NAVAL HEALTH RESEARCH CENTER

MEASUREMENT ERROR IN MAXIMAL OXYGEN UPTAKE TESTS

R. R. Vickers, Jr.

Reportt No. 04-03

NAVAL HEALTH RESEARCH CENTER
P. O. BOX 85122
SAN DIEGO, CA 92186-5122

BUREAU OF MEDICINE AND SURGERY (M2)
2300 E ST. NW
WASHINGTON, DC 20372-5300

Measurement Error in Maximal Oxygen Uptake Tests

Ross R. Vickers, Jr.

Human Performance Department
Naval Health Research Center

Report Number 04-03, supported by the U.S. Army Medical Research and Materiel Command, Ft. Detrick, Frederick, MD, under Work Unit No.60109. The views expressed in this article are those of the authors and do not necessarily reflect the official policy or position of the Department of the Navy, Department of Defense, or the U.S. Government. Approved for public release; distribution unlimited.

This research has been conducted in compliance with all applicable Federal Regulations governing the protection of human subjects in research. No human subjects were directly involved in this research.

SUMMARY

Background

Cardiorespiratory fitness is important for health, work, and athletic performance. Laboratory tests of maximal oxygen uptake (VO_{2max}) are the gold standard for assessing this aspect of fitness. VO_{2max} protocols with small measurement errors will provide the best estimates of relationships between fitness and its antecedents and consequences. For example, tests with smaller errors will provide better indications of how well running tests function as substitutes for laboratory tests.

Objective

Published studies of the reliability of VO_{2max} tests provide an empirical basis for estimating VO_{2max} test precision. This review employed meta-analysis procedures to model VO_{2max} test precision.

Approach

Studies of the test-retest reliability of VO_{2max} protocols were identified from previous reviews and searches of computerized databases for biomedical, behavioral, and sports research. Of 51 studies identified, 12 were dropped because long test-retest intervals made it likely that true VO_{2max} values changed during the study. The reported means, standard deviations, and test-retest correlations were used to compute the standard error of measurement (SEM) for VO_{2max} for the remaining 39 studies. The age and gender composition of the sample were coded along with the exercise mode (treadmill, cycle ergometer, other) and the test-retest interval for the protocol. Meta-analysis produced a predictive model for SEM based on sample and protocol attributes.

Results

Average SEM was 2.58 $ml \cdot kg^{-1} \cdot min^{-1}$. SEM was higher in samples with higher average VO_{2max}. Age, gender, test interval, and exercise mode were not related to SEM. After allowing for outliers, the final model to predict SEM was $\ln(SEM) = 0.661 + (.006 * VO_{2max})$.

Conclusions

SEM increases as the average VO_{2max} of the sample increases. Other population and protocol attributes were not related to SEM. The potential applications of the model for SEM include evaluating new VO_{2max} protocols, evaluating field tests (e.g., run tests, walk tests), and making allowances for measurement error when investigating the relationships of VO_{2max} with other variables.

Introduction

Cardiorespiratory fitness is important for health and work and athletic performance. Maximal oxygen uptake (VO_{2max}) is the accepted indicator of this physical capacity. Laboratory tests that directly measure oxygen uptake during heavy physical exertion are the gold standard for VO_{2max} measurement. These tests, which can be performed on treadmills or cycle ergometers, involve technical and performance factors that can introduce measurement errors (Howley, Bassett, & Welch, 1995). This paper summarizes the empirical evidence regarding the size of those errors.

Measurement error biases empirical estimates of relationships between VO_{2max} and other variables. The bias produces estimated associations that are less than the true population relationships (Nunnally & Bernstein, 1994). The technical term for this underestimation is attenuation due to measurement error. Better estimates of population parameters can be obtained by adjusting for this attenuation. The magnitude of error must be known to make the necessary corrections.

This review examines the measurement error for VO_{2max} tests when the results are expressed as milliliters of oxygen uptake per kilogram of body weight per minute ($ml \cdot kg^{-1} \cdot min^{-1}$). Meta-analysis provides a model to predict SEM based on the pooled evidence from available studies.

Methods

Data Sources

The PUBMED® computer database was searched to identify relevant studies. The search keywords were reliability or reproducibility combined with maximal oxygen uptake or VO_{2max}. The resulting set of articles was augmented with citations from Safrit, Hooper, Ehlert, Costa, and Patterson (1988) and Hopkins, Schabort, & Hawley (2001). The references in the articles identified in these first 2 steps were examined to identify additional studies.

The studies in this review met three criteria. First, oxygen uptake was expressed in units of $ml \cdot kg^{-1} \cdot min^{-1}$. This size-adjusted expression is the most common index of cardiorespiratory capacity in studies of health and performance. Second, the study reported SEM or sufficient information to compute SEM (i.e., the standard deviation and r_{xx} or intraclass correlation [ICC] for VO_{2max}). Third, the test-retest interval was no more than 3 weeks.[1]

[1] Twelve studies met the first two criteria but had retest intervals longer than 5 weeks. Preliminary analysis indicated that SEM was much larger in

Table 1. Descriptive Data

	Mean	Median	Minimum	Maximum
Age	30.3	27.2	9.8	79.8
VO_{2max}	45.3	46.2	13.1	68.5
Reliability (r_{xx})[a]	.890	.909	.620	.970
SD_t[b]	5.68	5.55	1.52	9.50
SEM[c]	2.64	2.53	1.30	5.00

Note. Table entries are the weighted statistics for the raw data. Sample size was the weighting factor. Values for r_{xx}, SEM, and SD_t reported in the text may differ from these because the raw data were transformed to approximate normal distributions before analysis.
[a] Test-retest reliability coefficient.
[b] Standard deviation of true scores.
[c] Standard error of measurement.

Thirty-one (31) studies covering 39 samples with 745 total participants met these criteria. The studies included 29 treadmill, 8 cycle ergometer, and 2 miscellaneous (e.g., tethered swimming) protocols. The protocols included 500 treadmill tests, 187 cycle ergometer tests, and 58 miscellaneous tests.

Coding Procedures

The mean and standard deviation for each test administration and the correlation between scores (i.e., r_{xx}) were recorded when reported. When raw data were reported, statistics in the paper were confirmed by repeating the basic data analysis. When the ICC was reported, ICC and the number of test administrations (*k*) were entered into the Spearman-Brown formula, $r_{ICC} = kr_{ij}/(1+r_{ij})$ where r_{ij} is the *average* correlation between VO_{2max} values for the i^{th} and j^{th} test administrations (Ghiselli, Campbell, & Zedeck, 1981, p. 232). The formula was solved for r_{ij}, which then was the study estimate of r_{xx}.

Additional information was extracted to examine factors that might modify SEM. Gender, age, and VO_{2max} were recorded as sample attributes that might indicate limits on the generalizability of SEM (see Table 1). Study design attributes were recorded to identify methodological factors that could be controlled to minimize SEM. Exercise mode coded as treadmill (k = 29 samples, N = 500 cases), cycle ergometer (k = 8, N = 187), and other (k = 2, N = 58).[2] Only 27

those studies than in the 39 studies retained for analysis. Changes in true VO_{2max} would be one possible explanation for the large errors. Dropping these studies helped ensure that the review evaluated protocol performance without the confounding effects of changes in VO_{2max}.

[2] The initial study plan for the review included coding protocol attributes (e.g., how initial work rate was determined, frequency and size of work rate

studies provided enough information to estimate test-retest interval. Typical descriptions referred to a range of times (e.g., “7 to 10 days,” “7 or more days”) between tests. This practice no doubt reflects the difficulty of maintaining precise scheduling when participants must return to the laboratory more than once. When a range was given, the midpoint of the range was recorded. When only a lower bound was given, this minimum value was recorded. The initial interval estimates were recoded into categories: “<1 week” (k = 9; N = 177), “7–10 days” (k = 11, N = 221), and “2–3 weeks” (k = 7; N = 134). Test interval could not be determined for 12 samples (N = 213). A missing data value was entered for those samples.

Analysis Procedures

SEM and SD_t were computed as follows:

$$SEM = \sqrt{}(1 - r_{xx}^2)*SD_{VO2max}$$
$$SD_t = r_{xx}*SD_{VO2max}$$

These variance components and r_{xx} were transformed to obtain normal distributions with known variances (Raudenbush & Bryk, 2002, p. 219, for the conversion formulae). The meta-analyses were conducted by applying standard regression and general linear model procedures (SPSS, Inc., 1998a, 1998b) to the transformed variables. In these analyses, the transformed variance component or correlation was weighted by the inverse of its known variance. Given this weighting, the sum of squares from the analyses provided Hedges’s Q (Hedges & Olkin, 1985, pp. 241-242). The Q statistic has a χ^2 distribution with k – 1 degrees of freedom (df) where k is the number of correlations or variance estimates being analyzed.

Preliminary analyses established two facts that affected decisions regarding the results reported here. First, the mean and standard deviation of the initial VO_{2max} test was an acceptable estimate of these statistics for both test administrations (Appendix A). Second, r_{xx} was substantially lower and SEM substantially higher when more than 3 weeks elapsed between tests. This difference was expected because testing conditions could change over the longer intervals. Possible changes include alterations in true VO_{2max} scores. The possibility of substantial changes in the person, the laboratory equipment, seasonal effects on physical activity and other factors would introduce major elements of uncertainty into attempts to evaluate SEM. Based on these preliminary analyses, the results reported here are based on the mean and standard deviation from the first test session for studies with test intervals $\leq$3 weeks.

increments, criteria for a valid VO_{2max}). Protocol details were missing from too many studies to support the planned analysis.

Table 2. Categorical Predictors of Reliability and Precision

	SD_t	SEM	r_{xx}
Test interval			
<7 days	5.39	2.56	.905
7-10 days	4.89	2.65	.883
2-3 weeks	5.58	3.00	.887
χ^2	3.06	3.65	1.06
p value	.216	.161	.588
Exercise mode			
Cycle ergometer	6.50	2.47	.936
Treadmill	5.15	2.59	.897
Other	6.67	2.87	.920
χ^2	17.06	1.96	7.71
p value	.001	.375	.022
Gender			
Missing	4.77	2.71	.875
Male	6.27	2.71	.920
Female	3.87	1.94	.898
M + F	7.20	3.58	.896
χ^2	50.46	30.83	4.37
p value	.001	.001	.225

Results

Bivariate Relationships

Test Interval. SEM increased slightly, but consistently, as the interval between tests increased, but the trend was not statistically significant (χ^2 = 3.65, 2 *df*, *p* > .161). SD_t was not related to test interval (χ^2 = 3.06, 2 *df*, *p* > .216). Test-retest reliability, r_{xx}, did not vary (χ^2 = 1.06, 2 *df*, *p* > .588).

The estimates of test interval effects may be biased. Studies for which interval could not be estimated had higher average SD_t (6.62 $ml \cdot kg^{-1} \cdot min^{-1}$) and lower average SEM (2.28 $ml \cdot kg^{-1} \cdot min^{-1}$) compared with studies with interval data. Test-retest reliability was higher reliability (r_{xx} = .946). The differences were statistically significant (*p* < .001 for each). The missing data would bias the estimates of interval effects if the studies with missing data all had approximately the same interval. The direction and magnitude of the bias would depend on where the cluster was located on the time continuum.

Exercise Mode. SEM was not related to exercise mode (χ^2 = 1.96, 2 *df*, *p* > .375). SD_t was lower for cycle ergometer protocols than for treadmill and other protocols (χ^2 = 17.06, 2 *df*, *p* < .001). Combining these trends produced small, but statistically significant differences in r_{xx} (χ^2 = 7.71, 2 *df*, *p* < .001).

Gender. SEM was larger for males (2.71) than for females (1.94; χ^2 = 22.03, 1 *df*, *p* < .001). SD_t was greater in samples of male (SD_t = 6.27) than in samples of females (SD_t = 3.87; χ^2 = 46.03, 1 *df*, *p* < .001). These opposite trends combined to yield comparable r_{xx} values for men and women (χ^2 = 1.62, 1 *df*, *p* > .204).

Age. Age was not related to SEM (*r* = -.15, χ^2 = 2.86, 1 *df*, *p* > .090) or SD_t (*r* = .12, χ^2 = 1.77, 1 *df*, *p* > .183). These weak opposing trends produced increasing r_{xx} with age (*r* = .27, χ^2 = 4.10, 1 *df*, *p* < .043).

VO_{2max}. VO_{2max} was positively related to SEM (*r* = .35, χ^2 = 16.59, 1 *df*, *p* < .001), but was not related to SD_t (*r* = .02, χ^2 = 0.04, 1 *df*, *p* > .841). The combined trends produced a negative relationship between r_{xx} and VO_{2max} (*r* = -.31, χ^2 = 6.07, 1 *df*, *p* < .014).

Multivariate Model for SEM

The general linear model routine of SPSS-PC was used to combine VO_{2max} and gender, the significant bivariate correlates of SEM, into a multivariate model. Each variable contributed independently to the prediction of SEM (VO_{2max}, χ^2 = 5.94, 1 *df*, *p* < .015; gender, χ^2 = 13.44, 1 *df*, *p* < .001). The regression formula was

$$\ln(SEM) = 0.985 + (.006^* VO_{2max}) - (.275^* \text{Gender}) \quad \text{(Equation 1)}$$

Sensitivity Analysis

Meta-analysis should attempt to evaluate the sensitivity of the results to assumptions embedded in the analysis (National Research Council, 1992). The potential bias associated with missing time interval data has been alluded to previously. The construction of the multivariate model therefore was followed by exploration of several factors that might have affected the content and structure of the model.

Gender Coding. The assumptions made in coding gender might have affected the model. Perhaps gender was less likely to be reported when the sample was composed of males. This assumption could be valid if male gender was an implicit default value in this research domain. Samples with missing data were reclassified as male to test this possibility.

The reclassification had little effect. Both gender and VO_{2max} were significantly related to SEM. The regression slope for VO_{2max} was unchanged. The slope for gender was .003 smaller. The intercept was .011 larger. These changes were reasonable given that the average values of SEM, SD_t, and r_{xx} for the samples with missing data were very similar to the average values for male samples.

Additional Female Data. The modest amount of data available for females was a second concern. The evidence included only 6 samples of women. Data from Katch, Sady, and Freedson (1982) were added to increase the total number of observations for women. That study included an intensive investigation of 4 women. Each of these four women completed between 10 and 21 tests with at least 1 day between tests. Each woman completed her series of tests in 2 to 4 weeks. The Katch et al. (1982) data had not been included in the analyses to this point because the set of tests for each woman comprised a time series. Correlations between errors could occur that would lead to underestimation of error variance (Ostrom, 1990). The possible lack of independence between observations also raises special statistical problems in meta-analysis (Becker & Schram, 1994). However, the data were used in this sensitivity analysis because the primary objective was to improve the estimate of average SEM rather than to make statistical inferences.

The unweighted average mean squared error for the four series was 3.2 $ml \cdot kg^{-1} \cdot min^{-1}$. This value was larger than the estimated average SEM for women in the 6 test-retest studies. In fact, this error was larger than the estimated value for men. Adding these data, the estimated SEM for females increased from 2.02 $ml \cdot kg^{-1} \cdot min^{-1}$ to 2.30 $ml \cdot kg^{-1} \cdot min^{-1}$. Although statistical inferences based on these data must be viewed with caution, it is worth noting that the gender difference still would be statistically significant ($\chi^2 = 3.90$, 1 *df*, $p < .049$) if the measurement errors were treated as independent from session to session.

Outlier/Influential Data Points. A point noted when coding the data was examined next. The standard deviation of VO_{2max} had been coded in two prior reviews of VO_{2max} as a predictor of running performance (Vickers, 2001a, 2001b). The distribution of standard deviations indicated a typical standard deviation of ~6.00 $ml \cdot kg^{-1} \cdot min^{-1}$. Very few values were <3.00 $ml \cdot kg^{-1} \cdot min^{-1}$ or >9.00 $ml \cdot kg^{-1} \cdot min^{-1}$. Thus, a small sample of studies such as that covered in this review should include very few standard deviations beyond the range from 3 to 9 $ml \cdot kg^{-1} \cdot min^{-1}$. Other things equal, values outside this range would produce extreme SEM values.

Analysis of the standard deviations from the prior reviews (Vickers, 2001a, 2001b) gave reason to believe the current set of studies was not broadly representative of men and women. The samples in this review accentuated a gender difference evident in the larger body of evidence. The male standard deviation was larger than expected (*SD* = 6.77 $ml \cdot kg^{-1} \cdot min^{-1}$ vs. *SD* = 6.16 $ml \cdot kg^{-1} \cdot min^{-1}$, χ^2 = 59.95, 26 *df*, *p* < .001). The female standard deviation was smaller than expected (*SD* = 4.31 $ml \cdot kg^{-1} \cdot min^{-1}$ vs. 5.66 $ml \cdot kg^{-1} \cdot min^{-1}$, χ^2 = 26.52, 6 *df*, *p* < .001). The gender difference in the data reviewed here was 5 times what would generally be expected (2.46 $ml \cdot kg^{-1} \cdot min^{-1}$ vs. 0.50 $ml \cdot kg^{-1} \cdot min^{-1}$). This trend produced a larger χ^2 for the male-female difference in the review data than in the larger body of evidence (χ^2 = 22.03 vs. χ^2 = 16.07) despite the smaller cumulative sample size in the present data.

The weighted average standard deviation for VO_{2max} was computed for 121 samples of men and 51 samples of women in the prior reviews. The average standard deviations for men and women then were the points of reference for computing z-scores for the studies in this review (i.e., [ln(Sample SD) – ln(Average SD)]*2f; f = N – 1, cf., Raudenbush & Bryk, 2002, p. 219). The computations produced $|z|$ > 3.00 for 6 studies. This frequency was >56 times the number that would be expected by chance. These samples, therefore, could be classified as outliers (Barnett & Lewis, 1978). Further analyses were undertaken to determine the impact of the outliers on the prior analysis findings (cf., Belsley, Kuh, & Welsch, 1980; Stevens, 1984).

The extreme values were not randomly distributed. Three of six female samples produced *z* < -3.00. For males, two samples produced *z* > 3.00; one sample yielded *z* < -3.00. In the context of this study, the extreme standard deviations strongly suggested that SEM might be underestimated for women. The implications for men were less clear, but it was possible that SEM was overestimated for males.

Each gender analysis described above was repeated after removing the extreme samples. When this was done, men and women had virtually identical SEM values (χ^2 < 0.40). The removal did not affect the VO_{2max}-SEM relationship. This association remained positive and statistically significant.

Re-examination of Exercise Mode. The effect of exercise mode on SEM was reexamined to complete the sensitivity checks. The question was whether exercise mode was related to SEM controlling for VO_{2max} and gender. This analysis was not based on any prior finding. The question was posed to check on a logical possibility that would be important if true. The initial analysis covered all of the exercise modes. The analysis was repeated for the subset of studies involving either the treadmill or cycle ergometer protocols. In each analysis, the χ^2 for exercise mode was less than would be expected by chance (i.e., χ^2/df < 1.00).

Revised Model

The sensitivity analyses suggested that the gender effects in Equation 1 were questionable. Therefore, a final predictive model for SEM was constructed with VO_{2max} as the only predictor. The model was

$$\ln(SEM) = 0.531 + (.009 * VO_{2max}) \qquad \text{(Equation 2a)}$$

This equation yields SEM = 2.23 when VO_{2max} = 30 $ml \cdot kg^{-1} \cdot min^{-1}$, SEM = 2.67 $ml \cdot kg^{-1} \cdot min^{-1}$ when VO_{2max} = 50 $ml \cdot kg^{-1} \cdot min^{-1}$, and 3.19 $ml \cdot kg^{-1} \cdot min^{-1}$ when VO_{2max} = 70 $ml \cdot kg^{-1} \cdot min^{-1}$. When the samples with exceptional standard deviations were removed, the equation was

$$\ln(SEM) = 0.661 + (.006 * VO_{2max}) \qquad \text{(Equation 2b)}$$

The relationship to VO_{2max} remained significant (r = .23, χ^2 = 4.43, 1 *df*, p < .036). The revised equation yields SEM estimates of 2.32 when VO_{2max} = 30 $ml \cdot kg^{-1} \cdot min^{-1}$, 2.61 when VO_{2max} = 2.67 at 50 $ml \cdot kg^{-1} \cdot min^{-1}$, and 2.95 when VO_{2max} = 70 $ml \cdot kg^{-1} \cdot min^{-1}$. Equation 2b can be seen as a robust version of Equation 2a because the outlier data points have been given less weight (Rousseeuw & Leroy, 1987).

Discussion

This review of the available evidence regarding the measurement precision of VO_{2max} tests provides a basis for addressing several general issues. First, the review produced reasonably firm conclusions about some factors that might influence precision. Second, the review defined some topics for continuing research by showing that additional evidence is needed to reach conclusions about other factors that might affect precision. Third, the review produced a simple model for estimating precision in studies that lack repeated measures. SEM estimates from this model can be useful when evaluating findings from past and future studies. Finally, the evidence illustrated that the standard error for a test should be the preferred statistical index of test performance. These points are discussed below.

Firm conclusions could be reached regarding 3 factors that might influence the precision of VO_{2max} measurements. Two findings are negative. Age does not affect precision. Treadmill and cycle ergometer tests have equal precision. The third finding provides the basis for a model to estimate SEM. Precision is lower when VO_{2max} is higher. Results obtained from the analysis of sample statistics must be extrapolated to individuals to reach this third conclusion, but the extrapolation seems reasonable. The applied uses of a model based on these conclusions are considered after summarizing the negative findings.

The evidence regarding 3 other potential influences on VO_{2max} test precision was ambiguous. Outliers made it impossible to estimate gender effects with precision. Lack of replication made it impossible to decide whether tests involving alternative exercise methods (e.g., tethered swimming) produced larger than average SEM values. Two methods have been tested and both produced larger than average SEM values. Neither result has been replicated to date. Also, those two methods are not necessarily representative of the universe of alternatives to treadmill and cycle ergometry protocols. Finally, SEM probably increases as the time between measurements increases, but this position cannot be adopted with certainty. The weak time trend shown in Table 1 and preliminary analyses showing larger SEM values in studies with intervals in excess of 5 weeks support the presence of a time effect. However, time interval could not be coded for a subset of studies. The studies in that subset had small SEM values. Their distribution along the time dimension could dramatically affect any temporal trend.

The established facts generate a simple model of VO_{2max} test precision. Precision decreases as VO_{2max} increases. Application of the model involves two steps. First, compute the natural logarithm of SEM, $\ln(y) = 0.661 + (.006 * VO_{2max})$. Second, compute the estimated SEM, $SEM' = \exp(y)$. The areas of uncertainty discussed in the preceding paragraph make it likely that this simple model is incomplete. There is a strong likelihood that a complete model would include time interval between measurements. It is less likely, but still possible, that a complete model also would include gender. However, the current model is based on the only association definitely supported by the available evidence.

The SEM estimates derived from the predictive model provide a frame of reference for evaluating 2 types of research results. The first type evaluates methods of assessing cardiorespiratory fitness. In this context, the model estimates provide a benchmark for new VO_{2max} protocols. When a new protocol is being evaluated, Equation 2b can be applied to estimate the treadmill or cycle ergometer SEM for the study sample. The observed SEM can be compared with the estimate by computing $z = (SEM – SEM')*(2N-4)$ (cf., Raudenbush & Bryk, 2002, p. 219). Standard statistical criteria ($p < .05$, one-tailed) can be applied to decide whether the SEM for the new protocol exceeds that for the reference standards. The model predictions also can have a role in the validation of field tests of cardiorespiratory fitness. In this context, VO_{2max} test results can be regressed on field test performance (e.g., run time) to obtain a standard error of estimate (SEE). If the z-score computations show that the field test is less precise than the laboratory test, as would be expected, the increase in error associated with the field test can be estimated by computing $\sqrt{(SEE_{Field}^2 – SEM_{Lab}^2)}$.

The second general application of the SEM model involves correcting for the effects of measurement error when studying relationships between VO_{2max} and other variables. SEM′ estimates from Equation 2 can be used to compute $r_{xx} = (1 - SEM'^2/SD^2)$. Inserting the computed r_{xx} into the formula $\rho_{xy} = r_{xy}/\sqrt{r_{xx}}$ provides an estimate of the true population correlation, ρ_{xy}, by correcting the observed correlation, r_{xy}, for unreliability of x variable, VO_{2max}. The computation treats the y variable as though it were measured without error (i.e., $r_{yy} = 1.00$). By doing so, the correction focuses specifically on the effects of measurement errors in VO_{2max}. The correction is not limited to correlational studies. Similar adjustments can be applied to a wide range of analysis procedures by transforming group differences into effect size estimates comparable to r_{xy}. For example, the difference between two groups in an experiment can be expressed as a point biserial correlation (cf., Hedges & Olkin, 1985). The effects of these corrections can be substantial (Appendix D).

The potential applications of SEM estimates are important. For example, consider a study relating VO_{2max} to some other variable (e.g., a training method, an ergogenic aid) in a small sample of endurance athletes. The true effect size is likely to be underestimated. Selection processes that determine who becomes an endurance athlete are likely to cause true differences in VO_{2max} to be smaller than in the general population. At the same time, the high average level of VO_{2max} implies higher than average SEM. The combination of restricted true score variance and large error variance will yield attenuated estimates of associations between VO_{2max} and other variables. Tests of statistical significance that combine small effect sizes with small samples have low statistical power and are not likely to reject the null hypothesis. Even moderate to strong true effects can fail to reach statistical significance under these conditions. Using the SEM model to estimate the effect of measurement error in such a study could reduce the risk of dismissing promising lines of work prematurely.

The review findings also demonstrated that SEM is preferable to other statistical indices when evaluating the measurement characteristics of VO_{2max} protocols. A distinction between absolute and scaled indices of test performance is the key issue here. SEM quantifies the reproducibility of individual VO_{2max} values in absolute terms. SEM is the average expected error. Other widely used statistics scale SEM by expressing it relative to some sample characteristic. Scaling is most evident for the coefficient of variation (CV). CV expresses SEM as a percentage of average VO_{2max} (i.e., CV = SEM/Average*100). Test-retest reliability, r_{xx}, scales SEM relative to the sample standard deviation. This reliability index is defined as the ratio of true score variance to total variance (i.e., $r_{xx} = SD_t^2/SD^2$; Nunnally & Bernstein, 1994). Because $SD^2 = SD_t^2 + SEM^2$, it is also true that $r_{xx} = 1 - SEM^2/SD^2$. Thus, r_{xx} scales SEM relative to the sample standard deviation. Both r_{xx} and CV are composite statistics in that each combines SEM with a sample characteristic. The analysis of SD_t in this review demonstrated that the sample characteristics sometimes are associated with factors when SEM is not (cf., Table 1). The discussion of Equation 2 showed that, in the present data, CV can decrease when SEM is increasing. In both cases, variations in the composite indices are a poor guide to the precision of test scores. In the final analysis, experts regard reproducibility of test results as the essence of reliability (e.g., American Psychological Association, 1994). SEM is the best index for assessing reliability because it separates test precision from population attributes.

The volume and quality of evidence must be considered when evaluating the points made in this discussion. Publication bias occurs when only statistically significant results are published (National Research Council, 1992). This practice inflates parameter estimates because studies that produce smaller values are missing from the published record. This form of bias is unlikely in the present case. In this domain, significance tests would be expected to evaluate the null hypothesis that $r_{xx} = .00$. Given the average value of r_{xx} for the studies reviewed here (i.e., $r_{xx} = .91$) and a sample size of 15 (i.e., the median for the studies reviewed), the null hypothesis will be rejected 99.99% of the time. Publication bias therefore does not appear likely to have had a major effect on the present findings.

Another trend in the evidence may appear to be a cause for concern. The analyses identified 15% of the studies as outliers. This rate is not exceptional. Outliers commonly comprise 10% to 20% of the data in fields as diverse as behavioral research and particle physics (Hedges, 1987; Hedges & Olkin, 1985). Furthermore, the conclusions drawn from the evidence have taken account of the outliers. The evidence regarding gender effects was treated as inconclusive because the outlier data points affected this element of the analyses.

To summarize, this review provided a simple model of the measurement precision (i.e., SEM) of VO_{2max} tests. The model is probably incomplete, but it represents the combined evidence from available studies of VO_{2max} reliability. Further research on the effects of time interval between tests and gender differences in SEM could refine this initial model. The model generates SEM estimates that can be used as benchmarks when evaluating new VO_{2max} protocols. The SEM estimates also can be applied to correct for measurement error when estimating associations between VO_{2max} and other variables. From a statistical perspective, the evidence indicates that SEM provides a better indication of test performance than r_{xx} or CV when evaluating VO_{2max} protocols. This summary sketch of the available evidence can be a framework for future studies of the measurement properties of VO_{2max} protocols.

References

American Psychological Association (1994). *Publication manual of the American Psychological Association* (4th ed.). Washington, DC: Author.

Barnett, V., & Lewis, T. (1978). *Outliers in statistical data*. Chichester: John Wiley.

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression diagnostics: identifying influential data and sources of collinearity*. New York: John Wiley.

Becker, B. J., & Schram, C. M. (1994). Examining explanatory models through research synthesis. In H. Cooper & L. V. Hedges (Eds.), *The Handbook of Research Synthesis* (pp. 357-382). New York: Russell Sage Foundation.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88*(3), 588-606.

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences*. San Francisco: Freeman.

Hedges, L. V. (1987). How hard is hard science, how soft is soft science? The empirical cumulativeness of research. *American Psychologist, 42*(2), 443-455.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Hoelter, J. W. (1983). The analysis of covariance structures: Goodness-of-fit indices. *Sociological Methods and Research, 11*, 325-344.

Hopkins, W. G., Schabort, E. J., & Hawley, J. A. (2001). Reliability of power in physical performance tests. *Sports Medicine*, *31*, 211-234.

Howley, E. T., Bassett, D. R., J., & Welch, H. G. (1995). Criteria for maximal oxygen uptake: Review and commentary. *Medicine & Science in Sports & Exercise, 27*(9), 1292-1301.

Joreskog, K. G., & Sorbom, D. (1996). *LISREL 8: User’s reference guide*. Chicago: Scientific Software International.

Katch, V. L., Sady, S. S., & Freedson, P. (1982). Biological variability in maximum aerobic power. *Medicine and Science in Sports and Exercise*, *14*, 21-25.

Mulaik, S. A., James, L. R., Van Alstine, J., Bennett, N., Lind, S., & Stilwell, C. D. (1989). Evaluation of goodness-of-fit indices for structural equation models. *Psychological Bulletin*, *105*(3), 430-445.

National Research Council. (1992). *Combining information: Statistical issues and opportunities for research*. Washington, DC: National Academy Press.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory (3rd ed.)*. New York: McGraw-Hill.

Ostrom, C. W., Jr. (1990). *Time series analysis: Regression techniques* (2nd ed.). Newbury Park, CA: Sage.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Rosenthal, R. (1978). Combining results of independent studies. *Psychological Bulletin, 85*, 185-193.

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection*. New York: John Wiley.

Safrit, M. J., Hooper, L. M., Ehlert, S. A., Costa, M. G., & Patterson, P. (1988). The validity generalization of distance run tests. *Canadian Journal of Sport Sciences, 13*(4), 188-196.

SPSS, Inc. (1998a). *SPSS Base 8.0 Applications Guide*. Chicago: SPSS, Inc.

SPSS, Inc. (1998b). *SPSS Advanced Statistics*. Chicago: SPSS, Inc.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, *25*, 173-180.

Stevens, J. P. (1984). Outliers and influential data points in regression analysis. *Psychological Bulletin*, *95*(2), 334-344.

Vickers, R. R., Jr. (2001a). *Running performance as an indicator of* VO_{2max}*: Distance effects* (Technical Report No. 01-20). San Diego, CA: Naval Health Research Center.

Vickers, R. R., Jr. (2001b). *Running performance as an indicator of* VO_{2max}*: A replication of distance effects* (Technical Report No. 01-24). San Diego, CA: Naval Health Research Center.

Appendix A

Studies Providing Data for the Meta-analysis

Aunola, S., & Rusko, H. (1984). Reproducibility of aerobic and anaerobic thresholds in 20-50 year old men. *European Journal of Applied Physiology*, *53*, 260-266.

Babcock, M. A., Paterson, D. H., & Cunningham, D. A. (1992). Influence of ageing on aerobic parameters determined from a ramp test. *European Journal of Applied Physiology*, *65*, 137-143.

Bar-Or, O., & Zwiren, L. D. (1975). Maximal oxygen consumption test during arm exercise--reliability and validity. *Journal of Applied Physiology, 38*(3), 424-426.

Boileau, R. A., Bonen, A., Heyward, V. H., & Massey, B. H. (1977). Maximal aerobic capacity on the treadmill and bicycle ergometer of boys 11-14 years of age. *Journal of Sports Medicine*, *17*, 153-162.

Brandon, L. J., & Boileau, R. A. (1987). The contribution of selected variables to middle and long distance run performance. *Journal of Sports Medicine*, *27*, 157-164.

Braun, B. A., Clarkson, P. M., Freedson, P. S., & Kohl, R. L. (1991). Effects of coenzyme Q10 supplementation on exercise performance, VO_{2max}, and lipid peroxidation in trained cyclists. *International Journal of Sports Nutrition*, *1*, 353-365.

Conley, K. E., Esselman, P. C., Jubrias, S. A., Cress, M. E., Inglin, B., Mogadam, C., & Schoene, R. B. (2000). Ageing, muscle properties and maximal O_2 uptake rate in humans. *Journal of Physiology*, *526*, 211-217.

Cunningham, D. A., MacFarlane van Waterschoot, B., Paterson, D. H., Lefcoe, M., & Sangal, S. P. (1977). Reliability and reproducibility of maximal oxygen uptake measurement in children. *Medicine & Science in Sports & Exercise, 9*(2), 104-108.

Cunningham, D. A., Telford, P., & Swart, G. T. (1976). The cardiopulmonary capacities of young hockey players: age 10. *Medicine & Science in Sports & Exercise, 8*(1), 23-25.

De Meersma, R. E. (1992). Respiratory sinus arrhythmia alteration following training in endurance athletes. *European Journal of Applied Physiology*, *64*, 434-436.

De Vito, G., Bernardi, Sproviero, E., & Figura, F. (1995). Decrease of endurance performance during Olympic triathlon. *International Journal of Sports Medicine*, *16*, 24-28.

Farrell, P. A., Wilmore, J. H., Coyle, E. F., Billing, J. E., & Costill, D. L. (1979). Plasma lactate accumulation and distance running performance. *Medicine and Science in Sports*, *11*, 338-344.

Fielding, R. A., Frontera, W. R., Hughes, V. A., Fisher, E. C., & Evans, W. J. (1997). The reproducibility of the Bruce protocol exercise test for the determination of aerobic capacity in older women. *Medicine & Science in Sports & Exercise, 29*(8), 1109-1113.

Foster, C. C., Jr. (1972). *Maximal aerobic power and the aerobic requirements of running in trained runners and trained non-runners*. Unpublished master's thesis, University of Texas at Austin.

Foster, V. L., Hume, G. J. E., Dickinson, A. L., Chatfield, S. J., & Byrnes, W. C. (1986). The reproducibility of VO_{2max}, ventilatory, and lactate thresholds in elderly women. *Medicine & Science in Sports & Exercise, 18*(4), 425-430.

Froelicher, V. F., Jr., Brammell, H., Davis, G., Noguear, I., Stewart, A., & Lancaster, M. C. (1974). A comparison of three maximal treadmill exercise protocols. *Journal of Applied Physiology, 36*(6), 720-725.

Harrison, M. H., Bruce, D. L., Brown, G. A., & Cochrane, L. A. (1980). A comparison of some indirect methods for predicting maximal oxygen uptake. *Aviation, Space, and Environmental Medicine*, *51*, 1128-1133.

Harrison, M. H., Brown, G. A., & Cochrane, L. A. (1980). Maximal oxygen uptake: Its measurement, application, and limitations. *Aviation, Space, and Environmental Medicine*, *51*, 1123-1127.

Hazard, A. A. (1982). *The effects of endurance training at 2,440m altitude on maximal oxygen uptake at altitude and sea level in young male and female middle distance runners.* Unpublished master's thesis, San Diego State University, San Diego, CA.

Huhn, R. R. (1975). The reliability, validity, and predictability of twelve- and fifteen-minute field tests in relation to laboratory maximal oxygen uptake tests. Unpublished master's thesis, San Diego State University, San Diego, CA.

Jackson, A. S., Beard, E. F., Wier, L. T., Ross, R. M., Stuteville, J. E., & Blair, S. N. (1995). Changes in aerobic power of men ages 25-70 yr. *Medicine and Science in Sports and Exercise, 27*, 113-120.

Jackson, A. S., Wier, L. T., Ayers, G. W., Beard, E. F., Stuteville, J. E., & Blair, S. N. (1996). Changes in aerobic power of women, ages 20-64 yr. *Medicine & Science in Sports & Exercise, 28*, 884-891.

Joreskog, K., & Sorbom, D. (1996). *LISREL 8: User's reference guide*. Chicago: SSI Scientific Software International.

Katch, F. I., McArdle, W. D., Czula, R., & Pechar, G. S. (1973). Maximal oxygen intake, endurance running performance, and body composition in college women. *Research Quarterly, 44*, 301-312.

Kohrt, W. M., Morgan, D. W., Bates, B., & Skinner, J. S. (1987). Physiological responses of triathletes to maximal swimming, cycling, and running. *Medicine and Science in Sports and Exercise, 19*(1), 51-55.

Kukkonen-Harjula, K., Laukkanen, R., Vuori, I., Oja, P., Pasanen, M., Nenonen, A., & Uusi-Rasi, K. (1998). Effects of walking running on health-related fitness in healthy middle-aged adults—A randomized controlled study. *Scandinavian Journal of Medicine and Science in Sports,* 8, 236-242. (Note: Source of descriptive statistics for samples in Laukkanen et al., 2000).

Kyle, S. B., Smoak, B. L., Douglass, L. W., & Deuster, P. A. (1989). Variability of responses across training levels to maximal treadmill exercise. *Journal of Applied Physiology*, *67*, 160-165.

Laukkanen, R. M. T., Kukkonen-Harjula, T. K., Oja, P., Pasanen, M. E., & Vuori, I. M. (2000). Prediction of change in maximal aerobic power by the 2-km walk test after walking training in middle-aged adults. *International Journal of Sports Medicine*, *21*, 113-116.

MacSween, A. (2001). The reliability and validity of the Astrand nomogram and linear extrapolation for deriving VO_{2max} from submaximal exercise data. *Journal of Sports Medicine and Physical Fitness, 41*, 312-317.

Magel, J. R., & Faulkner, J. A. (1967). Maximum oxygen uptakes of college swimmers. *Journal of Applied Physiology, 22*(5), 929-938.

McArdle, W. D., Katch, F. I., Pechar, G. S., Jacobson, L., & Ruck, S. (1972). Reliability and interrelationships between maximal oxygen intake, physical work capacity and step-test scores in college women. *Medicine and Science in Sports, 4*, 182-186.

Miller, F. R., & Manfredi, T. G. (1987). Physiological and anthropometrical predictors of 15-kilometer time trial cycling performance time. *Research Quarterly for Exercise and Sport*, *58*, 250-254.

Montgomery, D. L, Reid, G., & Koziris, L. P., (1992). Reliability and validity of three fitness tests for adults with mental handicaps. *Canadian Journal of Sports Science, 17*, 309-315.

Paterson, D. H., Cunningham, D. A., & Donner, A. (1981). The effect of different treadmill speeds on the variability of VO_{2max} in children. *European Journal Applied Physiology, 47*, 113-122.

Pawelcyzk, J. A., Kenney, W. L., & Kenney, P. (1988). Cardiovascular responses to head-up tilt after an endurance exercise program. *Aviation, Space, and Environmental Medicine*, 59, 107-112.

Pivarnik, J. M., Dwyer, M. C., & Lauderdale, M. A. (1996). The reliability of aerobic capacity (VO_{2max}) testing in adolescent girls. *Research Quarterly for Exercise and Sport, 67*(3), 345-348.

Shaver, L. G. (1975). Maximum aerobic power and anaerobic work capacity prediction from various running performances of untrained college men. *Journal of Sports Medicine, 15*, 147-150.

Sproule, J., Kunalan, C., McNeill, M., & Wright, H. (1993). Validity of 20-MST for predicting VO_{2max} of adult Singaporean athletes. *British Journal of Sports Medicine, 27*(3), 202-204.

Thomas, S. G., Cunningham, D. A., Rechnitzer, P. A., Donner, A. P., & Howard, J. H. (1987). Protocols and reliability of maximal oxygen uptake in the elderly. *Canadian Journal of Sport Sciences, 12*, 144-151.

Turley, K. R., Rogers, D. M., & Wilmore, J. H. Maximal testing in prepubescent children: Treadmill versus cycle ergometry. *Medicine & Science in Sports & Exercise, 25*(5), S9 (Abstract 49).

Walters, S. C. (1983). The physiological effects of a twenty week distance running program on teenage girls correlated with performance. Unpublished Master’s Thesis. Arizona State University, Tempe, AZ.

Ward, T. E., Hart, C. L., McKeown, B. C., & Kras, J. (1998). The Bruce treadmill protocol: Does walking or running during the fourth stage alter oxygen consumption values? *Journal of Sports Medicine and Physical Fitness, 38*, 132-137.

Weltman, A., Snead, D., Stein, P., Seip, R., Schurrer, R., Rutt, R., & Weltman, J. (1990). Reliability and validity of a continuous incremental treadmill protocol for the determination of lactate threshold, fixed blood lactate concentrations, and VO_{2max}. *International Journal of Sports Medicine, 11*(1), 26-32.

Appendix B
Data Coded from Studies

Table B-1. Descriptive Data

Sr. Author	1	2	3	4	5	6	7	8	9	10
Aunola	33	1	1	2	33.0	7.8	177	4.9	75.6	7.5
Babcock	79	1	1	.	50.1	13.9	180	6.0	83.9	12.2
Bar-Or	41	1	3	3	28.2	8.8	174	9.8	70.8	13.2
Boileau	21	1	2	2	12.8	1.1	159	13.3	49.1	12.8
Boileau	21	1	1	2	12.8	1.1	159	13.3	49.1	12.8
Brandon	26	1	2	.	26.7	4.4	178	4.6	69.5	7.9
Braun	12	1	1	4	21.9	1.8	178	11.9	72.6	8.3
Conley	12	.	1	.	68.8	5.9	.	.	72.1	2.1
Cunningham 76	15	1	1	1	10.6	.3	141	5.6	35.5	5.4
Cunningham 77	66	1	2	4	10.4	.3	140	6.5	33.5	5.3
De Meersma	9	1	1	4	20.0	.	.	.	.	.
De Vito	6	1	2	2	27.0	5.0	176	8.0	69.0	9.0
Farrell	18	1	2	.	28.0	9.0	180	6.7	70.2	8.1
Fielding	17	2	2	2	59.0	4.1	162	4.5	62.5	8.2
Foster	8	2	2	2	79.8	4.6	159	3.9	58.4	8.0
Froehlicher	15	.	2	3	32.0	.	178	.	78.0	.
Froehlicher	15	.	2	3	.	.	178	.	78.0	.
Froehlicher	15	.	2	3	.	.	178	.	78.0	.
Harrison 1	9	1	2	4	31.0	.	71.2	.	69.9	.
Harrison 2	5	1	2	2	31.1	.	.	.	.	.
Harrison 3	9	1	2	.	.	.	.	.	.	.
Harrison 4	10	1	2	1	.	.	.	.	.	.
Hazard	7	2	2	3	18.4	1.1	164	4.1	48.8	3.1
Hazard	21	1	2	3	19.0	2.3	175	5.1	61.2	7.2
Huhn	20	1	2	1	25.9	2.8	179	7.1	72.2	7.4
Jackson	156	1	2	4	45.6	5.0	.	.	82.3	13.6
Jackson	43	2	2	4	44.2	8.9	.	.	63.4	12.0
Katch	36	2	2	2	20.8	1.4	163	6.6	58.9	6.8
Kohrt	13	1	2	.	29.5	4.8	.	.	69.8	5.6
Kohrt	13	1	1	.	29.5	4.8	.	.	69.8	5.6
Kyle	17	1	2	1	31.9	4.6	181	6.3	78.0	9.7
Laukkanen	25	1	2	4	41.4	.	.	.	80.8	9.3
Laukkanen	26	1	2	4	41.4	.	.	.	84.0	10.4
Laukkanen	28	2	2	4	40.9	.	.	.	66.8	8.9
Laukkanen	29	2	2	4	40.9	.	.	.	68.6	8.6
MacSween	25	3	2	1	28.6	7.3	173	10.0	69.8	13.8
Magel	17	.	3	1	19.8	1.0	181	6.1	76.7	7.4
McArdle	41	2	2	1	20.9	1.3	163	6.0	58.2	6.9
Miller	5	1	1	.	26.7	5.8	178	5.3	73.9	7.1
Montgomery	10	.	2	.	24.8	4.0	172	7.1	75.1	15.9
Paterson	8	1	2	1	11.4	.	147	7.4	36.9	7.5
Pawelcyzk	10	3	.	4	22.8	3.3	176	10.1	71.7	17.1
Pivarnik	32	2	2	2	13.7	1.5	157	6.0	53.7	9.3

Table B-1. Descriptive Data (continued)

Sr. Author	1	2	3	4	5	6	7	8	9	10
Shaver	10	1	2	.	22.5	3.2	173	6.7	75.5	13.3
Sproule	20	1	2	3	23.0	3.3	171	8.3	60.8	7.8
Thomas	24	1	2	1	61.7	.	175	.	78.7	.
Turley	9	3	2	.	9.8	.	.	.	.	.
Turley	9	3	1	.	9.8	.	.	.	.	.
Walters	10	2	2	4	15.3	1.2	.	.	54.0	7.3
Ward	27	1	2	2	39.1	10.7	180	6.7	78.3	8.4
Weltman	15	1	2	2	27.2	8.2	175	7.5	69.1	8.3

Note. Columns are 1=Sample size (N); 2=Gender; 3=Protocol Type; 4=Interval Group; 5=Average Age; 6= SD Age; 7=Avgerage Height; 8= SD Height; 9=Average Weight; 10=SD Weight SD = standard deviation??

Table B-2. VO$_{2max}$ Statistics

Sr. Author	1	2	3	4	5	6	7	8	9
Aunola	47.60	6.20	48.20	6.50	.920	.958	5.1	2.49	1.81
Babcock	31.00	7.50	.	.	.960	.980	6.8	2.10	.
Bar-Or	30.37	9.05	31.04	9.01	.940	.969	10.2	3.08	2.21
Boileau	48.70	5.30	49.70	6.10	.870	.930	5.4	2.81	2.13
Boileau	44.90	6.30	46.30	6.60	.880	.936	6.7	3.06	2.24
Brandon	62.50	6.10	.	.	.910	.953	4.0	2.53	.
Braun	61.97	4.78	62.12	3.84	.909	.952	3.2	1.80	1.45
Conley	22.90	6.50	23.80	6.20	.959	.979	8.0	1.80	1.30
Cunningham 76	56.60	7.70	.	.	.760	.864	8.8	5.00	.
Cunningham 77	56.50	7.10	54.50	6.60	.530	.693	10.7	5.81	4.71
De Meersma	56.20	2.40	60.60	4.20	.722	.839	3.0	2.28	2.10
De Vito	68.50	5.21	64.50	6.47	.662	.797	5.7	4.38	3.49
Farrell	43.20	6.60	.	.	.950	.974	4.8	2.06	.
Fielding	27.50	4.50	28.30	5.40	.750	.857	10.8	3.27	2.55
Foster	13.10	2.00	13.40	1.80	.760	.864	9.9	1.23	.94
Froehlicher	43.90	5.70	44.60	6.20	.851	.920	6.8	3.12	2.32
Froehlicher	48.10	6.00	47.20	6.20	.941	.970	4.2	2.06	1.49
Froehlicher	43.60	4.80	43.30	5.70	.620	.765	8.6	4.12	3.29
Harrison 1	63.70	9.04	61.83	7.18	.887	.940	6.6	3.75	3.01
Harrison 2	58.66	9.90	59.84	10.27	.955	.977	5.0	3.00	2.15
Harrison 3	60.36	8.94	61.59	10.67	.916	.956	6.0	3.94	3.08
Harrison 4	58.09	7.36	58.78	9.34	.898	.946	5.6	3.67	3.00
Hazard	54.44	4.53	58.96	5.47	.769	.869	5.3	3.20	2.48
Hazard	64.98	4.86	68.71	5.56	.786	.880	4.6	3.22	2.46
Huhn	58.28	6.09	58.11	6.68	.960	.980	2.9	1.79	1.34
Jackson	37.20	7.30	37.20	7.00	.660	.795	14.7	5.37	4.17
Jackson	30.10	7.10	27.80	6.40	.855	.922	12.2	3.50	2.61
Katch	38.90	4.60	.	.	.950	.974	3.7	1.44	.
Kohrt	60.50	5.60	.	.	.970	.985	2.3	1.36	.
Kohrt	57.90	5.70	.	.	.930	.964	3.6	2.10	.
Kyle	56.90	10.00	.	.	.950	.974	5.5	3.12	.
Laukkanen	43.50	3.50	50.20	4.90	.795	.886	4.9	2.55	2.12
Laukkanen	43.30	3.60	47.00	4.80	.646	.785	6.3	3.21	2.61
Laukkanen	37.20	5.60	41.90	6.70	.834	.909	8.3	3.39	2.61
Laukkanen	36.80	5.20	38.70	5.70	.881	.937	6.7	2.58	1.91
MacSween	50.14	8.75	.	.	.895	.945	7.8	3.90	.
Magel	55.00	4.00	55.00	3.20	.830	.907	4.1	2.01	1.58
McArdle	38.14	3.87	38.70	4.02	.909	.952	4.2	1.64	1.21
Miller	59.70	6.70	.	.	.963	.981	3.0	1.81	.
Montgomery	45.50	8.10	.	.	.907	.951	7.5	3.41	.
Paterson	58.90	6.60	60.30	4.70	.864	.927	5.6	2.84	2.45
Pawelcyzk	41.20	10.51	44.80	9.96	.939	.969	8.8	3.52	2.56
Pivarnik	41.20	5.20	40.80	4.80	.870	.930	6.2	2.47	1.82
Shaver	53.50	5.60	.	.	.920	.958	4.1	2.19	.
Sproule	51.50	6.04	.	.	.900	.947	5.1	2.63	.

Table B-2. VO_{2max} Statistics (continued)

Sr. Author	1	2	3	4	5	6	7	8	9
Thomas	24.70	5.40	25.90	6.40	.900	.947	9.5	2.57	.98
Turley	51.70	5.90	.	.	.905	.950	4.9	2.51	.
Turley	46.20	6.70	.	.	.887	.940	6.7	3.09	.
Walters	45.15	3.62	49.34	5.19	.781	.877	5.0	2.75	2.31
Ward	53.00	9.20	53.40	9.60	.850	.919	9.1	4.95	3.65
Weltman	63.30	4.70	65.60	6.70	.710	.830	5.2	4.01	3.34

Note. Columns are 1. Average VO_{2max}, Time 1; 2=SD VO_{2max}, Time 1; 3=Average VO_{2max}, Time 2; 4=SD VO_{2max}, Time 2; 5=Test-Retest correlation (r_{xx}); 6=Intraclass Correlation (ICC); 7=Coefficient of variation (CV); 8=Standard Error of Measurement (SEM); 9=Maximum Likelihood SEM. See text for definitions.

Appendix C

Preliminary Evaluation of VO_{2max} as a Function of Test Occasion

A set of preliminary analyses evaluated the equivalence of the first and second VO_{2max} tests. The general hypothesis was that the tests were equivalent. This general hypothesis could be tested in the set of 24 studies for which the mean and standard deviation were available for both tests.

Average Scores

Average VO_{2max} values were highly correlated (r = .984). The regression equation to predict the second average based on the first average was $V_2' = 0.749 + 1.000*V_1$, where V_2'indicates the predicted VO_{2max} for the second measurement based on the first measurement. The standard errors for the coefficients were 1.394 and .031, respectively. Thus, the 95% confidence interval (CI) for the slope included 1.00 (CI = 0.939, 1.061). The 95% CI for the intercept included 0.00 (CI = -1.98, 3.47). The average VO_{2max} was higher for the first test than for the second test in 18 of 24 samples. However, the differences were uniformly small. The weighted average was 44.19 $ml \cdot kg^{-1} \cdot min^{-1}$ for the first test and 43.45 $ml \cdot kg^{-1} \cdot min^{-1}$ for the second test. The difference was too small to be significant in every individual sample ($|t| < 1.29$). The difference was not even significant when pooled across samples ($Z = -1.05$, $p > .293$, method of adding ts; cf. Rosenthal, 1978).

Standard Deviations

The sample estimates of the standard deviation were stable (r = .89). The regression to predict the standard deviation for the first test was $S_2' = 1.420 + 0.802*S_1$. The standard errors were 0.498 and 0.070, respectively. The 95% CI for the slope approached, but did not reach 1.00 (CI = .665, .939). The 95% CI for the intercept did not include 0.00 (CI = .444, 2.396). However, these results may be related to the fact that regression models assume that the predictor variable is measured without error. The confirmatory factor analysis (CFA) results clearly indicated that SEM was invariant across test occasions.

Confirmatory Factor Analysis (CFA) Models

LISREL 8.5 (Joreskog & Sorbom, 1996) was used to fit CFA models that tested 2 important hypotheses regarding the standard deviations. The CFA models treated each VO_{2max} test as an indicator of a single VO_{2max} latent trait. The model constrained the factor loading to be the same for both tests. This constraint embodied the assumption that a person's true VO_{2max} did not change between test sessions. If so, the

Table C-1. Comparison of CFA Models

							Crit	
Model	*df*	χ^2	Sig.	RMSEA	p(close)	NNFI	N	SRMR
Invariant	47	118.38	.0000	.296	.214	.944	264	.321
Mode Specific	45	84.83	.0004	.226	.315	.958	311	.309
Sample Specific	24	23.79	.4738	.000	.532	.997	697	.278

Note. *df* = degrees of freedom; RMSEA = root mean-square error of approximation (Steiger, 1990); NNFI = non-normed fit index (Bentler & Bonett, 1980); Crit N = critical N (Hoelter, 1983); SRMR = standardized root mean-square residual (Joreskog & Sorbom, 1996).

true score variance for the sample would be unchanged. The factor loadings are indicators of this true score variance, so it follows that they would be unchanged.

Alternative models were defined by imposing constraints on the standard errors. Every model imposed the constraint that SEM was the same for both tests within each sample. Different models were obtained by varying whether equality constraints were imposed across samples. The broadest constraint assumed that SEM was constant across all studies. A second model assumed that SEM differed between exercise modes, but was constant within modes. A third model assumed that each study produced a unique SEM that was constant across test sessions for that sample.

The LISREL analysis was limited to the 24 studies with standard deviation data for both tests. One model constrained the error to be the same across all tests (invariant). One model constrained the error to be the same within exercise mode (mode specific). One model permitted a distinct error for each sample (sample specific).

All 3 models were acceptable by several criteria (Table C-1). The *p*(close) values indicated that each model was within chance of the recommended RMSEA = .05 value. All 3 NNFI values exceeded .900. All 3 critical *N*s exceeded 200.

Criteria that differed between models generally favored the sample-specific model. First, the overall χ^2 decreased significantly moving from the invariant model to the group-specific model (χ^2 = 33.55, 2 *df*, *p* < .001) and then from the group-specific model to the sample-specific (χ^2 = 61.04, 21 *df*, *p* < .001). Second, the sample-specific model was the only one for which the overall χ^2 was nonsignificant. Third, the RMSEA estimate for the sample-specific model was .000. Fourth, the critical N sample-specific was more than twice as large as that for the group-specific model. Finally, the SRMR was smallest for the sample-specific model. However, if parsimony adjustments had been introduced the NNFI for the sample-specific model would have been substantially lower than that for the other two models.

The CFA led to 2 primary conclusions. First, SEM can be regarded as invariant across tests. This result means that analysis of the SEM for the first test is a reasonable basis for inferences about test errors. Second, SEM varies from sample to sample. This inference is supported by the general improvement in model fit when the sample-specific model is compared with the alternatives. The other models would be adequate by accepted modeling standards and could be preferred for their parsimony (Mulaik et al., 1989). However, an erroneous inference about the existence of sample-specific values for SEM should not cause problems. This review attempts to identify factors that affect SEM. If the sample-to-sample variation is truly the result of chance, there should be only a few chance associations between SEM and potential predictors. The modeling attempts should reinforce the inference that SEM differences are chance.

Appendix D

Using SEM to Correct Effect Size Estimates

VO_{2max} SEM estimates can be used to assess the effects of measurement error on estimates of associations between VO_{2max} and other variables. For example, the correlation between two variables, x and y, is

$$r_{xy} = C_{xy}/(S_x*S_y)^{1/2} \qquad \text{(Equation D-1)}$$

Given the usual assumption that errors are uncorrelated, C_{xy}, the covariance of x and y, is determined entirely by the correspondence between true scores. S_x, the standard deviation of x, and S_y, the standard deviation of y, are composites of true score and error variance.

SEM can be used to correct effect size estimates because eliminating measurement error does not affect C_{xy}. This parameter is not affected because SEM, by definition, does not contribute to C_{xy}. However, eliminating measurement error does reduce S_x and/or S_x. The elimination of measurement error increases r_{xy} because the denominator of Equation 3 is reduced while the numerator remains constant.

SEM estimates can be used to correct for attenuation due to measurement error. The first step in using SEM estimates true score variance, SD_t, by computing $S' = \sqrt{}(S^2 - \text{SEM}^2)$. An SEM estimate based on Equation 2 can be used for this computation. Substituting S' into Equation 3 then provides an adjusted estimate of r_{xy}.

To illustrate the correction process, consider the association of running performance with VO_{2max}. The expected association would be r = .82 with an associated SD = 6.2 for VO_{2max} (Vickers, 2001a, 2001b). Adopting Equation 2b as a robust estimate of the relationship between SEM and average VO_{2max}, SEM = 2.7 at 50 $ml \cdot kg^{-1} \cdot min^{-1}$. The estimated true score variance is $S' = 5.6$ $ml \cdot kg^{-1} \cdot min^{-1}$. If VO_{2max} is the y variable, the denominator for Equation 3, S_x*S_y', is 10% smaller than the original denominator, S_x*S_y. C_{xy} remains constant, so the smaller denominator increases the estimated run time-VO_{2max} correlation increases to $r = .91$. Although this illustration involves a correlation coefficient, the general approach extends to most common measures of effect size because effect size indicators generally can be converted into correlations (Hedges & Olkin, 1985).

The correction procedure illustrated above is equivalent to applying a well-known equation to correct for attenuation due to measurement error. The equation is based on r_{xx}, but the equivalence follows from the relationships between r_{xx}, SD_t, and SEM (cf., Nunnally & Bernstein, 1994, pp. 260-262).

REPORT DOCUMENTATION PAGE

1. Report Date (DD MM YY)
14 Nov 03

2. Report Type
Interim

3. DATES COVERED (from - to)
1 Jan 03 to 14 Nov 03

4. TITLE AND SUBTITLE
(U) Measurement Error in Maximal Oxygen Uptake Tests

5a. Contract Number:
5b. Grant Number:
5c. Program Element: 63706N
5d. Project Number: M0096
5e. Task Number: 001
5f. Work Unit Number: 6418
USAMMRC Reimbursable 60109

6. **AUTHORS**
Vickers, Ross R., Jr.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Health Research Center
P.O. Box 85122
San Diego, CA 92186-5122

9. PERFORMING ORGANIZATION REPORT NUMBER
Report No. 04-03

8. SPONSORING/MONITORING AGENCY NAMES(S) AND ADDRESS(ES)

Office of Naval Research
800 North Quincy St.
Arlington, VA 22217-5600

Chief, Bureau of Medicine and Surgery
Code: BUMED-26
2300 E Street NW
Washington, D.C.

10. Sponsor/Monitor's Acronyms(s)

11. Sponsor/Monitor's Report Number(s)

12 DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT (maximum 200 words)

Cardiorespiratory fitness is important for health, work, and athletic performance. Laboratory tests of maximal oxygen uptake (VO_{2max}) are the gold standard for assessing cardiorespiratory fitness. Test protocols with small measurement errors accurately estimate current fitness levels of individuals, thereby permitting test results to accurately describe relations between VO_{2max} and potential antecedents and consequences of fitness. This review applied meta-analysis methods to summarize the evidence from 39 studies of the test-retest reliability of VO_{2max} measurements. The standard error of measurement (SEM) was the statistical index of test precision. SEM was higher in samples with higher average values of VO_{2max}. SEM may be higher for men than for women, but the evidence was equivocal. SEM was not related to age or to the exercise mode for the test protocol. SEM increased as the interval between tests increased, but the trend is not statistically significant. These SEM findings provide a frame of reference for evaluating new laboratory protocols for VO_{2max}, determining the loss of precision when a field test is substituted for a laboratory protocol, and correcting for error when estimating associations between VO_{2max} and other variables. Further study of gender and test interval effects on the accuracy of VO_{2max} measurements could help resolve uncertainties about the range of factors affecting VO_{2max} test precision.

15. SUBJECT TERMS
aerobic capacity, measurement methods, measurement error, test reliability, statistical model

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Commanding Officer
a. REPORT	**b.ABSTRACT**	**b. THIS PAGE**			
UNCL	UNCL	UNCL	UNCL	25	**19b. TELEPHONE NUMBER (INCLUDING AREA CODE)** COMM/DSN: (619) 553-8429