ORIGINAL ARTICLE

The repeatability and criterion related validity of the 20 m multistage fitness test as a predictor of maximal oxygen uptake in active young men

S-M Cooper, J S Baker, R J Tong, E Roberts, M Hanford

Br J Sports Med 2005;**39**:e19 (http://www.bjsportmed.com/cgi/content/full/39/4/e19). doi: 10.1136/bjsm.2004.013078

See end of article for authors' affiliations

Correspondence to: S-M Cooper, University of Wales Institute Cardiff, School of Sport, PE and Recreation, Cyncoed Campus, Cyncoed Road, Cyncoed, Cardiff CF 23 6XD, UK; smcooper@uwic.ac.uk

Accepted 8 June 2004

Objective: To investigate the repeatability and criterion related validity of the 20 m multistage fitness test (MFT) for predicting maximal oxygen uptake (Vo_{2max}) in active young men.

Methods: Data were gathered from two phases using 30 subjects ($\bar{x} \pm s$; age = 21.8 ± 3.6 years, mass = 76.9 ± 10.7 kg, stature = 1.76 ± 0.05 m). MFT repeatability was investigated in phase 1 where 21 subjects performed the test twice. The MFT criterion validity to predict Vo_{2max} was investigated in phase 2 where 30 subjects performed a continuous incremental laboratory test to volitional exhaustion to determine Vo_{2max} and the MFT.

Results: Phase 1 showed non-significant bias between the two applications of the MFT ($\bar{x}_{diff} \pm s_{diff} = -0.4 \pm 1.4$ ml kg^{-1} min^{-1}; $t = -1.37$, $p = 0.190$) with 95% limits of agreement (LoA) ±2.7 ml kg^{-1} min^{-1} and heteroscedasticity 0.223 ($p = 0.330$). Log transformation of these data reduced heteroscedasticity to 0.056 ($p = 0.808$) with bias −0.007 ± 0.025 ($t = -1.35$, $p = 0.190$) and LoA ± 0.049. Antilogs gave a mean bias on the ratio scale of 0.993 and random error (ratio limits) ×/÷1.050. Phase 2 showed that the MFT significantly underpredicted Vo_{2max} ($\bar{x}_{diff} \pm s_{diff} = 1.8 \pm 3.2$ ml kg^{-1} min^{-1}; $t = 3.10$, $p = 0.004$). LoA were ±6.3 ml kg^{-1} min^{-1} and heteroscedasticity 0.084 ($p = 0.658$). Log transformation reduced heteroscedasticity to −0.045 ($p = 0.814$) with LoA ± 0.110. The significant systematic bias was not eliminated ($\bar{x}_{diff} \pm s_{diff} = 0.033 \pm 0.056$; $t = 3.20$, $p = 0.003$). Antilogs gave a mean bias of 1.034 with random error ×/÷1.116.

Conclusions: These findings lend support to previous investigations of the MFT by identifying that in the population assessed it provides results that are repeatable but it routinely underestimates Vo_{2max} when compared to laboratory determinations. Unlike previous findings, however, these results show that when applying an arguably more appropriate analysis method, the MFT does not provide valid predictions of Vo_{2max}.

It is widely recognised that the most valid physiological indicator of a subject's cardiovascular function is a laboratory determination of maximal oxygen uptake (Vo_{2max}).[1] Such determinations require the use of sophisticated technical equipment and are expensive in terms of the financial cost of this equipment, the training of assessors, the time that it takes to make each estimate of Vo_{2max}, and the accurate analyses of expired gases. As a consequence, exercise scientists have continued to pursue the idea of estimating Vo_{2max} from maximal or sub-maximal tests conducted in non-laboratory environments, via walking protocols[2–4], cycling protocols[5 6], and running protocols.[7–9]

The most common field test for the prediction of Vo_{2max} is the 20 m multistage fitness test (MFT). Originally developed for adults by Léger and Lambert[8] and modified later for children, by reducing the stages from 2 min to 1 min, by Léger *et al*[10], it aims to simulate a continuous incremental exercise test to volitional exhaustion. The MFT was included in the *Eurofit provisional handbook*[11], and after subsequent developmental work by Ramsbottom *et al*,[9] it has been marketed commercially in the form of an audiocassette tape, or CD diskette, and accompanying instruction booklet that includes a table for the conversion of MFT performances into predicted Vo_{2max}.[12] The test is widely used by sports scientists, teachers, coaches, and fitness advisors because it requires limited equipment, is relatively easy to administer, and is suitable for the assessment of large numbers of subjects.

As is the case with all tests and measurements used to assess the components of physical fitness, critical questions must be asked concerning the repeatability and validity of the MFT. A number of studies have been conducted, each of which purport to have investigated the repeatability and/or validity of the MFT when used with children or adolescents[10 13–17] and with adults.[8 9 18 19] In the case of each of these previously published investigations, the authors have used analytical methods such as Pearson's inter-class correlation coefficient and hypothesis tests such as the dependent (paired) *t* test or repeated measures ANOVA as indices of the MFT's repeatability and/or validity. Bland and Altman,[20] and more recently Nevill and Atkinson,[21] have criticised the reliance upon these methods, in particular the correlation coefficients, as being primarily indicants of relationship rather than of agreement. In the estimation of both test repeatability and test validity, the preferred analysis of choice should be the 95% limits of agreement (LoA) method introduced by Bland and Altman in 1986.[22]

The aim of the present study was therefore twofold: (i) to examine the repeatability of the MFT, as described by Brewer *et al*,[12] by applying 95% LoA to predicted Vo_{2max} gathered from repeat applications of the test, and (ii) to consider the criterion related validity of the MFT by calculating the 95% LoA between predicted Vo_{2max} and Vo_{2max} measured directly in a laboratory, in a group of active young men.

Abbreviations: LoA, limits of agreement; MFT, 20 m multistage fitness test

METHODS

Subjects

Measurements were made on 30 male undergraduates ($\bar{x} \pm s$; age = 21.8±3.6 years, body mass = 76.9±10.7 kg, and stature = 1.76±0.05 m) who were all pursuing sports studies degrees at a British university. Before data collection began the relevant University Research Ethics Sub-Committee approved both phases of the proposed study, and all participants gave written informed consent and volunteered to act as subjects. Each was also screened to verify that he was a non-smoker and was not suffering from an injury. None had any history of cardiovascular disease or other health risks, and none were taking medication known to influence oxygen uptake.

Data collection procedures

The first phase of the study aimed to establish the repeatability of the MFT in a group of 21 subjects drawn randomly from the 30 volunteers. Each subject performed the MFT twice, with a minimum of 7 days and a maximum of 14 days elapsing between the test and the retest. In phase 2, the 30 subjects performed both the MFT and a laboratory assessment to determine Vo_{2max}. Each assessment was performed randomly on separate days with a minimum of 7 days and a maximum of 14 days elapsing between assessments. All subjects were fully familiarised with both measurement protocols before data collection. In order to avoid the affects of diurnal variations, data were collected from the subjects in both phases of the study at approximately the same time of day. Because of the difficulties involved in ensuring compliance, no attempt was made to control the diet of the subjects (this included their consumption of alcohol) nor was an attempt made to control the pre-testing exercise condition of the subjects.

Laboratory determined Vo_{2max}

Maximal oxygen uptake was defined as the maximum rate at which a subject could take up and utilise oxygen while breathing air at sea level (Bird and Davidson[23], page 64) and was determined during a continuous incremental exercise test to volitional exhaustion while running on a motorised treadmill (Ergo ELG2, Woodway, Weil am Rhein, Germany). Each assessment was preceded by a standardised 5 min warm up on the treadmill where subjects ran at a speed of 2.22 m s^{-1} and zero (0%) gradient. Subjects began the test by running at a speed of 3.06 m s^{-1} and 0% gradient, after which the inclination of the treadmill was increased by 2.5° every 3 min. This increase in treadmill inclination continued until the subject indicated that he could run no further. During the last minute of each 3 min exercise period, expired air was channelled into pre-emptied 150 l Douglas bags via a two way low resistance respiratory valve (Hydraulic Transmission Services, Salford, UK) with 80 ml dead space and a short length of 32 mm bore respiratory tubing. Towards the end of the Vo_{2max} assessment, a sample of expired air was collected when the subject indicated that he could continue for only one more minute. All subjects were verbally encouraged to perform maximally throughout the assessment. After being assessed, all subjects participated in a 5 min cool down that included prescribed jogging and stretching.

Subsequently, each Douglas bag was analysed for volume, using a dry gas meter (Harvard Apparatus, Edenbridge Kent, UK), oxygen consumption, and carbon dioxide production in order to determine oxygen uptake. Oxygen and carbon dioxide concentrations were obtained from a Servomex 1440C dual gas analyser (Servomex International, Crowborough, UK) that was calibrated before each assessment using gases of known concentration. In deciding whether individual subjects had achieved Vo_{2max}, three of the criteria provided by the British Association of Sport and Exercise Sciences were used: (i) subjective fatigue and volitional exhaustion, (ii) a plateau in the oxygen uptake/exercise intensity relationship, and (iii) a final respiratory exchange ratio of 1.15 or above (Bird and Davidson,[23] page 64).

Maximal oxygen uptake was expressed relative to body mass for each subject. Relative performance was derived using the ratio standard where Vo_{2max} (ml min^{-1}) was divided by body mass (ml kg^{-1} min^{-1}). It is fully acknowledged that this method of scaling these data might be considered inappropriate, and further that allometric modelling of these data might be more appropriate in partitioning out differences in body size.[24] In order for the requisite comparisons with data gathered from performance on the MFT to be made, however, it was considered that this was the necessary approach to take.

Multistage fitness test (MFT)

The protocol for the MFT was identical to that described by Brewer *et al*.[12] Briefly, this consisted of shuttle running between two parallel lines set 20 m apart, running speed cues being indicated by signals emitted from a commercially available pre-recorded audiocassette tape. The audiocassette tape dictated that subjects started running at an initial speed of 2.36 m s^{-1} and that running speed increased by 0.14 m s^{-1} each minute. This increase in running speed is described as a change in test level.[9] The speed of the cassette player was checked for accuracy in accordance with the manufacturer's instructions before each application. All subjects performed a 10 min warm up that included prescribed jogging and stretching. The MFT was conducted in a gymnasium with sprung wooden flooring where subjects ran in groups of five in order to add an element of competition and to aid maximal effort. All were verbally encouraged to perform maximally during each assessment. After finishing the MFT, all subjects participated in a 5 min cool down that also included prescribed jogging and stretching. MFT results for each subject were expressed as a predicted Vo_{2max} (ml kg^{-1} min^{-1}) obtained by cross-referencing the final level and shuttle number (completed) at which the subject volitionally exhausted with that of the Vo_{2max} table provided in the instruction booklet accompanying the MFT. Only fully completed 20 m shuttle runs were considered.

Statistical analyses

The normality of appropriate data sets (that is, residual errors) was confirmed via the Anderson-Darling normality test.[25] It was considered appropriate therefore to test stated hypotheses using parametric statistical techniques. A maximum a priori α level of 0.05 was applied throughout.

In phase 1 of the study the agreement between repeat performances on the MFT (test-retest) was quantified using the 95% LoA method originally described by Bland and Altman.[20] This included plotting a graph (Bland-Altman plot) of the mean for subjects' test and retest results [(test+retest)/2] on the x axis corresponding to the difference between each subject's test and retest results (test−retest) on the y axis. To investigate systematic bias, a dependent *t* test was conducted to test the hypothesis of no difference between the sample mean score for the test versus the sample mean score for the retest. Provided the differences between subjects' test and retest scores (residual errors) were normally distributed, the 95% LoA (indicative of random error) were expressed as ±1.96 multiplied by the standard deviation of the residual errors (that is, $\pm 1.96 \times s_{diff}$). When the systematic bias is not statistically significant, there is a rationale for expressing the

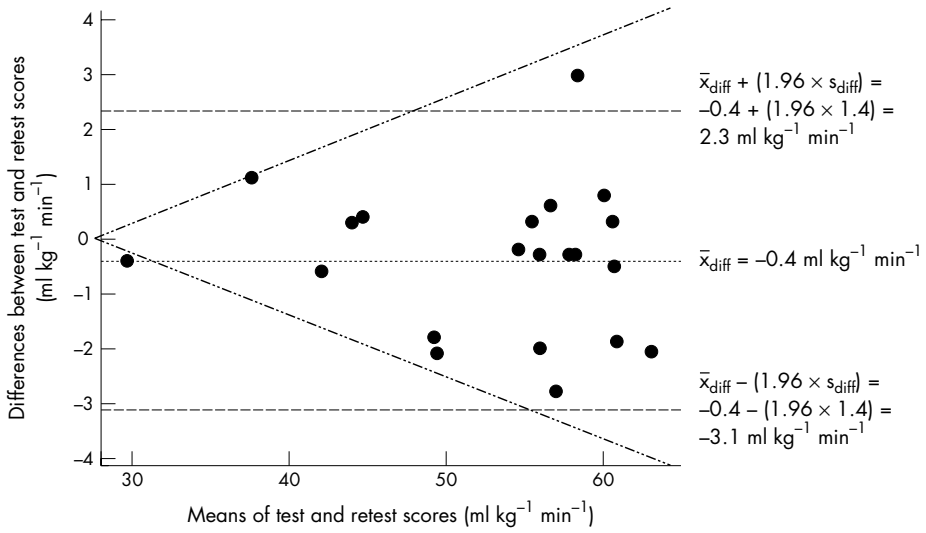

Figure 1 Bland-Altman plot summarising the results from phase 1.

95% LoA as ± the value of this bias, thus, $\bar{x}_{diff} \pm (1.96 \times s_{diff})$. In which case the results could therefore be described in the actual units of measurement.[26]

Heteroscedasticity occurs in test data when the amount of random error increases as the measured values increase.[26] Heteroscedasticity was investigated in the present study by calculating the zero order correlation coefficient (heteroscedasticity coefficient) between the means of subjects' test and retest scores (indicative of the size of measured values) and the absolute differences between subjects' test and retest scores (indicative of random error). Bland and Altman[20] originally proposed that the solution to establishing a positive, statistically significant heteroscedasticity coefficient ($p<0.05$) was to transform the original test data into natural logarithms and then to repeat the limits of agreement methods described above with these log transformed data. Subsequently, Nevill and Atkinson[21] have suggested that if the correlation between absolute residual errors and individual means is positive, but not necessarily statistically significant, there is some benefit in reducing heteroscedasticity by transforming test data into natural logarithms and recalculating the limits of agreement. This suggestion was followed in the present study so that when antilogs of these results were taken, the outcomes could be expressed as the mean bias ×/÷ by the 95% agreement component (random error) on the ratio scale.

In phase 2, the criterion related validity of the MFT was investigated by quantifying the agreement between subjects' laboratory determined Vo_{2max} and their predicted Vo_{2max} from performing the MFT. Both laboratory determined Vo_{2max} and MFT predicted Vo_{2max} data were exposed to exactly the same diagnostic statistical tests as those described for calculating the 95% LoA for the data collected in phase 1 of the study.

RESULTS

Phase 1 test-retest repeatability of MFT scores (table 1 and fig 1)

Two administrations (test-retest) of the MFT were performed by a group of 21 subjects ($\bar{x} \pm s$; age = 22.1 ± 3.9 years, body mass = 77.1 ± 8.4 kg, stature = 1.78 ± 0.05 m). The mean MFT performance for the test was 52.9 ± 8.8 ml kg^{-1} min^{-1}, and for the retest it was 53.3 ± 8.9 ml kg^{-1} min^{-1}. The dependent *t* test conducted to test the hypothesis of equality of means showed no significant bias. The residual errors between the test and the retest were normally distributed and the bias ± the 95% LoA was −0.4 ± 2.7 ml kg^{-1} min^{-1}.

Figure 1 shows that there is some evidence of heteroscedasticity present in these data. While the computed heteroscedasticity coefficient was not statistically significant, however, it was positive ($r = 0.223$, $p = 0.330$). Transformation of the test and retest data into natural logarithms

Table 1 Performance characteristics and analysis summary from phase 1: repeatability of MFT scores (n = 21)

Variable	Units	$\bar{x} \pm s$	t Ratio	p Value
Test				
20 m MFT Vo_{2max}	ml kg^{-1} min^{-1}	52.9 ± 8.8		
Retest				
20 m MFT Vo_{2max}	ml kg^{-1} min^{-1}	53.3 ± 8.9		
Differences*†	ml kg^{-1} min^{-1}	−0.4 ± 1.4	−1.37	0.190
Test				
20 m MFT Vo_{2max}	Logarithms¶	3.95 ± 0.19		
Retest				
20 m MFT Vo_{2max}	Logarithms¶	3.96 ± 0.19		
Differences‡§	Logarithms¶	−0.007 ± 0.025	−1.35	0.190

*Heteroscedasticity coefficient, $r = 0.223$ ($p = 0.330$); †data are normally distributed (Anderson-Darling test, $p = 0.116$); ‡heteroscedasticity coefficient, $r = 0.056$ ($p = 0.808$); §data are normally distributed (Anderson-Darling test, $p = 0.388$); ¶natural logarithms.
$t_{20}(0.05) = 2.086$ (two tailed test).

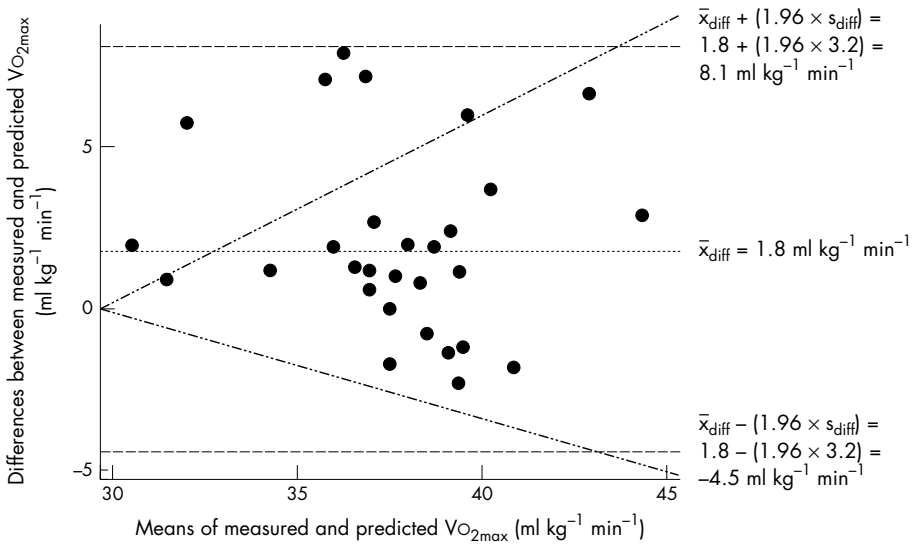

Figure 2 Bland-Altman plot summarising the results from phase 2.

reduced the heteroscedasticity to $r = 0.056$ ($p = 0.808$). The dependent t test performed between the log transformed mean score for the test (3.95 ± 0.19) and the log transformed mean score for the retest (3.96 ± 0.19) showed no significant bias. Residual errors between test and retest log transformed data were normally distributed. The mean difference $\pm$ the 95% LoA was -0.007 ± 0.049. Taking antilogs of these values gave a mean bias of 0.993 with a random error component of $\times/\div 1.050$.

Phase 2 criterion related validity of the MFT (table 2 and fig 2)

A total of 30 subjects (age = 21.8 ± 3.6 years, body mass = 76.9 ± 10.7 kg, stature = 1.76 ± 0.05 m) performed a laboratory test to determine Vo_{2max} and the MFT from which Vo_{2max} was predicted. Table 2 shows that the mean laboratory determined Vo_{2max} was 57.5 ± 4.5 ml kg^{-1} min^{-1} and the mean predicted Vo_{2max} from performing the MFT was 55.7 ± 5.0 ml kg^{-1} min^{-1}. Residual errors were normally distributed and the mean bias (1.8 ml kg^{-1} min^{-1}) was statistically significant ($t_{29} = 3.10$, $p = 0.004$). The 95% LoA were ± 6.3 ml kg^{-1} min^{-1}.

Figure 2 shows that there was very little evidence of heteroscedasticity present in these data. However, the computed coefficient was positive ($r = 0.084$, $p = 0.658$). Data were therefore transformed into natural logarithms and the 95% LoA method repeated. This reduced heteroscedasticity to $r = -0.045$ ($p = 0.814$). The dependent t test performed between the mean log transformed score for laboratory determined Vo_{2max} (4.05 ± 0.08) and the mean log transformed score for the MFT predicted Vo_{2max} (4.02 ± 0.09) continued to show a significant systematic bias ($\bar{x}_{diff} = 0.033$; $t_{29} = 3.20$, $p = 0.003$). Residual errors were normally distributed and the 95% ratio limits of agreement were ± 0.110. Taking antilogs of these values gave a mean bias of 1.034 with a random error component of $\times/\div 1.116$.

DISCUSSION

It was not possible to compare the results of the present MFT repeatability 95% LoA directly, as none were available in the current literature. In terms of the statistics that could be compared however, we computed, post hoc, the zero order correlation between the test and the retest results from phase 1 and found there to be a high and statistically significant linear relationship ($r = 0.988$, $p = 0.0005$). In addition, there was no significant difference between the mean scores for the test and the retest ($\bar{x}_{diff} = -0.4$ ml kg^{-1} min^{-1}; $t_{20} = -1.37$, $p = 0.190$). These results are similar to those reported by

Table 2 Performance characteristics and analysis summary from phase 2: criterion related validity of the MFT (n = 30)

Variable	Units	$\bar{x} \pm s$	t Ratio	p Value
Measured Vo_{2max}	ml kg^{-1} min^{-1}	57.5 ± 4.5		
Predicted 20 m MFT Vo_{2max}	ml kg^{-1} min^{-1}	55.7 ± 5.0		
Differences*†	ml kg^{-1} min^{-1}	1.8 ± 3.2	3.10	0.004
Measured Vo_{2max}	Logarithms¶	4.05 ± 0.08		
Predicted 20 m MFT Vo_{2max}	Logarithms¶	4.02 ± 0.09		
Differences‡§	Logarithms¶	0.033 ± 0.056	3.20	0.003

*Heteroscedasticity coefficient, $r = 0.084$ ($p = 0.658$); †data are normally distributed (Anderson-Darling test, $p = 0.079$); ‡heteroscedasticity coefficient, $r = -0.045$ ($p = 0.814$); §data are normally distributed (Anderson-Darling test, $p = 0.057$); ¶natural logarithms.
$t_{29}(0.05) = 2.045$ (two tailed test).

Léger *et al*[10] in their original study, where subjects also ran to volitional exhaustion during a 20 m multistage shuttle run test ($r = 0.95$, $p<0.01$ and unspecified t, $p>0.05$).

The test sample used by Léger *et al*[10] consisted of 81 men and women whose ages ranged from 20 to 45 years, and who were in varying states of physical condition. In contrast, the sample used in phase 1 were active male undergraduate students, all of a similar age, physical condition, and training status. The strength of the test-retest correlation in the present data is surprising therefore as normally the calculation of the numerical value of the coefficient is highly influenced by the range of the characteristic being analysed, that is data heterogeneity.[27–29] Indeed, this observation is often cited as one of the major weaknesses of the correlation coefficient as a measure of repeatability.[20 22] The strong test-retest correlation in the present data might have been due to the fact that all of the subjects who participated in the study were sports studies students, all of whom were used to performing the MFT as part of their programmes of study and were therefore also better able to gauge the intensity of their performances.

In considering the approach to the design of phase 1, it was intended to identify the stability reliability[30] of the MFT. Regardless of the source of the error, however, there are two components of variability associated with the assessment of measurement error—systematic bias and random error—that need to be considered in detail.[26] Inspection of the Bland-Altman plot presented as fig 1 provides a visual indication of both systematic bias and random error in the raw data. It can be seen from both the direction and the size of the raw data scatter around the zero line (y axis) that there is evidence of a slight tendency towards a negative bias as well as random variation in these data. From fig 1 there is also visual evidence to suggest that these raw data show some evidence of heteroscedasticity. Natural log transformation of the test and retest raw data reduced the heteroscedasticity coefficient and gave a mean bias ± the 95% LoA of -0.007 ± 0.049. Taking antilogs resulted in a mean bias on the ratio scale of 0.993 and an agreement (random error) component of $\times/\div 1.050$. That is, 95% of the ratios for the sample (log transformed test score divided by log transformed retest score) should be contained between the values 0.946 ($0.993 \div 1.050$) and 1.043 (0.993×1.050). In fact, in the present data, 100% of the ratios for the 21 subjects assessed were contained between these two values. For any new individual from the studied population therefore, assuming the bias present (0.007%) to be negligible, any two tests would differ due to measurement error by no more than 5% in a positive or negative direction.[21] It is interesting to note that this latter result is very similar to the 95% coefficient of variation of 5.2% calculated for the original (non-transformed) data in the arguably simpler manner $[100 \times ((1.96 \times s_{diff})/\text{grand } \bar{x})]$ identified by Bland.[31]

These ratio limits of agreement are not common indices in the sport and exercise sciences. To put them into some practical context therefore, if a new subject from the studied population presented with an estimated MFT performance of 30 ml kg^{-1} min^{-1} on the first application of the test, the worse case scenario (a 95% probability) is that this subject on the second occasion could score an estimated score as low as $30 \times 0.946 = 28.4$ ml kg^{-1} min^{-1}, or as high as $30 \times 1.043 = 31.3$ ml kg^{-1} min^{-1}. Most sports scientists would probably consider these limits of agreement to be acceptable. However, for a subject with a higher estimated performance on the test of, for instance, 70 ml kg^{-1} min^{-1}, there is a 95% probability that their retest performance might be as low as $70 \times 0.946 = 66.2$ ml kg^{-1} min^{-1} or as high as $70 \times 1.043 = 73.0$ ml kg^{-1} min^{-1}. These ratio limits of agreement might vary in absolute terms, but they remain a constant ratio in performance from the test to the retest. While these scores are probably acceptable for the repeatability of a field test of one of the physiological aspects of physical fitness, they are also more realistic in the manner in which they are allowed to vary depending upon the standards of performance of the subjects.[21]

The term calibration refers to the development of a model that facilitates the prediction of measured criterion values from related predictor values (Atkinson and Nevill,[32] page 812). In the development of useful calibration models the regression equation developed on one sample of the chosen population should be cross-validated against results provided by another equivalent sample. Without cross-validation to test the accuracy of the prediction, results will always be suspect.[22 26 33] Indeed, Atkinson and Nevill[26] believe that many of the most commonly used field tests of physiological fitness that provide tables for the prediction of the directly measured physiological parameter from indirect measures lack this key element of validity. The MFT is a prime example of such a test, and the design of phase 2, and the manner in which the resultant data were analysed using the 95% LoA method, was an attempt to address this issue directly.

In order to develop the Vo_{2max} table found in the booklet that accompanies the MFT, Brewer *et al*[12] used linear regression methods on the data of Ramsbottom *et al*[9] to produce a calibration model that predicted Vo_{2max} from MFT performances expressed as maximum level and shuttle number achieved. Regrettably the authors' of those studies available in the literature that have investigated the validity of this calibration model have reported their results in terms of correlation coefficients and/or hypothesis tests rather than applying limits of agreement to measured and predicted data gathered from equivalent samples. Consequently, it was not possible to compare the 95% LoA results from phase 2 directly, as none were currently available in the literature.

Out of interest therefore, we computed, post hoc, the magnitude of the zero order correlation between the predicted Vo_{2max} from the MFT and the laboratory determined Vo_{2max}. Although this correlation was statistically significant ($r = 0.785$, $p = 0.0005$) it was disappointingly low when compared to others available in the literature. For example, McNaughton *et al*[19] have reported that for 32 male undergraduates, the correlation coefficient between MFT predicted Vo_{2max} and a laboratory determined Vo_{2max} was far stronger than that forthcoming from the present data ($r = 0.82$, $p<0.05$). Indeed, in the original validation study of the MFT[9] from which Brewer *et al*[12] subsequently developed the version of the MFT used in our study, the correlation between the shuttle run test and laboratory determined Vo_{2max} for 36 males was also $r = 0.82$ ($p<0.01$).

The Bland-Altman plot presented as fig 2 provides a visual indication of both the systematic bias and the random error between MFT predicted Vo_{2max} and laboratory determined Vo_{2max} in the raw data drawn from the present sample. From both the direction and the size of the scatter of these data around the zero line (y axis) there is evidence of a substantial positive systematic bias. Additionally, there seems to be limited random variation in these data. Atkinson and Nevill[26] have shown that a significant difference between means is more likely when there is limited random variation amongst the raw scores, and vice versa.

The statistical analyses conducted on these data as part of the limits of agreement method confirmed the situations relating to both systematic bias and random error. The mean of the residual errors between laboratory determined Vo_{2max} and MFT predicted Vo_{2max} was statistically significant ($\bar{x}_{diff} \pm s_{diff} = 1.8 \pm 3.2$ ml kg^{-1} min^{-1}; $t_{29} = 3.10$, $p = 0.004$). This resulted from the mean MFT predicted Vo_{2max} being 3.1% below that for laboratory determined Vo_{2max}. This result

is similar to that reported by McNaughton *et al*[19] where the mean ($\pm s_{\bar{x}}$) Vo_{2max} predicted from the MFT (58.1 ± 4.9 ml kg^{-1} min^{-1}) was 3% lower than that for a laboratory determination (59.7 ± 5.9 ml kg^{-1} min^{-1}). This difference was not reported as being statistically significant ($p>0.05$). It is also interesting to report the similarity between the present results and those reported originally by Ramsbottom *et al*[9] with respect to such differences. It is unfortunate that Ramsbottom *et al*[9] did not report the results of a hypothesis test of equality of means that would have quantified the systematic bias between measured and predicted values, but the mean ($n = 36$, males) MFT predicted Vo_{2max} (55.4 ml kg^{-1} min^{-1}) was 5.2% lower than that recorded for the laboratory determination (58.5 ml kg^{-1} min^{-1}).

Even though it showed statistical significance, most exercise physiologists would probably consider that a mean difference between measured and predicted Vo_{2max} in the order of 1.8 ml kg^{-1} min^{-1} would not be significant from a practical perspective. Considering the criticisms levelled at hypothesis tests when used as the sole method in the assessment of test validity in the literature,[26] we decided to interrogate our data further (post hoc) in an attempt to identify the practical significance of this bias. Cohen[34] considers the effect size to be a reasonable index of the meaningfulness of a statistical outcome. In the present study the effect size index (d) for the *t* test for means was computed: $d = [|\bar{x}_1 - x_2| s_P]$, where: $\bar{x}_1$ is the sample mean for the laboratory measured Vo_{2max}, $\bar{x}_2$ is the sample mean for the MFT predicted Vo_{2max}, and s_p is the pooled standard deviation $= \sqrt{[((s_1^2(n_1-1))+(s_2^2(n_2-1)))/(n_1+n_2-2)]}$. Here s_1^2 and n_1 are, respectively, the sample variance and the sample number for the laboratory measured Vo_{2max}, and s_2^2 and n_2 are the sample variance and the sample number for the MFT predicted Vo_{2max}. In the present data $d = 0.4$ which is described by Cohen[34] (page 40) as only a small to medium sized difference. Indeed, the statistical power of this analysis in rejection of the null hypothesis of equality of means in the population from which this sample of subjects was drawn was only 33%.

When measurements are made in the sport and exercise sciences there are often multiple sources of error. While we attempted to account for many sources of error in our research design, we can speculate that a mean under-prediction in Vo_{2max} by the MFT when compared to a laboratory determination of the magnitude 1.8 ml kg^{-1} min^{-1} might well have been due to the error inherent in a different gas analysis system being used in the present study to that used by Ramsbottom *et al*[9] in their study. Unfortunately, Ramsbottom *et al*[9] do not identify the gas analysis system that they used. Consequently, we could not perform a study to compare the Servomex analysis system that was used in the present research with that used by Ramsbottom *et al*.[9]

It is clear from the Bland-Altman plot (fig 2) generated from the phase 2 data that there is no substantial increase in variability in these scores as the size of the measured values increases. The statistical examination of heteroscedasticity resulted in a coefficient of $r = 0.084$ ($p = 0.658$). If confirmation of the presence of heteroscedasticity in these data was based solely on the size of the coefficient therefore, it can be concluded that the assumption that the limits of agreement remain constant throughout the range of measurements can be accepted.[20] Even though the heteroscedasticity coefficient was close to zero, it was still positive. Consequently the raw data were transformed into natural logarithms and the limits of agreement method was applied to these transformed scores.

Log transformation reduced heteroscedasticity to $r = -0.045$ ($p = 0.814$) but it did not improve the normality of the distribution of residual errors between laboratory determined Vo_{2max} and MFT predicted Vo_{2max} ($p = 0.057$). Once again, the mean difference between these two data sets was found to be statistically significant ($\bar{x}_{diff} \pm s_{diff} = 0.033 \pm 0.056$; $t_{29} = 3.20$, $p = 0.003$) and the 95% LoA were ± 0.110. Taking antilogs of these values gave a mean bias on the ratio scale of 1.034 and a random error component of $\times/\div 1.116$.

In terms of the ratio limits of agreement the 3.3% bias present (the 0.2% difference between this logarithmic value and that calculated from the raw data (3.1%) is probably due to rounding errors) cannot be considered to be negligible. The two methods of determining Vo_{2max} differ due to measurement error by a substantial 11.6% in a positive or negative direction. Interestingly, when Bland's[31] calculation was applied to the original data before log transformation, it gave a 95% coefficient of variation of 11.1%. Indeed, 95% of the ratios for the sample (log transformed laboratory determined Vo_{2max} divided by log transformed MFT prediction of Vo_{2max}) should be contained between the limits 0.927 ($1.034 \div 1.116$) and 1.154 (1.034×1.116). In the present data, 100% of the ratios for the 30 subjects assessed were actually contained between these two values.

To help interpret these ratio limits of agreement: if a new subject from the studied population presented with a laboratory determined Vo_{2max} of 30 ml kg^{-1} min^{-1}, there is a 95% probability that their predicted performance from the MFT calibration model could be as low as $30 \times 0.927 = 27.8$ ml kg^{-1} min^{-1} or as high as $30 \times 1.154 = 34.6$ ml kg^{-1} min^{-1}. For a subject with a higher laboratory determined performance of 70 ml kg^{-1} min^{-1} the prediction from the MFT calibration model could result (a 95% probability) in a score as low as $70 \times 0.927 = 64.9$ ml kg^{-1} min^{-1} or as high as $70 \times 1.154 = 80.8$ ml kg^{-1} min^{-1}. We consider these ratio limits of agreement to be more realistic in the way that they are allowed to vary depending upon the levels of subjects' performances. Considering that the MFT is a field test, the ratio limits for the lower performing subject are probably just on the border of acceptability, while the ratio limits for the higher performer are too wide to be acceptable for most sports scientists. As has previously been stated, however, the fact that a significant systematic bias was identified in these data indicates that the MFT cannot be considered as a valid predictor of laboratory determined Vo_{2max} in male undergraduates, regardless of the calculated limits of agreement.

CONCLUSIONS

From these results it was possible to conclude that the calculated bias and 95% LoA are narrow enough for the MFT to be considered repeatable when used with active male undergraduates. However, while the MFT might prove useful in predicting the more substantial effect that might accompany aerobic training conducted by a less well trained subject, there is some doubt as to whether the test is sensitive enough to monitor the small changes in performance that might accompany the improved training status of a subject who already has a highly developed aerobic fitness.

These findings also lend support to previous validations of the MFT by identifying that it routinely underestimates Vo_{2max} when compared to laboratory determinations. Unlike previous findings, however, these results also show that when applying an arguably more appropriate analysis method (95% LoA), the MFT does not provide valid predictions of Vo_{2max}. The results of the cross-validation of the calibration model developed by Brewer *et al*[12] which provided the Vo_{2max} table that accompanies the commercially available MFT, showed a significant systematic bias in

What is already known on this topic

The most common field test for the prediction of Vo_{2max} is the 20 m multistage fitness test (MFT). However, critical questions must be asked concerning the repeatability and validity of the MFT.

What this study adds

While the MFT is a well established and ubiquitous field test of cardiovascular function, the results of this study show that it is not a valid test for the accurate prediction of Vo_{2max} in active young men.

underestimating Vo_{2max} when compared to a laboratory determined assessment. While the MFT is a well established and ubiquitous field test of cardiovascular function, these results show that it is not a valid test for the accurate prediction of Vo_{2max} in active male undergraduates at least. Additionally, and arguably more importantly, these findings highlight the need for sport and exercise scientists to appraise the repeatability and validity of frequently used measurement protocols by applying more appropriate statistical methods.

Authors' affiliations
S-M Cooper, R J Tong, E Roberts, M Hanford, University of Wales Institute Cardiff, Cardiff, UK
J S Baker, University of Glamorgan, Pontypridd, UK

Competing interests: none declared

REFERENCES

1 **Armstrong N**. A critique of fitness testing. In: Biddle S, ed. *Foundations of health-related fitness in physical education*. London: Ling, 1987:136–8.
2 **Hermiston R**, Faulkner J. Prediction of maximal oxygen uptake by a step-wise regression technique. *J Appl Physiol* 1971;**30**:833–7.
3 **Kline G**, Porcari J, Hintermaster R, *et al*. Estimation of VO_{2max} from a one-mile track walk, gender, age and body weight. *Med Sci Sport Exerc* 1987;**19**:253–9.
4 **Ebbeling C**, Ward A, Puleo A. Development of a single-stage submaximal treadmill walk test. *Med Sci Sport Exerc* 1991;**23**:966–73.
5 **Åstrand P-O**, Rhyming I. A nomogram for calculation of aerobic capacity from pulse rate during submaximal work. *J Appl Physiol* 1954;**7**:32.
6 **Sinconolfi S**, Cullinane E, Carleton R, *et al*. Assessing VO_{2max} in epidemiological studies: modification of the Åstrand-Rhyming test. *Med Sci Sports Exerc* 1982;**14**:335–8.
7 **Burke E**. Validity of selected laboratory and field tests of working capacity. *Res Q* 1976;**47**:95–104.
8 **Léger L**, Lambert J. A maximal multistage 20 m shuttle run test to predict VO_{2max}. *Eur J Appl Physiol Occup Physiol* 1982;**49**:1–12.
9 **Ramsbottom R**, Brewer J, Williams C. A progressive shuttle run test to estimate maximal oxygen uptake. *Br J Sports Med* 1988;**22**(4):141–4.
10 **Léger L**, Mercier D, Gadoury C, *et al*. The multistage 20 metre shuttle run test for aerobic fitness. *J Sports Sci* 1988;**6**:93–101.
11 **Council of Europe**. *Eurofit provisional handbook: testing physical fitness*. London: HMSO, 1983.
12 **Brewer J**, Ramsbottom R, Williams C. *Multistage fitness test*. Leeds: National Coaching Foundation, 1988.
13 **van Mechelen W**, Hlobil H, Kemper H. Validation of two running tests as estimates of maximal aerobic power in children. *Eur J Appl Physiol* 1986;**55**:503–6.
14 **Armstrong N**, Williams J, Ringham D. Peak oxygen uptake and progressive shuttle run performance in boys aged 11–14 years. *Br J Phys Educ* 1988;**19**(4):10–11.
15 **Boreham C**, Paliczka V, Nichols A. A comparison of the PWC_{170} and the 20-MST tests of aerobic fitness in adolescent children. *J Sports Med Phys Fitness* 1990;**30**:19–23.
16 **Liu N**, Plowman S, Looney M. The reliability and validity of the 20-meter shuttle test in American students 12 to 15 years old. *Res Q Exerc Sport* 1992;**63**(4):360–5.
17 **McVeigh S**, Payne A, Scott S. The reliability and validity of the 20-meter shuttle test as a predictor of peak oxygen uptake in Edinburgh school children, age 13 to 14 years. *Pediatr Exerc Sci* 1995;**7**:69–79.
18 **Palickza V**, Nichols A, Boreham C. A multistage shuttle run as a predictor of running performance and maximal oxygen uptake in adults. *Br J Sports Med* 1987;**21**(4):163–4.
19 **McNaughton L**, Hall P, Cooley D. Validation of several methods of estimating maximal oxygen uptake in young men. *Percept Mot Skills* 1998;**87**:575–84.
20 **Bland J**, Altman D. Statistical methods for assessing agreement between two methods of clinical measurements. *Lancet* 1986;**i**:307–10.
21 **Nevill A**, Atkinson G. Assessing agreement between measurements recorded on a ratio scale in sports medicine and sports science. *Br J Sports Med* 1997;**31**:314–18.
22 **Nevill A**. Validity and measurement agreement in sports performance [editorial]. *J Sports Sci* 1996;**14**:199.
23 **Bird S**, Davidson R, eds. *Guidelines for the physiological testing of athletes*, 3rd ed. Leeds, UK: British Association of Sport & Exercise Sciences, 1997.
24 **Winter E**, Nevill A. Scaling: adjusting for differences in body size. In: Eston R, Reilly T, eds. *Kinanthropometry and exercise physiology laboratory manual*. Vol 1. *Anthropometry*, 2nd ed. London: E & FN Spon, 2001:275–93.
25 **Minitab Inc**. *MINITAB reference manual*. State College, PA: Minitab, 1995.
26 **Atkinson G**, Nevill A. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Med* 1998;**26**(4):217–38.
27 **Altman D**, Bland J. Measurement in medicine: the analysis of method comparison studies. *Statistician* 1983;**32**:307–17.
28 **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 & FN Spon, 1995:218–22.
29 **Bates B**, Zhang S, Dufek J. The effects of sample size and variability on the correlation coefficient. *Med Sci Sport Exerc* 1996;**28**(3):386–91.
30 **Baumgartner T**. Norm-referenced measurement: reliability. In: Safrit M, Wood T, eds. *Measurement concepts in physical education and exercise science*. Champaign, IL: Human Kinetics, 1989:45–72.
31 **Bland J**. *An introduction to medical statistics*, 3rd ed. Oxford: Oxford University Press, 2000.
32 **Atkinson G**, Nevill A. Selected issues in the design and analysis of sport performance research. *J Sports Sci* 2001;**19**:811–27.
33 **Vincent W**. *Statistics in kinesiology*, 2nd ed. Champaign, IL: Human Kinetics, 1999.
34 **Cohen, J**. *Statistical power analysis*, 2nd ed. New Jersey: Lawrence Erlbaum Associates, 1988.