Br J Sports Med 1997;**31**:314–318

Assessing agreement between measurements recorded on a ratio scale in sports medicine and sports science

Alan M Nevill, Greg Atkinson

Abstract

***Objective*—The consensus of opinion suggests that when assessing measurement agreement, the most appropriate statistic to report is the "95% limits of agreement". The precise form that this interval takes depends on whether a positive relation exists between the differences in measurement methods (errors) and the size of the measurements—that is, heteroscedastic errors. If a positive and significant relation exists, the recommended procedure is to report "the ratio limits of agreement" using log transformed measurements. This study assessed the prevalence of heteroscedastic errors when investigating measurement agreement of variables recorded on a ratio scale in sports medicine and sports science.**

***Methods*—Measurement agreement (or repeatability) was assessed in 13 studies (providing 23 examples) conducted in the Centre for Sport and Exercise Sciences at Liverpool John Moores University over the past five years.**

***Results*—The correlation between the absolute differences and the mean was positive in all 23 examples (median r=0.37), eight being significant ($P<0.05$). In 21 of 23 examples analysed, the correlation was greater than the equivalent correlation using log transformed measurements (median r=0.01). Based on a simple meta-analysis, the assumption that no relation exists between the measurement differences and the size of measurement must be rejected ($P<0.001$).**

***Conclusions*—When assessing measurement agreement of variables recorded on a ratio scale in sports medicine and sports science, this study (23 examples) provides strong evidence that heteroscedastic errors are the norm. If the correlation between the absolute measurement differences and the means is positive (but not necessarily significant) and greater than the equivalent correlation using log transformed measurements, the authors recommend reporting the "ratio limits of agreement".**

(*Br J Sports Med* 1997;**31**:314–318)

Keywords: heteroscedastic errors; ratio limits of agreement; dimensionless ratio

School of Human Sciences, Liverpool John Moores University
A M Nevill
G Atkinson

Correspondence to:
A M Nevill, Centre for Sport and Exercise Sciences, School of Human Sciences, Liverpool John Moores University, Byrom Street, Liverpool L3 3AK,United Kingdom.

Accepted for publication
22 July 1997

Sports scientists are always seeking better or simpler methods, or both, of measuring variables associated with sports performance. In the past, various statistical techniques have been adopted to assess whether a new method of measuring a variable is either repeatable or equivalent to an established method. These techniques include Pearson's correlation coefficient, intra-class correlation coefficients, the correlated (paired) t test, and repeated measures analysis of variance.

More recently however, Altman and Bland[1] and Bland and Altman[2] have criticised the use of many of these techniques, in particular, the use of correlation coefficients because they are measures of relation rather than agreement and are highly influenced by the range of subjects' measurements. For example, when comparing the results of a new test to measure V_{O_2} max with an existing test, if the chosen sample contains young and old, male and female, and heavy and light subjects, the correlation is likely to be high. If, on the other hand, the same two tests are to be compared using only male subjects, all of approximately the same age and similar body mass, the correlation between the results of the two tests will be relatively small, but not necessarily less valid.

Similarly, the correlated (paired) t test or repeated measures analysis of variance are equally inconclusive when assessing repeatability or agreement between two or more measurement methods. Although such tests of significance will formally examine the hypothesis that no bias exists between the repeated measurements—that is, H_0: $\mu_1 = \mu_2 = ... = \mu_n$ versus H_1: $\mu_1 \neq \mu_2 \neq ... \neq \mu_n$, if the variance within subjects (the residual mean square) is large, the null hypothesis (H_0) could be accepted but the repeated measurements will still display unacceptable random variation.

Bland and Altman[2] propose an alternative approach, based on the differences between the two measurement methods (measurement errors), using simple calculations and graphical techniques. Provided no obvious relation can be detected between the measurement errors and the mean of the two observations, Bland and Altman[2] recommended reporting a simple interval known as the "limits of agreement", based on the standard deviation of the differences between measurements. Assuming the differences are normally distributed, the limits should contain 95% of such differences. Bland and Altman[2] argued that the scientist or clinician can then use their own judgement to assess the acceptability or agreement associated with this interval and hence the measurement methods.

To assess whether the errors depend on the size of the measurements (usually a larger error being associated with a larger measurement mean—that is, heteroscedastic errors), Bland and Altman[2] recommended a scatter diagram of the differences (errors) against the measurement means. If a possible relation is detected, it can be confirmed by calculating the correlation between the absolute differences and the mean. If found to be positive and significant, the authors recommend taking logarithms of both measurement methods and proceed as before by reporting the "limits of agreement" but using the log scale. However, the authors acknowledge that by reporting the limits of agreement on the log scale, the resulting antiloged interval becomes the difference between two dimensionless ratios.

When modelling measurements such as maximum oxygen uptake, strength and power, various authors[3-5] observed the presence of heteroscedastic errors. This is not too surprising as the range of all variables, recorded on a ratio scale (variables that cannot be negative and have a natural zero point) is forced to remain non-negative at the lower end of the scale but is theoretically unbounded at the other. This will naturally lead to heteroscedastic errors when two measurement methods, both on the same ratio scale, are to be compared and assessed for measurement agreement. For example, when one measurement method is plotted against the other, the spread of data at the bottom left hand corner of the plot is constrained by the origin (coordinate X=0, Y=0), but no such constraint occurs with the data in the top right hand corner of the plot. In his book, Bland[6] reported two examples, both using a variable recorded on a ratio scale (peak expiratory flow rates), that provide positive heteroscedastic errors, although the author rejects the need to take logarithms as neither of the correlations (between the absolute differences and the mean) prove to be significant. This lack of significance may simply be caused by the comparatively small sample sizes. Hence, the purpose of this study is to examine a large number of similar studies, performed at the Centre for Sport and Exercise Sciences, Liverpool John Moores University, to assess the extent or prevalence of heteroscedastic errors and the appropriateness of using the log transformation when assessing measurement agreement of variables recorded on a ratio scale in sports medicine and sports science.

Methods

Assuming no relation is found between the measurement differences (errors) and their mean, the "95% limits of agreement" are obtained as follows: (1) calculate the mean (d) and the standard deviation (s) of the differences that indicates the level of bias and the random variation between the two methods, respectively. (2) Provided the differences are normally distributed, the 95% "limits of agreement" are given by $d \pm (1.96 \times s)$.

Bland and Altman[2] argued that provided differences within these limits are not clinically important, the two measurement methods can be used interchangeably.

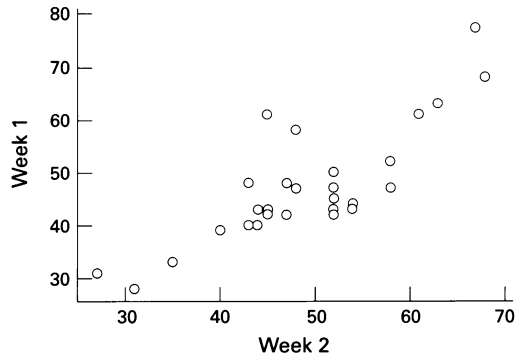

Figure 1 The relation between estimates of maximum oxygen uptake ml/kg/min, obtained on two consecutive weeks (weeks 1 and 2), using the Fitech step test ($r=0.80$, $P<0.001$).

To examine whether a positive relation exists between the measurement error and the mean, Bland and Altman[2] recommend a plot of the differences (errors) against the measurement mean (known as the Bland and Altman plot). This can be confirmed by calculating the correlation between the absolute differences and the mean. If a positive relation is observed, the analysis described above should be applied to the log transformed measurements (once again provided the differences between the natural log transformed measurements are normally distributed). (Note that the difference between two log transformed measurements is equivalent to the log transformation of the ratio between the two measurements—that is, $\log_e(X_1)-\log_e(X_2)=\log_e(X_1/X_2)$). By taking antilogs of the resulting "limits of agreement", we obtain an average (the geometric mean) dimensionless ratio (obtained by dividing one measurement method by the second) that describes the measurement bias, multiplied or divided by a second ratio that indicates the level of agreement. The latter ratio is not dissimilar to the concept of a coefficient of variation except the new ratio limits should contain 95% of the observed ratios. Note that if the "agreement ratio" were equal to 1, we would have perfect agreement between the measurement methods.

Over the past five years, the School of Human Sciences at Liverpool John Moores University has carried out a number of studies to assess measurement agreement or repeatability. The present work will examine 13 such studies that provide a total of 23 examples. The subjects were all recreationally active male and female students, aged between 18 and 30 years—that is, relatively heterogeneous samples.

Results

To illustrate the alternative methods used to assess measurement agreement, we shall examine the repeatability of the Fitech step test (study 3), designed to estimate maximum oxygen uptake ml/kg/min, measured on two consecutive weeks. The two weeks' results are plotted in figure 1 ($r=0.8$; $P<0.001$).

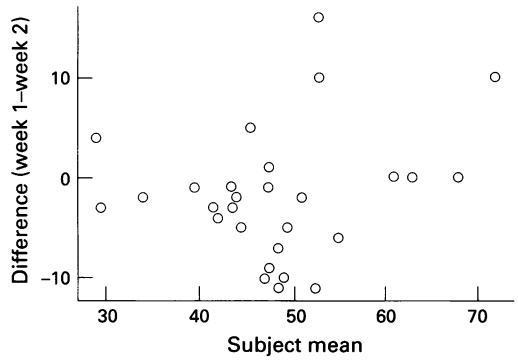

Figure 2 The differences (errors) between the estimated maximum oxygen uptake (ml/kg/min) results from weeks 1 and 2, plotted against the subjects' measurement means (mean of week 1 and 2 for each subject).

The mean (SD) difference between estimates on the first and second occasion, −1.47 (6.61), was not significantly biased (t=−1.22, P>0.05). Assuming the differences are normally distributed, 95% of the differences should lie between the limits −1.47 (13.0)—that is, from −14.47 to 11.53 ml/kg/min regardless of the subjects' mean performance.

To help the reader interpret these limits, if the subject's estimated maximum oxygen uptake performance was 30 ml/kg/min on the first week, it is possible (worst case scenario) that the same subject could obtain an estimate as low as 15.53 ml/kg/min, or as high as 41.53 ml/kg/min, on the second week. Indeed, if the subject's performance on the first week had been much higher at 70 ml/kg/min, the second weeks performance might have been as low as 55.53 ml/kg/min or as high as 81.53 ml/kg/min. This range is unlikely to be acceptable to many sports scientists involved in exercise/performance testing, especially when the subjects' performances tend towards the lower end of the ratio scale, for example 30 ml/kg/min.

Unfortunately, the observed differences from study 3 were not normally distributed, using the Anderson-Darling test for normality as implemented in MINITAB.[7] Furthermore there was evidence that a positive relation exists between the differences and the mean. To illustrate the recommended process to assess whether a positive relation exists between the differences and the mean, the differences (errors) between the estimated maximum oxygen uptake results from weeks 1 and 2 were plotted against the subjects' measurement means (fig 2).

The positive relation was confirmed when the correlation between the absolute differences and the mean was found to be r=0.179. Adopting the alternative approach of taking natural logarithms of the measurements, the mean (SD) difference was found to be −0.0356 (0.131). Once again, no significant bias was observed (t=−1.49; P>0.05), but the differences between the log transformed estimates were now normally distributed. Hence, 95% of the differences should lie between the limits −0.0356 ((0.131)×1.96)—that is, from −0.293 to 0.221. Taking antilogs of these values, the mean (bias) ratio was estimated as 0.97, multiplied or divided by the agreement ratio (*/÷ 1.29)—that is, 95% of the ratios (measurement 1 divided by measurement 2) should be contained between 0.75 and 1.25. The histogram of the ratios (week 1 divided by week 2) is given in figure 3. All but one of the ratios (n=30) appears to be contained between the calculated limits (0.75 to 1.25).

To help interpret these "ratio limits of agreement", if the subject's estimated maximum oxygen uptake performance was 30 ml/kg/min on the first week, it is possible (worst case scenario) that the same subject could obtain an estimate as low as 30×0.75 =22.5 ml/kg/min, or as high as 30×1.25= 37.5 ml/kg/min, on the second week. For a subject with a higher performance on the first week at 70 ml/kg/min, the second weeks performance might be as low as 70×0.75=52.5 ml/kg/min or as high as 70×1.25=87.5 ml/kg/min. As can be seen, these ratio limits now vary in absolute terms but remain a constant ratio or percentage change in performance from week 1 to week 2. Even though these ratio limits are still too wide to be acceptable for most sports scientists, they are more realistic in the way they are allowed to vary depending on the level of the subjects' performance.

Table 1 summarises the results from all 13 studies (23 examples), including the sample size, the "limits of agreement", together with the correlation between the absolute differences and the mean. The table also indicates whether the correlation is significant (P<0.05) and whether the differences are normally distributed.

Note that the differences were not normally distributed in five of the 23 examples (Anderson-Darling normality test) and the correlation between the absolute differences and the mean is positive in all 23 examples, eight being significant (P<0.05). By combining the results of all 23 examples using a non-parametric sign test (a simple meta-analysis), the assumption that no relation exists between the measurement differences (errors) and their mean, must be rejected (P<0.001). The median correlation was r=0.37.

Assuming a positive relation exists between the measurement differences (errors) and the mean, the analysis was repeated for all 23 examples, using log transformed measurements. Table 2 summarises the results, including the sample size, the mean of the log transformed measurement, and their mean

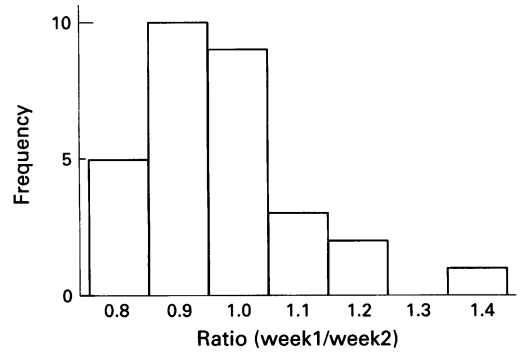

Figure 3 The histogram of the ratios, obtained by dividing the estimates of maximum oxygen uptake from week 1 by those from week 2.

Table 1 The sample size, the measurement means and differences, the absolute "limits of agreement", together with the correlation between the absolute differences amd the mean

Study (units)	Measurements					Correlation (abs (diff) v mean)
	Sample size	Mean 1	Mean 2	Difference (SD)	Absolute limits	
1 Astrand rhyming (ml/kg/min)	30	54.9	55.8	−0.90 (8.48)	−0.90 (16.6)	0.05
2 Back strength (kg)	12	93.0	91.0	1.98 (8.73)	1.98 (17.1)	0.42
3 Fitech step test (ml/kg/min)	30	47.4	48.9	−1.47 (6.61)†	−1.47 (13.0)	0.18
4 Grip strength left (kg)	12	36.6	37.2	−0.62 (1.09)	−0.62 (2.14)	0.24
5 Grip strength right (kg)	12	39.9	40.5	−0.62 (3.04)	−0.62 (5.96)	0.33
6 Isokinetic leg strength (peak torque) (Nm)						
Extension 1.05 rad/s	10	217.3	225.0	−7.70 (21.3)	−7.70 (41.8)	0.44
Flexion 1.05 rad/s	10	128.0	129.2	−1.20 (16.3)	−1.20 (31.9)	0.31
Extension 3.13 rad/s	10	162.0	162.2	−0.20 (8.4)†	−0.20 (15.8)	0.32
Flexion 3.13 rad/s	10	105.5	101.6	3.90 (8.52)	3.90 (16.7)	0.45
Extension 5.22 rad/s	10	126.1	126.2	−0.10 (9.16)	−0.10 (18.0)	0.40
Flexion 5.22 rad/s	10	89.6	86.5	3.10 (6.95)	3.10 (13.6)	0.69*
7 Lactate threshold (mmol/l)	9	1.78	1.75	−0.032 (1.02)	−0.032 (2.00)	0.75*
8 Leg strength (kg)	12	152.3	149.5	2.79 (17.0)	2.79 (33.3)	0.59*
9 Margaria (W)	14	1022.7	1005.9	16.9 (55.9)†	16.9 (109.5)	0.24
10 Isokinetic trunk strength (peak torque) (Nm)						
Flexion 1.05 rad/s	31	179.6	198.8	−19.16 (48.1)	−19.16 (94.3)	0.60*
Extension 1.05 rad/s	31	147.6	165.4	−17.71 (35.8)	−17.71 (70.2)	0.37*
Flexion 1.57 rad/s	23	197.5	206.3	−8.8 (52.7)	−8.8 (103.3)	0.31
Extension 1.57 rad/s	23	164.6	176.8	−12.17 (45.0)	−12.17 (88.1)	0.34
Flexion 2.09 rad/s	31	190.8	202.0	−11.19 (49.7)	−11.19 (97.4)	0.56*
Extension 2.09 rad/s	31	153.2	159.5	−6.39 (47.3)	−6.39 (92.7)	0.51*
11 Vertical jump (W)	15	1132.4	1115.7	16.8 (102.2)†	16.8 (200.2)	0.03
12 Wingate mean power (W)	13	657.1	663.2	−6.1 (83.5)†	−6.1 (163.7)	0.86*
13 Wingate peak power (W)	13	867.7	903.9	−36.2 (182.6)	−36.2 (182.6)	0.34

* Significant correlation ($P < 0.05$).
† Differences not normally distributed ($P < 0.05$); using the Anderson-Darling normality test.

differences, but the "limits of agreement" were expressed as a dimensionless ratio, multiplied or divided by the second ratio as a measure of agreement having already taken antilogs. The table also gives the correlation between the absolute differences and the mean (using the log transformed data), whether the correlation is significant ($P<0.05$) and whether the log transformed differences are normally distributed.

Having taken logarithms, the differences were not normally distributed in just two of the 23 examples, described in table 2. The correlation between the absolute differences and the mean was positive in 14 examples but negative in nine, none of which were significant. Again by combining the results of all 23 examples and using a non-parametric sign test, the assumption that no relation exists between the measurement differences (errors) and their mean, could be accepted ($P>0.10$). The median correlation was $r=0.01$.

Discussion

Based on the significant test-retest correlation ($r=0.80$, $P<0.001$) and non-significant paired t test ($t=-1.22$, $P>0.05$) from study 3, researchers in the past might have concluded that the Fitech step test was repeatable. However, if on the first week a subject's estimated maximum

Table 2 The sample size, the log transformed (ln) measurement means and differences, the "ratio limits of agreement", together with the correlation between the absolute differences amd the mean (log transformed)

Study (units)	Log transformed measurements					Correlation (abs (diff) v mean)
	Sample size	Mean 1	Mean 2	Difference (SD)	Ratio limits	
1 Astrand rhyming (ml/kg/min)	30	3.986	4.002	−0.0156 (0.157)	0.98 (*/÷ 1.36)	−0.08
2 Back strength (kg)	12	4.509	4.491	0.0182 (0.089)	1.02 (*/÷ 1.19)	0.23
3 Fitech step test (ml/kg/min)	30	3.835	3.871	−0.0356 (0.131)	0.97 (*/÷ 1.29)	0.01
4 Grip strength left (kg)	12	3.567	3.582	−0.0156 (0.031)	0.98 (*/÷ 1.06)	−0.18
5 Grip strength right (kg)	12	3.665	3.673	−0.0082 (0.073)	0.99 (*/÷ 1.15)	0.01
6 Isokinetic leg strength (peak torque) (Nm)						
Extension 1.05 rad/s	10	5.349	5.387	−0.0383 (0.088)	0.96 (*/÷ 1.19)	0.16
Flexion 1.05 rad/s	10	4.811	4.912	−0.0118 (0.129)	0.99 (*/÷ 1.29)	−0.20
Extension 3.13 rad/s	10	5.063	5.069	−0.0060 (0.047)	0.99 (*/÷ 1.10)	0.03
Flexion 3.13 rad/s	10	4.631	4.598	0.0337 (0.076)	1.03 (*/÷ 1.16)	0.34
Extension 5.22 rad/s	10	4.806	4.814	0.0072 (0.069)	0.99 (*/÷ 1.14)	0.01
Flexion 5.22 rad/s	10	4.472	4.435	0.0370 (0.071)†	1.04 (*/÷ 1.15)	0.35
7 Lactate threshold (mmol/l)	9	0.341	0.349	−0.0970 (0.562)	0.91 (*/÷ 3.01)	0.01
8 Leg strength (kg)	12	5.003	4.983	0.0196 (0.100)	1.02 (*/÷ 1.22)	0.51
9 Margaria (W)	14	6.901	6.886	0.0155 (0.055)	1.02 (*/÷ 1.11)	0.01
10 Isokinetic trunk strength (peak torque) (Nm)						
Flexion 1.05 rad/s	31	5.129	5.218	−0.0895 (0.234)	0.91 (*/÷ 1.58)	0.21
Extension 1.05 rad/s	31	4.918	5.048	−0.1298 (0.220)	0.88 (*/÷ 1.54)	−0.16
Flexion 1.57 rad/s	23	5.222	5.284	−0.0617 (0.247)	0.94 (*/÷ 1.62)	−0.03
Extension 1.57 rad/s	23	5.038	5.125	−0.0869 (0.245)	0.92 (*/÷ 1.62)	−0.03
Flexion 2.09 rad/s	31	5.176	5.251	−0.0751 (0.233)	0.93 (*/÷ 1.58)	−0.02
Extension 2.09 rad/s	31	4.957	5.007	−0.0501 (0.273)	0.95 (*/÷ 1.71)	0.23
11 Vertical jump (W)	15	7.003	6.983	0.0203 (0.094)†	1.02 (*/÷ 1.20)	−0.04
12 Wingate mean power (W)	13	6.415	6.419	−0.0033 (0.080)	1.00 (*/÷ 1.17)	0.54
13 Wingate peak power (W)	13	6.707	6.745	−0.0384 (0.111)	0.96 (*/÷ 1.24)	−0.11

* Significant correlation ($P < 0.05$).
† Differences not normally distributed ($P < 0.05$); using the Anderson-Darling normality test.

oxygen uptake performance was 30 ml/kg/min, reporting the absolute limits of agreement, it is possible (the worst case scenario) that the same subject could obtain an estimate as low as 15.53 ml/kg/min or as high as 41.53 ml/kg/min on the second week. Even when the ratio limits of agreement were calculated, it is possible that the same subject could obtain an estimate as low as 22.5 ml/kg/min, or as high as 37.5 ml/kg/min, on the second week. Clearly, this range is unlikely to be acceptable to most sport scientists involved in exercise/performance testing, especially if the absolute limits of agreement are reported and the subject's performance tend towards the lower end of the ratio scale.

We must acknowledge that the 23 examples of measurement agreement carried out over the past five years in the School of Human Sciences at Liverpool John Moores University, do not represent a random sample of such studies in sports medicine and sports science. Nevertheless, the combined results provide strong evidence that a positive relation exists between the measurement differences (errors) and the size of measurements of variables recorded on a ratio scale taken from such disciplines—that is, where a larger measurement error is associated with a larger measurement mean (heteroscedastic errors).

This is not too surprising, as most human performance variables are recorded on a ratio scale (variables with a natural zero point) and, as such, the range of measurements must remain non-negative in one direction but are theoretically unbounded in the other direction. This will naturally lead to heteroscedastic errors when two measurement methods, both recorded on a ratio scale, are to be compared and assessed for measurement agreement.

The major advantage of using the log transformation approach to measurement assessment, is that the resulting "ratio limits of agreement" enables the scientist to compare the quality of measurement agreement from a variety of studies using a dimensionless ratio as measure of bias, multiplied or divided by a second ratio that indicates the level of agreement.

By observing the agreement ratios in table 2, we can immediately compare the quality of measurement agreement or precision of all 23 examples regardless of the units involved. Comparing the agreement ratios observed in table 2, we see that study 4 shows the greatest agreement, with little bias, 0.98 and an excellent agreement ratio (*/÷ 1.06)—that is, 95% of ratios are constrained between approximately 6% of the mean bias ratio, 0.98/1.06=0.925 and 0.98*1.06=1.039. (Note that by multiplying the mean ratio by 1.06, this will increase the mean by 6%). The worst agreement was found with study 7, where two methods of measuring lactate at the "lactate threshold" were compared (the Lactate minimum versus the D-max method). Although the bias ratio is not great, given as 0.91, the agreement ratio (*/÷ 3.01) implies that 95% of ratios will lie between 301% of the mean bias ratio—that is, from (0.91/3.01=0.303) to (0.91*3.01=2.74). In this example, we might expect some of the lactate measurements taken using the first method to be three times larger (or smaller) than the lactate measurements using the second method.

In summary, based on 13 studies (providing 23 examples) designed to assess measurement agreement in sports medicine and sports science, the assumption that no relation exists between the measurement differences (errors) and the size of measurement must be rejected ($P<0.001$). These examples, plus additional examples provided by Bland[6] and Atkinson,[8] provide strong evidence that when assessing measurement agreement of variables recorded on a ratio scale in sports medicine and sports science, heteroscedastic errors are the norm and, as such, advocate the use of the log transformation when assessing measurement agreement. Not only did the log transformation reduce the correlation between the absolute measurement differences and the mean in 21 of 23 examples (see tables 1 and 2), but the differences of the log transformed measurements were normally distributed in all but two of the examples (compared with five examples that were found not to be normal distributed using the differences of the untransformed measurements).

When assessing measurement agreement or repeatability of variables recorded on a ratio scale, the present authors recommend taking natural logarithms of the measurement methods if the correlation between the absolute measurement differences and the means is positive (not necessarily significant, especially with small sample sizes) and, the correlation is reduced numerically (irrespective of the sign) having first taken logarithms of both measurement methods. This proved to be the case in 21 of 23 examples described above and of the remaining two other cases (examples 1 and 11), the differences in correlations were numerically trivial.

We wish to thank Dr Don MacLaren, Dr Diana Leighton, Miss Julie Greeves, and Mr Steve Winterburn for access and permission to publish their measurement assessment data recorded in the School of Human Sciences at Liverpool John Moores University.

1 Altman D, Bland JM. Measurement in medicine: the analysis of method comparison studies. *Statistician* 1983;**32**:307–17.
2 Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;i:307–10 .
3 Nevill AM, Holder RL. Modelling maximum oxygen uptake: A case study in non-linear regression formulation and comparison. *Journal of the Royal Statistical Society (Series C)* 1994;**43**:653–66.
4 Jolicoeur P, Heusner AA. The allometry equation in the analysis of the standard oxygen consumption and body weight of the white rat. *Biometrics* 1971;**27**:841–55.
5 Nevill AM, Ramsbottom R, Williams C. Scaling physiological measurements for individuals of different body size. *Eur J Appl Physiol* 1992;**65**:110–17.
6 Bland M. *An introduction to medical statistics*. Oxford: Oxford University Press, 1995: 269, 272–3.
7 Minitab Inc. *MINITAB reference manual*. State College, PA: Minitab Inc, 1995.
8 Atkinson G. A comparison of statistical methods for assessing measurement repeatability in ergonomics research. In: Atkinson G, Reilly T, eds. *Sport, leisure and ergonomics*. London: E and FN Spon, 1995: 218–22.