still sensitive to population heterogeneity (i.e. not a true indicator of absolute reliability at all). There is also a general lack of exploration of associated assumptions with absolute reliability statistics and disagreement on the described proportion of measurement error (68 *vs* 95%).

While statistics will never be more important than a well designed reliability study itself, it is sensible that there should be a standardised statistical analysis for any reliability study involving ratio of interval measurements. This may entail using several reliability statistics, so that different researchers can interpret the one they are most accustomed to. To this end, we would suggest:

- The inclusion in any reliability study of an examination of the assumptions surrounding the choice of statistics, especially the presence or absence of heteroscedasticity.
- A full examination of any systematic bias in the measurements coupled with practical recommendations for future researchers on the number of pretest familiarisation sessions to employ and the advised recovery time between tests so that any bias due to fatigue is minimised.
- The inclusion of intraclass correlation analysis, but with full details of which type of ICC has been calculated and the citation of confidence intervals for the ICC. This analysis could be supplemented with an examination of relative

reliability through the test-retest stability of sample ranks or the relation of the degree of absolute reliability to the interindividual or between-centile differences in a population. This is recommended even if a high ICC (>0.9) has been obtained.

- The citation of the most popular measures of absolute reliability, depending on whether heteroscedasticity is present (CV, 'ratio limits of agreement') or absent (SEM, 'absolute limits of agreement'). It is preferable that these are calculated from the mean square error term in a repeated-measures ANOVA model. The described percentile of measurement error (68 or 95%) should also be stated.
- The arrival at an eventual decision of reliability (or not) based on the extrapolation of the measurement error to the realisation of 'analytical goals'. These may include the effectiveness of use of the measurement tool on individual cases, a meaningful degree of relative reliability and the implications of the measurement error for sample size estimation in experiments.

Acknowledgements

We are grateful to Dr Don MacLaren for providing, from his laboratory work, the data used to illustrate the various statistical analyses in this review. The comments of Professor Thomas Reilly, Dr Jim Waterhouse and Dr Richard Tong were also appreciated during the preparation of this manuscript.

References

1. Yeadon MR, Challis JH. The future of performance-related sports biomechanics research. J Sports Sci 1994; 12: 3-32
2. Jakeman PM, Winter EM, Doust J. A review of research in sports physiology. J Sports Sci 1994; 12: 33-60
3. Hardy L, Jones G. Current issues and future directions for performance-related research in sport psychology. J Sports Sci 1994; 12: 61-92
4. Nevill AM. Statistical methods in kinanthropometry and exercise physiology. In: Eston R, Reilly T, editors. Kinanthropometry and exercise physiology laboratory manual. London: E and FN Spon, 1996: 297-320
5. Safrit MJ. An overview of measurement. In: Safrit MJ, Wood TM, editors. Measurement concepts in physical education and exercise science. Champaign (IL): Human Kinetics, 1989: 3-20
6. Zar JH. Biostatistical analysis. London: Prentice Hall, 1996
7. Mathews JN. A formula for the probability of discordant classification in method comparison studies. Stat Med 1997; 16 (6): 705-10
8. Bates BT, Dufek JS, Davis HP. The effects of trial size on statistical power. Med Sci Sports Exerc 1992; 24 (9): 1059-65
9. Dufek JS, Bates BT, Davis HP. The effect of trial size and variability on statistical power. Med Sci Sports Exerc 1995; 27: 288-95
10. Atkinson G. [Letter]. British Association of Sports Sciences Newsletter, 1995 Sep: 5
11. Nevill AM. Validity and measurement agreement in sports performance [abstract]. J Sports Sci 1996; 14: 199
12. Ottenbacher KJ, Stull GA. The analysis and interpretation of method comparison studies in rehabilitation research. Am J Phys Med Rehab 1993; 72: 266-71
13. Hollis S. Analysis of method comparison studies. Ann Clin Biochem 1996; 33: 1-4
14. Liehr P, Dedo YL, Torres S, et al. Assessing agreement between clinical measurement methods. Heart Lung 1995; 24: 240-5
15. Ottenbacher KJ, Tomcheck SD. Measurement variation in method comparison studies: an empirical examination. Arch Phys Med Rehabil 1994; 75 (5): 505-12
16. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; I: 307-10
17. Safrit MJ, Wood TM, editors. Measurement concepts in physical education and exercise science. Champaign (IL): Human Kinetics, 1989
18. Baumgarter TA. Norm-referenced measurement: reliability. In: Safrit MJ, Wood TM, editors. Measurement concepts in physical education and exercise science. Champaign (IL): Human Kinetics, 1989: 45-72
19. Atkinson G. Reilly T. Circadian variation in sports performance. Sports Med 1996; 21 (4): 292-312
20. Morrow JR, Jackson AW, Disch JG, et al. Measurement and evaluation in human performance. Champaign (IL): Human Kinetics, 1995
21. Morrow JR. Generalizability theory. In: Safrit MJ, Wood TM, editors. Measurement concepts in physical education and exercise science. Champaign (IL): Human Kinetics, 1989: 73-96
22. Roebroeck ME, Harlaar J, Lankhorst GJ. The application of generalizability theory to reliability assessment: an illustration using isometric force measurements. Phys Ther 1993; 73 (6): 386-95
23. Chatburn RL. Evaluation of instrument error and method agreement. Am Assoc Nurse Anesthet J 1996; 64 (3): 261-8
24. Coldwells A, Atkinson G, Reilly T. Sources of variation in back and leg dynamometry. Ergonomics 1994; 37: 79-86
25. Hickey MS, Costill DL, McConnell GK, et al. Day-to-day variation in time trial cycling performance. Int J Sports Med 1992; 13: 467-70
26. Nevill A. Why the analysis of performance variables recorded on a ratio scale will invariably benefit from a log transformation. J Sports Sci 1997; 15: 457-8
27. Bland JM, Altman DG. Transforming data. BMJ 1996; 312 (7033): 770
28. Schultz RW. Analysing change. In: Safrit MJ, Wood TM, editors. Measurement concepts in physical education and exercise science. Champaign (IL): Human Kinetics, 1989: 207-28
29. Morrow JR, Jackson AW. How 'significant' is your reliability? Res Q Exerc Sport 1993; 64 (3): 352-5
30. Altman DG. Practical statistics for medical research. London: Chapman and Hall, 1991: 396-403
31. Mathews JNS, Altman DG, Campbell MJ, et al. Analysis of serial measurements in medical research. BMJ 1990; 300: 230-5

32. Vincent J. Statistics in kinesiology. Champaign (IL): Human Kinetics Books, 1994
33. Ross JW, Fraser MD. Analytical goals developed from the inherent error of medical tests. Clin Chem 1993; 39 (7): 1481-93
34. Fraser CG, Hyltoft Peterson P, et al. Setting analytical goals for random analytical error in specific clinical monitoring situations. Clin Chem 1990; 36 (9): 1625-8
35. Zehr ER, Sale DG. Reproducibility of ballistic movement. Med Sci Sports Exerc 1997; 29: 1383-8
36. Hofstra WB, Sont JK, Sterk PJ, et al. Sample size estimation in studies monitoring exercise-induced bronchoconstriction in asthmatic children. Thorax 1997; 52: 739-41
37. Schabort EJ, Hopkins WG, Hawley JA. Reproducibility of self-paced treadmill performance of trained endurance runners. Int J Sports Med 1998; 19: 48-51
38. Hopkins W. A new view of statistics. Internet site, 1997, http://www.sportsci.org/resource/stats/index.html
39. Bland M. An introduction to medical statistics. Oxford: University Press, 1995
40. Proceedings of the 43rd Meeting of the American College of Sports Medicine. Med Sci Sports Exerc 1996; 28: S1-211
41. Altman DG, Bland JM. Measurement in medicine: the analysis of method comparison studies. Statistician 1983; 32: 307-17
42. Bland JM, Altman DG. Comparing two methods of clinical measurement: a personal history. Int J Epidemiol 1995; 24 Suppl. 1: S7-14
43. Bland JM, Altman DG. Measurement error. BMJ 1996; 312 (7047): 1654
44. Bland JM, Altman DG. Measurement error proportional to the mean. BMJ 1996; 313 (7049): 106
45. Thomas JR, Nelson JK. Research methods in physical activity. Champaign (IL): Human Kinetics, 1990
46. Nevill AN, Atkinson G. Assessing measurement agreement (repeatability) between 3 or more trials [abstract]. J Sports Sci 1998; 16: 29
47. Coolican H. Research methods and statistics in psychology. London: Hodder and Stoughton, 1994
48. Sale DG. Testing strength and power. In: MacDougall JD, Wenger HA, Green HJ, editors. Physiological testing of the high performance athlete. Champaign (IL): Human Kinetics, 1991: 21-106
49. Bates BT, Zhang S, Dufek JS, et al. The effects of sample size and variability on the correlation coefficient. Med Sci Sports Exerc 1996; 28 (3): 386-91
50. Perrin DH. Isokinetic exercise and assessment. Champaign (IL): Human Kinetics, 1993
51. Glass GV, Hopkins KD. Statistical methods in education and psychology. 2nd ed. Englewood Cliffs (NJ): Prentice-Hall, 1984
52. Estelberger W, Reibnegger G. The rank correlation coefficient: an additional aid in the interpretation of laboratory data. Clin Chim Acta 1995; 239 (2): 203-7
53. Nevill AN, Atkinson G. Assessing agreement between measurements recorded on a ratio scale in sports medicine and sports science. Br J Sports Med 1997; 31: 314-8
54. Atkinson G, Greeves J, Reilly T, et al. Day-to-day and circadian variability of leg strength measured with the lido isokinetic dynamometer. J Sports Sci 1995; 13: 18-9
55. Bailey SM, Sarmandal P, Grant JM. A comparison of three methods of assessing inter-observer variation applied to measurement of the symphysis-fundal height. Br J Obstet Gynaecol 1989; 96 (11): 1266-71
56. Sarmandal P, Bailey SM, Grant JM. A comparison of three methods of assessing inter-observer variation applied to ultrasonic fetal measurement in the third trimester. Br J Obstet Gynaecol 1989; 96 (11): 1261-5
57. Atkinson G, Coldwells A, Reilly T, et al. Does the within-test session variation in measurements of muscle strength depend on time of day? [abstract] J Sports Sci 1997; 15: 22
58. Charter RA. Effect of measurement error on tests of statistical significance. J Clin Exp Neuropsychol 1997; 19 (3): 458-62
59. Muller R, Buttner P. A critical discussion of intraclass correlation coefficients. Stat Med 1994; 13: 23-4, 2465-76
60. Eliasziw M, Young SL, Woodbury MG, et al. Statistical methodology for the concurrent assessment of inter-rater and intra-rater reliability: using goniometric measurements as an example. Phys Ther 1994; 74 (8): 777-88
61. Krebs DE. Declare your ICC type [letter]. Phys Ther 1986; 66: 1431
62. Atkinson G. A comparison of statistical methods for assessing measurement repeatability in ergonomics research. In: Atkinson G, Reilly T, editors. Sport, leisure and ergonomics. London: E and FN Spon, 1995: 218-22
63. Bland JM, Altman DG. A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. Comput Biol Med 1990; 20: 337-40
64. Myrer JW, Schulthies SS, Fellingham GW. Relative and absolute reliability of the KT-2000 arthrometer for uninjured knees. Testing at 67, 89, 134 and 178 N and manual maximum forces. Am J Sports Med 1996; 24 (1): 104-8
65. Quan H, Shih WJ. Assessing reproducibility by the within-subject coefficient of variation with random effects models. Biometrics 1996; 52 (4): 1195-203
66. Lin LI-K. A concordance correlation coefficient to evaluate reproducibility. Biometrics 1989; 45: 255-68
67. Nickerson CAE. A note on 'A concordance correlation coefficient to evaluate reproducibility'. Biometrics 1997; 53: 1503-7
68. Atkinson G, Nevill A. Comment on the use of concordance correlation to assess the agreement between two variables. Biometrics 1997; 53: 775-7
69. Stratford PW, Goldsmith CH. Use of the standard error as a reliability index of interest: an applied example using elbow flexor strength data. Phys Ther 1997; 77 (7): 745-50
70. Payne RW. Reliability theory and clinical psychology. J Clin Psychol 1989; 45 (2): 351-2
71. Strike PW. Statistical methods in laboratory medicine. Oxford: Butterworth-Heinemann, 1991
72. Fetz CJ, Miller GE. An asymptotic test for the equality of coefficients of variation from k populations. Stat Med 1996; 15 (6): 646-58
73. Allison DB. Limitations of coefficient of variation as index of measurement reliability [editorial]. Nutrition 1993; 9 (6): 559-61
74. Yao L, Sayre JW. Statistical concepts in the interpretation of serial bone densitometry. Invest Radiol 1994; 29 (10): 928-32
75. Detwiler JS, Jarisch W, Caritis SN. Statistical fluctuations in heart rate variability indices. Am J Obstet Gynecol 1980; 136 (2): 243-8
76. Stokes M. Reliability and repeatability of methods for measuring muscle in physiotherapy. Physiother Pract 1985; 1: 71-6
77. Bishop D. Reliability of a 1-h endurance performance test in trained female cyclists. Med Sci Sports Exerc 1997; 29: 554-9
78. Bland JM, Altman DG. Comparing methods of measurement: why plotting difference against the standard method is misleading. Lancet 1995; 346 (8982): 1085-7

79. British Standards Institution. Precision of test methods I. Guide for the determination and reproducibility for a standard test method. BS5497: Pt 1. London: BSI, 1979
80. de Jong JS, van Diest PJ, Baak JPA. In response [letter]. Lab Invest 1996; 75 (5): 756-8
81. Wisen AG, Wohlfart B. A comparison between two exercise tests on cycle; a computerised test versus the Astrand test. Clin Physiol 1995; 15: 91-102
82. Wilmore JH, Costill DL. Physiology of sport and exercise. Champaign (IL): Human Kinetics, 1994
83. Pollock ML. Quantification of endurance training programmes. Exerc Sports Sci Rev 1973; 1: 155-88
84. Doyle JR, Doyle JM. Measurement error is that which we have not yet explained. BMJ 1997; 314: 147-8
85. Schaefer F, Georgi M, Zieger A, et al. Usefulness of bioelectric impedance and skinfold measurements in predicting fat-free mass derived from total body potassium in children. Pediatr Res 1994; 35: 617-24
86. Webber J, Donaldson M, Allison SP, et al. Comparison of skinfold thickness, body mass index, bioelectrical impedance analysis and x-ray absorptiometry in assessing body composition in obese subjects. Clin Nutr 1994; 13: 177-82
87. Fuller NJ, Sawyer MB, Laskey MA, et al. Prediction of body composition in elderly men over 75 years of age. Ann Hum Biol 1996; 23: 127-47
88. Gutin B, Litaker M, Islam S, et al. Body composition measurement in 9-11 year old children by dual energy x-ray absorptiometry, skinfold thickness measures and bioimpedance analysis. Am J Clin Nutr 1996; 63: 287-92
89. Reilly JJ, Wilson J, McColl JH, et al. Ability of bioelectric impedance to predict fat-free mass in prepubertal children. Pediatr Res 1996; 39: 176-9
90. Wood TM. The changing nature of norm-referenced validity. In: Safrit MJ, Wood TM, editors, Measurement concepts in physical education and exercise science. Champaign (IL): Human Kinetics, 1989: 23-44

Correspondence and reprints: Dr *Greg Atkinson,* Research Institute for Sport and Exercise Sciences, Trueman Building, Webster Street, Liverpool John Moores University, Liverpool L3 2ET, England.
E-mail: g.atkinson@livjm.ac.uk