Journal of Sports Sciences, July 2008; 26(9): 967 – 976

Routledge
Taylor & Francis Group

Do regression-based computer algorithms for determining the ventilatory threshold agree?

PANTELEIMON EKKEKAKIS[1], ERIK LIND[1], ERIC E. HALL[2], & STEVEN J. PETRUZZELLO[3]

[1]*Department of Kinesiology, Iowa State University, Ames, Iowa,* [2]*Department of Health and Human Performance, Elon University, Elon, North Carolina and* [3]*Department of Kinesiology and Community Health, University of Illinois at Urbana-Champaign, Illinois, USA*

(Accepted 11 January 2008)

Abstract
The determination of the ventilatory threshold has been a persistent problem in research and clinical practice. Several computerized methods have been developed to overcome the subjectivity of visual methods but it remains unclear whether different computerized methods yield similar results. The purpose of this study was to compare nine regression-based computerized methods for the determination of the ventilatory threshold. Two samples of young and healthy volunteers ($n = 30$ each) participated in incremental treadmill protocols to volitional fatigue. The ventilatory data were averaged in 20-s segments and analysed with a computer program. Significant variance among methods was found in both samples (Sample 1: $F = 11.50$; Sample 2: $F = 11.70$, $P < 0.001$ for both). The estimates of the ventilatory threshold ranged from 2.47 litres · min^{-1} (71% $\dot{V}O_{2max}$) to 3.13 litres · min^{-1} (90% $\dot{V}O_{2max}$) in Sample 1 and from 2.37 litres · min^{-1} (67% $\dot{V}O_{2max}$) to 3.03 litres · min^{-1} (83% $\dot{V}O_{2max}$) in Sample 2. The substantial differences between methods challenge the practice of relying on any single computerized method. A standardized protocol, likely based on a combination of methods, might be necessary to increase the methodological consistency in both research and clinical practice.

Keywords: *Gas exchange threshold, anaerobic threshold, computer algorithms, limits of agreement*

Introduction

The ventilatory or gas exchange threshold has been a concept immersed in controversy since its inception. Researchers have questioned whether it is a valid marker of the so-called ''anaerobic threshold'', whether it is causally linked to the lactate threshold, and whether it occurs concurrently with the lactate threshold (Brooks, 1985; Myers & Ashley, 1997; Svendahl & MacIntosh, 2003). Nevertheless, the ventilatory threshold is considered a useful index of functional capacity for patients suffering from cardiovascular and pulmonary conditions (Meyer, Lucia, Earnest, & Kindermann, 2005; Wasserman, 1997), has been shown to be of prognostic value (Gitt *et al.*, 2002; Older, Hall, & Hader, 1999), and is used as a measure of the effectiveness of exercise interventions (Gaskill *et al.*, 2001).

A factor that limits the practicality of the ventilatory threshold is the difficulty associated with its determination. Methods based on the visual inspection of ventilatory data plots have been widely criticized for showing poor inter-evaluator and inter-method agreement (Caiozzo *et al.*, 1982; Dickstein, Barvik, Aarsland, Snapinn, & Karlsson, 1990a; Gladden, Yates, Stremel, & Stamford, 1985; Shimizu *et al.*, 1991; Yeh, Gardner, Adams, Yanowitz, & Crapo, 1983). Consequently, computerized methods have gained popularity, as they offer the promise of objectivity, reliability, and automation. However, none of the computerized methods seems to have provided a definitive solution and thus a *de facto* standard has yet to emerge.

Nevertheless, the complexity of the computational problem might not be fully appreciated. Marketing materials for software programs that accompany metabolic analysis systems, which are usually based on a single regression-based algorithm, promise ''automatic determination'' of the ventilatory threshold. Reliance on single computerized algorithms is increasingly common in both research and clinical practice. However, whether published regression-based

Correspondence: P. Ekkekakis, Department of Kinesiology, Iowa State University, 235 Barbara E. Forker Building, Ames, IA 50011, USA.
E-mail: ekkekaki@iastate.edu

ISSN 0264-0414 print/ISSN 1466-447X online © 2008 Taylor & Francis
DOI: 10.1080/02640410801910269

algorithms produce similar estimates has not been investigated. Thus, the purpose of the present study was to compare estimates derived from nine computerized methods of determining the ventilatory threshold, all of which are based on various types of regression analysis and have been published in the literature.

Methods

Participants

Two groups, each consisting of 30 young and healthy adult volunteers, participated in the study. Sample 1 consisted of 13 women (mean age 22.8 years, $s = 3.0$; body mass 63.7 kg, $s = 9.8$; $\dot{V}O_{2max}$ 46.9 ml $\cdot$ kg^{-1} $\cdot$ min^{-1}, $s = 4.1$) and 17 men (mean age 24.4 years, $s = 4.1$; body mass 78.1 kg, $s = 7.1$; $\dot{V}O_{2max}$ 51.5 ml $\cdot$ kg^{-1} $\cdot$ min^{-1}, $s = 7.0$). Sample 2 consisted of 14 women (mean age 21.2 years, $s = 2.0$; body mass 60.6 kg, $s = 6.6$; $\dot{V}O_{2max}$ 47.7 ml $\cdot$ kg^{-1} $\cdot$ min^{-1}, $s = 7.6$) and 16 men (mean age 21.5 years, $s = 2.5$; body mass 78.5 kg, $s = 9.2$; $\dot{V}O_{2max}$ 56.6 ml $\cdot$ kg^{-1} $\cdot$ min^{-1}, $s = 7.3$). All participants signed an informed consent form approved by the university's institutional review board.

Procedure

The two groups participated in a different treadmill protocol. The purpose of using different samples and protocols was to determine (a) whether there was a tendency for certain pairs of methods to yield consistent estimates regardless of the data set, suggesting a systematic agreement, and (b) whether there was a tendency for certain methods to consistently (regardless of data set) yield lower or higher estimates compared with the other methods. For Sample 1, after a 3-min warm-up (4.8 km $\cdot$ h^{-1}, 0% incline), the workload was increased in 2-min stages, by alternating between increases in speed of 1.6 km $\cdot$ h^{-1} or incline of 2%, starting from 8 km $\cdot$ h^{-1} and 0% incline. For Sample 2, after a 5-min warm-up (4.8 km $\cdot$ h^{-1}, 0% incline), the workload was increased in 1-min stages, by alternating between increases in speed of 0.8 km $\cdot$ h^{-1} or incline of 1%, again starting from 8 km $\cdot$ h^{-1} and 0% incline.

The increases in workload continued until each participant reached volitional fatigue. Attainment of $\dot{V}O_{2max}$ was confirmed by at least two of the following three criteria: (a) a plateau in $\dot{V}O_2$ (changes < 2 ml $\cdot$ kg^{-1} $\cdot$ min^{-1} following an increase in workload); (b) respiratory exchange ratio > 1.1; and (c) reaching age-predicted maximum heart rate (i.e. 220 $-$ age).

The oxygen analyser (S-3A/I, Ametek Applied Electrochemistry, Naperville, IL) and carbon dioxide analyser (CD-3A, Ametek Applied Electrochemistry) sampled from a mixing chamber at 120 Hz for Sample 1 and with each breath for Sample 2. The analysers and a turbine volume measurement system (KTC3, Ametek Applied Electrochemistry) were calibrated before each test.

Computerized determination of the ventilatory threshold

A computer program (WinBreak 3.7) that incorporates all the algorithms of interest was developed for this study. Data preparation involved three steps. First, non-physiological (e.g. negative) values were removed. Second, the data were averaged every 20 s (Gaskill *et al.*, 2001). Third, lower and upper boundaries for the ventilatory threshold calculations were set. The lower boundary was set at the end of the warm-up (minute 3 for Sample 1, minute 5 for Sample 2) or at the end of the initial decrease in analyses involving the ventilatory equivalents. The upper boundary was set either at the end of the test or at the respiratory compensation point, if one was found. The respiratory compensation point was determined by a slightly modified version of the procedure of Beaver and colleagues (Beaver, Wasserman, & Whipp, 1986) based on $\dot{V}_E$ by $\dot{V}CO_2$ data (see Figure 1). Specifically, Beaver and colleagues recommended dividing the $\dot{V}_E$ by $\dot{V}CO_2$ data into two linear segments and locating the point at which the slope of the second segment exceeded the slope

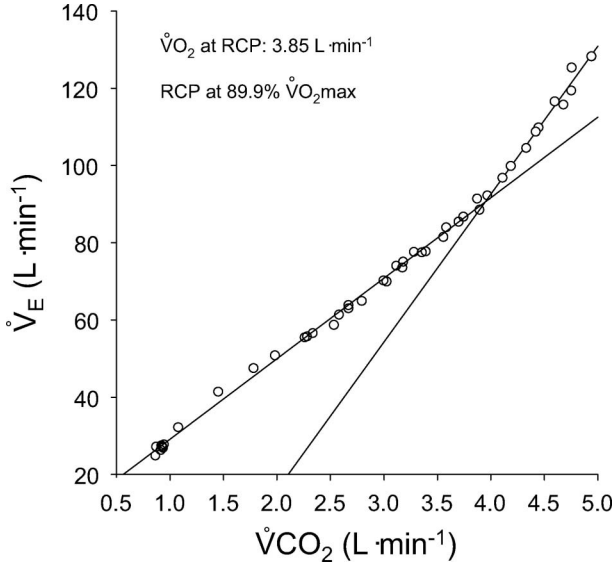

Figure 1. An illustration of the method used for the determination of the respiratory compensation point (RCP), which was used as the upper boundary in the ventilatory threshold calculations. The $\dot{V}_E$ by $\dot{V}CO_2$ data were divided into two segments fitted with linear regressions. If a significant departure from linearity was found, the data point demarcating the segments with the largest slope difference was designated as the respiratory compensation point.

of the first by a pre-selected amount (15%). They did note, however, that a respiratory compensation point does not always occur and our experience confirms this. Thus, we first examined whether the data showed a significant departure from linearity. If they did not (which occurred in 20 of the 60 cases), the end of the treadmill test was used as the upper boundary for the ventilatory threshold calculations. If they did, we set the respiratory compensation point at the point of the largest slope difference between the two segments, because using a fixed slope difference (such as the 15% mentioned by Beaver *et al.*) occasionally resulted in untenable results (e.g. points below 50% $\dot{V}O_{2max}$). After these preliminary data processing steps, the ventilatory threshold was estimated using the following nine algorithms (see Figure 2).

Method 1. The first method consisted of using the general-purpose "breakpoint" algorithm developed by Jones and Molitoris (1984) in conjunction with $\dot{V}CO_2$ by $\dot{V}O_2$ data, as implemented by Schneider and colleagues (Schneider, Phillips, & Stoffolano, 1993). The algorithm considers two regression equations, one before and one after the breakpoint x_0: $y_i = b_0 + b_1x$, where $x \leq x_0$, and $y_i = b_2 + b_3x$, where $x > x_0$. The following constraint forces the two regression lines to join at x_0: $b_0 + b_1x_0 = b_2 + b_3x_0$. By solving for b_2, $b_2 = b_0 + b_1x_0 - b_3x_0$, the two equations can be rewritten as: $y_i = b_0 + b_1x$, where $x < x_0$, and $y_i = b_0 + b_1x_0 + b_3(x - x_0)$, where $x \geq x_0$. This is a four-parameter regression, consisting of three linear parameters (b_0, b_1, b_3) and one non-linear parameter (x_0). The method searches for the value of x_0 that minimizes the residual sum of squares.

Method 2. The second method consisted of a two-segment version of the "brute force" algorithm proposed by Orr and colleagues (Orr, Green, Hughson, & Bennett, 1982) in conjunction with $\dot{V}CO_2$ by $\dot{V}O_2$ data. The algorithm calculates regression parameters for all possible divisions of the data into two contiguous groups and the pair of regressions yielding the least pooled residual sum of squares is selected as the best-fitting solution.

Method 3. The third method consisted of using the "V-slope" algorithm of Beaver *et al.* (1986) in conjunction with $\dot{V}CO_2$ by $\dot{V}O_2$ data. The algorithm for locating the ventilatory threshold consists of (a) systematically dividing the $\dot{V}CO_2$ by $\dot{V}O_2$ data into two contiguous segments, (b) fitting each segment with a linear regression, (c) calculating the location of the intersection between the two regression lines, (d) identifying the pair of regression lines that maximizes the ratio of the distance of the intersection

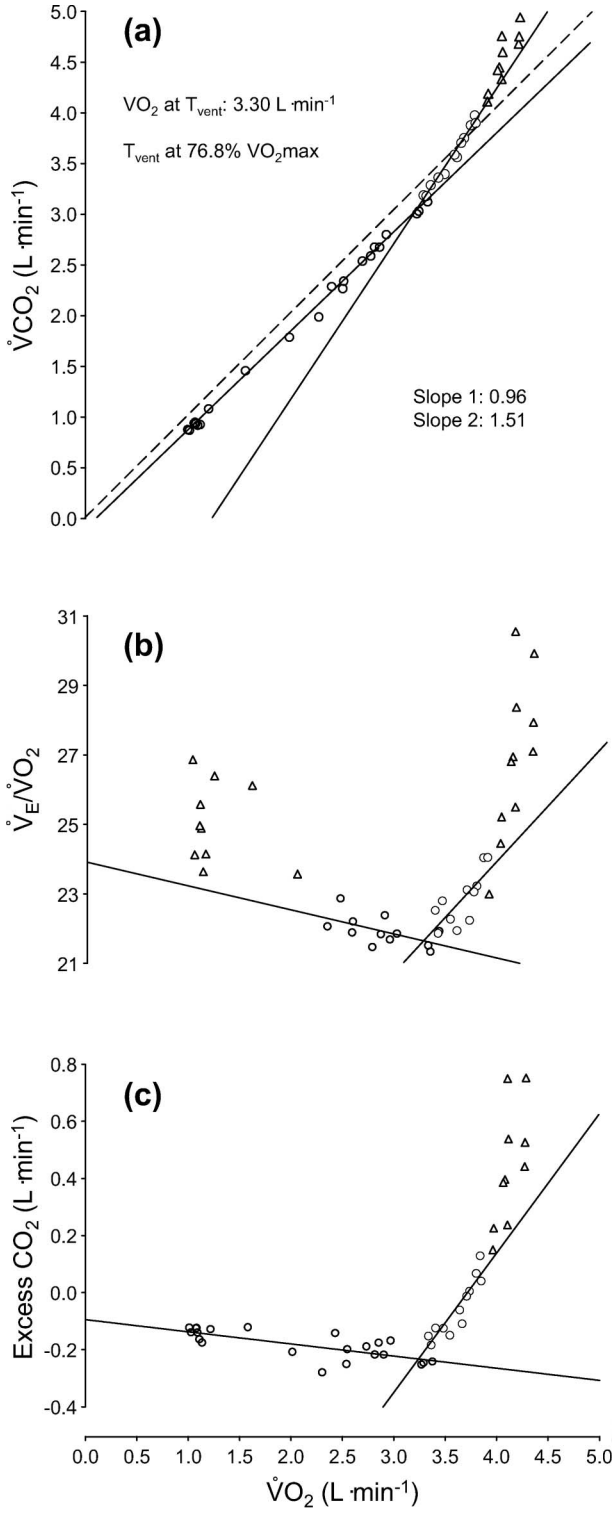

Figure 2. An example of computer plots showing agreement between methods for the determination of the ventilatory threshold (T_{vent}) based on different data sets. Open triangles symbolize data points below the lower boundary or above the upper boundary of T_{vent} calculations. Open circles symbolize data points before and dark circles symbolize data points after the calculated T_{vent}. Solid lines indicate regression lines fitted to the sub-T_{vent} and supra-T_{vent} data segments. Panel (a) shows $\dot{V}CO_2$ by $\dot{V}O_2$ data. The dotted line is the line of identity (i.e. slope of 1). Panel (b) shows $\dot{V}_E/\dot{V}O_2$ by $\dot{V}O_2$ data. Panel (c) shows excess CO_2 by $\dot{V}O_2$ data.

point from a single regression line through the entire data set to the mean square error of regression, and (e) ensuring that the slope of the first regression line is >0.6 and the change in slope from the first regression to the second is >0.1.

Method 4. The fourth method consisted of using the "Dmax" algorithm of Cheng *et al.* (1992) in conjunction with $\dot{V}CO_2$ by $\dot{V}O_2$ data. The algorithm involves (a) fitting a third-order curvilinear regression to the $\dot{V}CO_2$ by $\dot{V}O_2$ data, (b) drawing a straight line connecting the first and last points of the curve, (c) calculating the distances of all the data points along the curve to the straight line, and (d) identifying the point that yields the maximal distance (Dmax) from the straight line.

Method 5. The fifth method consisted of using the "simplified V-slope" algorithm developed by Sue and colleagues (Sue, Wasserman, Moricca, & Casaburi, 1988) and computerized by Dickstein and co-workers (Dickstein, Barvik, Aarsland, Snapinn, & Millerhagen, 1990b) in conjunction with $\dot{V}CO_2$ by $\dot{V}O_2$ data. In the implementation used in the present study, (a) regression parameters are calculated for all possible divisions of the data into two contiguous groups, and (b) a point is identified at which the slope of the first regression is between 0.6 and 1.0, the slope of the second regression is >1.0, and the difference between the slopes of the first and second regression is >0.1.

Method 6. The sixth method consisted of using the "breakpoint" algorithm of Jones and Molitoris (1984) in conjunction with $\dot{V}_E/\dot{V}O_2$ by $\dot{V}O_2$ data. The data points containing the initial decreasing trend in $\dot{V}_E/\dot{V}O_2$ were removed before analysis.

Method 7. The seventh method was similar to the sixth, with the only difference being that $\dot{V}_E/\dot{V}O_2$ by time data were used.

Method 8. The eighth method consisted of using the "breakpoint" algorithm of Jones and Molitoris (1984) in conjunction with excess CO_2 by $\dot{V}O_2$ data. Excess CO_2 was defined as $(\dot{V}CO_2{}^2/\dot{V}O_2) - \dot{V}CO_2$, following Gaskill *et al.* (2001).

Method 9. The ninth method was similar to the eighth, with the only difference being that the excess CO_2 by time data were used.

Statistical analyses

The data were analysed by (a) analyses of variance and paired *t*-tests to determine whether different methods yielded significantly different estimates, (b) Pearson bivariate correlations to examine the relationships between different methods, (c) Spearman rho to examine the relationship between the rank order of estimates derived from the two samples, and (d) Bland and Altman's (1986, 2003) limit of agreement analysis to examine the degree of agreement between methods.

Results

Indeterminate cases

As shown in Table I, Methods 3 and 5, because of the restrictions they place on the viability of the solutions, resulted in some indeterminate cases. Specifically, Method 3 led to two and one, and Method 5 led to four and one, indeterminate cases in Samples 1 and 2, respectively.

Differences between methods

Analyses of variance on the estimates of the ventilatory threshold from the nine methods were significant in both samples (Sample 1: $F_{8,192} = 11.50$, $P < 0.001$; Sample 2: $F_{8,224} = 11.70$, $P < 0.001$). The estimates of the ventilatory threshold ranged from 2.47 litres $\cdot$ min^{-1} (71% $\dot{V}O_{2max}$) (Method 5) to 3.13 litres $\cdot$ min^{-1} (90% $\dot{V}O_{2max}$) (Method 2) in Sample 1 and from 2.37 litres $\cdot$ min^{-1} (67% $\dot{V}O_{2max}$) (Method 5) to 3.03 litres $\cdot$ min^{-1} (83% $\dot{V}O_{2max}$) (Method 7) in Sample 2.

The mean inter-method difference (average of all 36 pairwise comparisons) was similar in the two samples at 7% ($s = 4$) $\dot{V}O_{2max}$ for Sample 1 and 7% ($s = 4$) $\dot{V}O_{2max}$ for Sample 2. For Sample 1, the lowest inter-method difference was 0.4% $\dot{V}O_{2max}$ (Method 3 vs. Method 4), whereas the highest was 19% $\dot{V}O_{2max}$ (Method 2 vs. Method 5). For Sample 2, the lowest inter-method difference was 0.2% $\dot{V}O_{2max}$ (Method 1 vs. Method 4), whereas the highest was 16% $\dot{V}O_{2max}$ (Method 5 vs. Method 7). The method that yielded the lowest average difference with all other methods was Method 3 in Sample 1 (5% $\dot{V}O_{2max}$) and Method 2 in Sample 2 (5% $\dot{V}O_{2max}$). However, averaged across both samples, the method that yielded the lowest average inter-method difference was Method 4 (5% $\dot{V}O_{2max}$). On the other hand, the comparisons involving Method 5 consistently yielded the highest inter-method differences across the two samples (11% $\dot{V}O_{2max}$ in both samples).

Because Method 5 appeared to yield consistently discrepant estimates, the analyses of variance were repeated for the remaining eight methods, excluding Method 5. However, the variance was again found to be significant

Table I. Estimates of the ventilatory threshold, expressed as $\dot{V}O_2$ and $\%\dot{V}O_{2max}$ ($\pm$s), mean square residuals, and the number of indeterminate cases associated with each method.

	Method								
	1	2	3	4	5	6	7	8	9
Sample 1									
$\dot{V}O_2$ (litres · min^{-1})	2.60 ± 0.67	3.13 ± 0.68	2.85 ± 0.67	2.85 ± 0.65	2.47 ± 0.40	2.81 ± 0.64	2.96 ± 0.66	2.78 ± 0.70	3.03 ± 0.70
$\%\dot{V}O_{2max}$	73.84 ± 7.90	89.75 ± 8.54	80.76 ± 9.09	81.16 ± 6.85	70.99 ± 9.24	79.24 ± 7.34	83.65 ± 9.72	78.29 ± 9.75	85.34 ± 7.99
Mean square residuals	0.016 ± 0.012	0.009 ± 0.006	0.015 ± 0.011	0.014 ± 0.009	0.017 ± 0.010	0.399 ± 0.187	0.454 ± 0.496	0.004 ± 0.002	0.003 ± 0.002
Indeterminate (n)	0	0	2	0	4	0	0	0	0
Sample 2									
$\dot{V}O_2$ (litres · min^{-1})	2.68 ± 0.76	2.87 ± 0.94	2.63 ± 0.85	2.66 ± 0.67	2.37 ± 0.73	2.96 ± 0.70	3.03 ± 0.77	2.89 ± 0.72	2.95 ± 0.70
$\%\dot{V}O_{2max}$	72.69 ± 6.79	77.27 ± 10.72	70.77 ± 12.70	72.93 ± 6.43	66.77 ± 16.57	81.67 ± 7.04	82.64 ± 5.44	79.51 ± 6.89	81.18 ± 5.78
Mean square residuals	0.007 ± 0.007	0.006 ± 0.005	0.007 ± 0.006	0.008 ± 0.007	0.010 ± 0.009	0.872 ± 0.453	0.835 ± 0.413	0.003 ± 0.002	0.003 ± 0.002
Indeterminate (n)	0	0	1	0	1	0	0	0	0

(Sample 1: $F_{7,182} = 8.52$, $P < 0.001$; Sample 2: $F_{7,203} = 13.54$, $P < 0.001$). Excluding all comparisons that involved Method 5, the mean inter-method difference was reduced to 6% ($s = 4$) $\dot{V}O_{2max}$ for both samples.

Of the eight comparisons of each method with the other methods, those between Methods 3 and 7 in Sample 1 and between Methods 4 and 6 in Sample 2 yielded significantly different estimates of the ventilatory threshold. On the other hand, some methods yielded estimates that did not differ in either sample (i.e. Methods 1 vs. 5, 2 vs. 9, 3 vs. 4, 6 vs. 8, and 7 vs. 9). Of these, some differences were consistently within 2% $\dot{V}O_{2max}$ (i.e. Methods 3 vs. 4, 6 vs. 8, and 7 vs. 9).

Inter-sample rank-order comparisons

The rank order of the nine estimates of the ventilatory threshold from the two samples was not consistent, as indicated by a Spearman rho rank-order correlation coefficient of only 0.43 ($P = 0.25$). Although Method 5 yielded the lowest estimates in both samples, the methods yielding the highest estimates varied. However, Method 7 (ranked third in Sample 1 and first in Sample 2) and Method 9 (ranked second in Sample 1 and third in Sample 2) appeared at the high end of the rank order in both samples.

Correlations between methods

As shown in Table II, most of the inter-method correlations were significant. The mean inter-method correlation was 0.76 ($s = 0.12$) for Sample 1 and 0.81 ($s = 0.20$) for Sample 2. Method 5 was the only method that yielded non-significant correlations and the correlations involving this method were consistently the lowest for any of the methods. When the correlations involving Method 5 were removed, the mean inter-method correlation was raised to 0.81 ($s = 0.05$) for Sample 1 and 0.91 ($s = 0.06$) for Sample 2. The highest correlations were between Methods 1 and 8 in Sample 1 ($r = 0.90$) and between Methods 7 and 9 in Sample 2 ($r = 0.97$). The correlation coefficients were fairly consistent between the two samples. The mean difference between correlation coefficients referring to the same pair of methods in Sample 1 and Sample 2 was 0.11 ($s = 0.06$), the largest being 0.27.

Limits of agreement

An examination of all 72 Bland-Altman plots showed that the assumptions for applying the 95% limits of agreement method were met (i.e. the means and standard deviations of the inter-method differences were constant throughout the range of

Table II. Inter-method product–moment correlation coefficients.

	Method								
	1	2	3	4	5	6	7	8	9
Method 1		0.914***	0.889***	0.938***	0.449*	0.944***	0.939***	0.955***	0.951***
Method 2	0.860***		0.830***	0.934***	0.440*	0.839***	0.905***	0.885***	0.935***
Method 3	0.843***	0.802***		0.849***	0.363	0.889***	0.808***	0.856***	0.835***
Method 4	0.871***	0.890***	0.809***		0.445*	0.912***	0.949***	0.897***	0.958***
Method 5	0.521**	0.553**	0.630***	0.701***		0.452*	0.499**	0.517**	0.517**
Method 6	0.859***	0.760***	0.829***	0.799***	0.606***		0.956***	0.959***	0.928***
Method 7	0.795***	0.742***	0.703***	0.759***	0.391	0.834***		0.940***	0.967***
Method 8	0.897***	0.762***	0.713***	0.786***	0.526**	0.851***	0.885***		0.944***
Method 9	0.810***	0.727***	0.763***	0.807***	0.771***	0.855***	0.815***	0.858***	

Note: The results from Sample 1 appear below and the results from Sample 2 appear above the main diagonal.
*$p<0.05$, **$p<0.01$, ***$p<0.001$.

measurements and these differences were from an approximately normal distribution). There was no sign that the inter-method differences tended to increase as the value of the estimates increased. The 95% limits of agreement between the methods are shown in Table III. An examination of these data shows not only substantial discrepancies, but also considerable differences in the limits of agreement. In several cases, the upper or lower limit exceeded 1.0 litre $\cdot$ min^{-1} and, in comparisons involving Method 5, they even exceeded 2.0 litres $\cdot$ min^{-1}. The closest agreement, consistently across the two samples, was again between Methods 3 and 4, Methods 6 and 8, and Methods 7 and 9. In these comparisons, the mean inter-method difference was <0.1 litres $\cdot$ min^{-1} and the 95% limits of agreement were below or close to 0.8 litres $\cdot$ min^{-1}.

Discussion

The present investigation has shown that different computerized methods for determining the ventilatory threshold can yield substantially different results. In many cases, the differences were large enough to raise concerns for both research and clinical practice. Clearly, Method 5 showed the most prominent signs of divergence from the other methods, with statistically significant differences (being consistently lower than other methods), low correlations, and very broad 95% limits of agreement. Method 5 is unique among the methods tested here in that it scans the data (slopes) from left to right and stops as soon as its criterion is satisfied (the slope of the second regression through the $\dot{V}CO_2$ by $\dot{V}O_2$ data exceeding 1.0). This pattern of searching for a solution (i.e. not considering all possibilities) probably accounts for the consistently lower estimates associated with Method 5 compared with the other methods.

Even with Method 5 removed, however, most of the remaining eight methods showed discrepancies as high as 0.5 litres $\cdot$ min^{-1} or 16% $\dot{V}O_{2max}$ (Method 1 vs. Method 2) in Sample 1 and 0.4 litres $\cdot$ min^{-1} or 12% $\dot{V}O_{2max}$ (Method 3 vs. Method 7) in Sample 2. With the exception of Method 5, the most and least discrepant pairs were not consistent across the two samples, possibly indicating that, to some extent, most methods are susceptible to the idiosyncrasies of different data sets. On the other hand, there were also some signs of convergence between methods, consistently across both samples. Specifically, the discrepancies of the estimates derived from Method 3 versus Method 4 (0.4% and 2.2% $\dot{V}O_{2max}$), Method 6 versus Method 8 (0.9% and 2.2% $\dot{V}O_{2max}$), and Method 7 versus Method 9 (1.7% and 1.5% $\dot{V}O_{2max}$) were within 2.5% $\dot{V}O_{2max}$ and not statistically significant. The convergence between Methods 3 and 4 is perhaps not surprising given that both methods are based on $\dot{V}CO_2$ by $\dot{V}O_2$ data and they seek to identify the point that is farthest away either from the single regression line through the data (Method 3) or the line connecting the first and last data points (Method 4). The reasons for the close convergence between Methods 6 and 8 and between Methods 7 and 9 are less obvious, particularly since these methods are based on different physiological variables, namely $\dot{V}_E/\dot{V}O_2$ (Methods 6 and 7) and $(\dot{V}CO_2{}^2/\dot{V}O_2) - \dot{V}CO_2$ (Methods 8 and 9). What the methods do have in common is that they use the same data in the abscissa, namely $\dot{V}O_2$ (Methods 6 and 8) and time (Methods 7 and 9) and the same computational algorithm for finding the breakpoint (Jones & Molitoris, 1984). However, it is unlikely that these common elements can fully account for the consistent convergence of the methods across both samples, pointing instead to a substantive agreement.

Only two previous studies have examined multiple computerized methods for determining the ventilatory threshold, neither of which can be compared directly with the present study due to differences in the algorithms and data manipulation methods used.

Table III. Results of the 95% limits of agreement (LoA) analyses, including the mean inter-method differences and the low and high limits of agreement ($\bar{x} \pm 1.96$ s).

	Method								
	1	2	3	4	5	6	7	8	9
Method 1									
$\bar{x}$		−0.198	0.072	0.016	0.266	−0.294	−0.349	−0.222	−0.277
LoA		−0.969/0.573	−0.688/0.832	−0.505/0.536	−1.234/1.765	−0.781/0.194	−0.871/0.173	−0.660/0.217	−0.736/0.182
Method 2									
$\bar{x}$	−0.534		0.276	0.214	0.467	−0.096	−0.151	−0.024	−0.079
LoA	−1.232/0.164		−0.756/1.308	−0.553/0.980	−1.280/2.214	−0.964/0.773	−0.949/0.647	−0.907/0.859	−0.819/0.660
Method 3									
$\bar{x}$	−0.219	0.299		−0.047	0.199	−0.378	−0.422	−0.301	−0.338
LoA	−0.948/0.511	−0.533/1.131		−0.927/0.833	−1.530/1.927	−1.152/0.396	−1.411/0.568	−1.159/0.557	−1.250/0.575
Method 4									
$\bar{x}$	−0.248	0.286	0.000		0.249	−0.309	−0.365	−0.237	−0.293
LoA	−0.905/0.408	−0.327/0.899	−0.797/0.798		−1.168/1.666	−0.875/0.256	−0.849/0.119	−0.861/0.386	−0.684/0.098
Method 5									
$\bar{x}$	0.160	0.663	0.381	0.377		−0.569	−0.620	−0.495	−0.559
LoA	−0.922/1.241	−0.358/1.683	−0.602/1.364	−0.502/1.256		−2.021/0.883	−2.069/0.829	−1.873/0.884	−1.931/0.812
Method 6									
$\bar{x}$	−0.215	0.320	0.014	0.033	−0.349		−0.055	−0.072	−0.016
LoA	−0.896/0.466	−0.578/1.217	−0.726/0.753	−0.770/0.837	−1.258/0.560		−0.502/0.391	−0.473/0.329	−0.535/0.502
Method 7									
$\bar{x}$	−0.359	0.176	−0.121	−0.111	−0.510	−0.144		−0.127	−0.072
LoA	−1.188/0.470	−0.762/1.114	−1.117/0.875	−1.001/0.779	−1.654/0.633	−0.876/0.588		−0.640/0.385	−0.463/0.319
Method 8									
$\bar{x}$	−0.186	0.348	0.019	0.062	−0.370	−0.029	−0.173		−0.055
LoA	−0.801/0.428	−0.588/1.284	−0.986/1.023	−0.814/0.938	−1.505/0.766	−0.758/0.701	−0.818/0.473		−0.523/0.412
Method 9									
$\bar{x}$	−0.436	0.099	−0.238	−0.187	−0.642	0.221	0.077	−0.249	
LoA	−1262/0.390	−0.896/1.093	−1.134/0.657	−1.013/0.639	−1.434/0.150	−0.493/0.934	−0.732/0.855	−0.981/0.482	

Note: The results from Sample 1 appear below and the results from Sample 2 appear above the main diagonal. The data are shown in litres · min^{-1}.

Fukuba and colleagues (Fukuba, Munaka, Usui, & Sasahara, 1988) examined six methods. However, they did not report the results of inter-method comparisons but rather the comparisons of each method to the lactate threshold, using this marker as the ''gold standard'' criterion of the ''anaerobic threshold''. They reported that the methods yielded variable differences from the lactate threshold (from 0.015 to 0.822 litres $\cdot$ min^{-1}) and variable and modest correlations with it (from $r = 0.211$ to $r = 0.637$). More recently, Santos and Giannella-Neto (2004) compared five computerized methods with visual inspection (the average determinations of two judges). They reported that the computerized methods yielded estimates ranging from 73% to 83% $\dot{V}O_{2max}$, whereas the visual methods yielded estimates ranging from 70% to 82% $\dot{V}O_{2max}$. Despite these broad ranges, an analysis of variance showed that the differences were not statistically significant. Similarly, the inter-method product–moment correlations ranged from 0.88 to 0.96 for the computerized methods and from 0.85 to 0.97 for the visual methods. Although these results could indicate somewhat stronger agreement than that found in the present study, several differences prohibit a direct comparison (stationary cycling instead of treadmill running, breath-by-breath data instead of 20-s averages, use of unpublished algorithms, and reportedly finding a respiratory compensation point in all cases).

The present investigation was limited to comparisons between methods and, unlike other comparative evaluations of methods for determining the ventilatory threshold, lacked a comparison to a ''gold standard''. This obviously precludes determining which method is the ''best'', a question that is admittedly more interesting than whether the different computerized methods agree. However, at the present time, we see no viable alternative, since no method can legitimately be considered the ''gold standard'' for determining the ventilatory threshold.

Both criteria that have served as ''gold standards'' in previous research – namely, the lactate threshold and visual detection – have extensively documented problems that undermine their utility as validation criteria. Specifically, regarding the use of the lactate threshold as a ''gold standard'' (Caiozzo *et al.*, 1982; Dickstein *et al.*, 1990a; Fukuba *et al.*, 1988; Gaskill *et al.*, 2001; Gladden *et al.*, 1985), there is evidence that the ventilatory and lactate thresholds reflect distinct, or perhaps partly distinct, underlying mechanisms (Brooks, 1985; Myers & Ashley, 1997; Svendahl & MacIntosh, 2003). The two thresholds do not necessarily occur simultaneously and respond differently to training (Meyer *et al.*, 1996; Powers, Dodd, & Garner, 1984). Moreover, the determination of the lactate threshold itself is subject to similar challenges as those associated with the determination of the ventilatory threshold.

Using visual detection as a validation criterion is also problematic. Numerous comparative studies have established that, although the reliability of such methods is generally acceptable (Aunola & Rusko, 1984; Gladden *et al.*, 1985; Hebestreit, Staschen, & Hebestreit, 2000; Meyer *et al.*, 1996), there are other problems. These include: (a) the number of indeterminate cases (Cohen-Solal *et al.*, 1991; Fawkner, Armstrong, Childs, & Welsman, 2002; Foster, Hume, Dicinson, Chatfield, & Byrnes, 1986; Hebestreit *et al.*, 2000; Shimizu *et al.*, 1991; Simonton, Higginbotham, & Cobb, 1988); (b) low agreement between reviewers (Gladden *et al.*, 1985; Hebestreit *et al.*, 2000; Meyer *et al.*, 1996; Shimizu *et al.*, 1991; Sullivan *et al.*, 1984; Yeh, Gardner, Adams, Yanowitz, & Crapo, 1983); and (c) low agreement between determinations made on the basis of plots of different respiratory variables (Caiozzo *et al.*, 1982; Cohen-Solal *et al.*, 1991; Garrard & Das, 1987; Shimizu *et al.*, 1991). Perhaps not surprisingly, studies that have compared visual methods with a computer algorithm, such as the V-slope, have shown poor agreement (Dickstein *et al.*, 1990a; Fawkner *et al.*, 2002; Gladden *et al.*, 1985).

If declaring a ''best'' computerized method is unattainable at this stage, what is the novel contribution of the present investigation and where do we go from here? To our knowledge, ours is the first study to compare multiple, published, computerized, regression-based methods for determining the ventilatory threshold. The significant variance found between methods suggests that reliance on any *single* computerized method might be imprudent. Single, regression-based computer algorithms, such as those incorporated in commercial metabolic analysis software programs, should be viewed as useful aides but not capable of providing ''automatic'' or definitive solutions. Similarly, stating that the ventilatory threshold was determined by a ''computer algorithm'' without providing additional quality-control assurances, as seen occasionally in published reports, should not be regarded as adequate.

On the basis of the results reported here, we join the chorus of voices calling for a systematic effort towards the development of a standard protocol for determining the ventilatory threshold, accompanied by a set of universally agreed and explicitly defined quality-control criteria (e.g. Gaskill *et al.*, 2001; Meyer *et al.*, 1996). A standard protocol, although difficult to achieve, could bring much-needed consistency in this notoriously inconsistent area of investigation. Such a standardization effort could benefit greatly from the experience gained in previous undertakings that faced similar, or greater, challenges. A prominent example is the international

project ''Common Standards for Quantitative Electrocardiography'' (Willems *et al*., 1987, 1990), which dealt with computerized methods of detecting a variety of pathologies in the electrocardiogram. The establishment of an expert panel, working under the auspices of a respected scientific organization, is the necessary first step. The development of an extensive, internet-accessible reference data base, against which computerized methods could be validated, would probably be a reasonable next step. Ultimately, as several researchers have predicted (e.g. Gaskill *et al*., 2001; Kara, Gökbel, & Bediz, 1999), the protocol that will emerge will probably combine multiple algorithms. In this sense, the finding of consistent convergence between methods across two samples (Method 3 vs. Method 4, Method 6 vs. Method 8, Method 7 vs. Method 9) in the present investigation could aid in the development of such a protocol.

The limitations of the present study include the following. First, the generalizability of the findings is limited by the characteristics of the participants. It is possible that the results might tend to be more consistent for young and healthy individuals. For example, previous studies have shown that the determination of the ventilatory threshold in certain populations, such as children (Fawkner *et al*., 2002; Hebestreit *et al*., 2000), elderly women (Foster *et al*., 1986), and chronic heart failure patients (Meyer *et al*., 1996; Simonton *et al*., 1988), can present additional challenges. Second, the results may be specific to treadmill exercise and the incremental protocols used in this investigation. Ventilatory data collected during other exercise modalities may exhibit different characteristics (e.g. signal to noise ratio) and may, therefore, behave differently when analysed by the computational algorithms examined here. Moreover, in the absence of test–retest information on the test protocols, whether the results would be reliable over repeated test sessions is unclear. Third, even though the size of our samples was larger than most previous studies comparing different methods of determination of the ventilatory threshold, samples larger than 30 could further improve the stability of estimates. Fourth, the computerized methods examined in the present study are all based on various types of regression. However, the literature also contains methods that are based on fundamentally different approaches, as diverse as time series with hidden Markov chains, smoothing splines, neuro-fuzzy logic, and neural networks. These methods are still at an early experimental stage, their complex computational details remain largely undisclosed, and they have yet to find their way into commercially available software. Yet, given the continuing absence of a universal standard method based on regression techniques, the dissemination and testing of these more computationally advanced methods should be encouraged.

References

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

Beaver, W. L., Wasserman, K., & Whipp, B. J. (1986). A new method for detecting anaerobic threshold by gas exchange. *Journal of Applied Physiology*, *60*, 2020–2027.

Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, *8476*, 307–310.

Bland, J. M., & Altman, D. G. (2003). Applying the right statistics: Analyses of measurement studies. *Ultrasound in Obstetrics and Gynecology*, *22*, 85–93.

Brooks, G. A. (1985). Anaerobic threshold: Review of the concept and directions for future research. *Medicine and Science in Sports and Exercise*, *17*, 22–31.

Caiozzo, V. J., Davis, J. A., Ellis, J. F., Azus, J. L., Vandagriff, R., Prietto, C. A., *et al*. (1982). A comparison of gas exchange indices used to detect the anaerobic threshold. *Journal of Applied Physiology*, *53*, 1184–1189.

Cheng, B., Kuipers, H., Snyder, A. C., Keizer, H. A., Jeukendrup, A., & Hesselink, M. (1992). A new approach for the determination of ventilatory and lactate thresholds. *International Journal of Sports Medicine*, *13*, 518–522.

Cohen-Solal, A., Zannad, F., Kayanakis, J. G., Gueret, P., Aupetit, J. F., & Kolsky, H. (1991). Multicentre study of the determination of peak oxygen uptake and ventilatory threshold during bicycle exercise in chronic heart failure: Comparison of graphical methods, interobserver variability and influence of the exercise protocol. *European Heart Journal*, *12*, 1055–1063.

Dickstein, K., Barvik, S., Aarsland, T., Snapinn, S., & Karlsson, J. (1990a). A comparison of methodologies in detection of the anaerobic threshold. *Circulation*, *81*, II38–II46.

Dickstein, K., Barvik, S., Aarsland, T., Snapinn, S., & Millerhagen, J. (1990b). Validation of a computerized technique for detection of the gas exchange anaerobic threshold in cardiac disease. *American Journal of Cardiology*, *66*, 1363–1367.

Fawkner, S. G., Armstrong, N., Childs, D. J., & Welsman, J. R. (2002). Reliability of the visually identified ventilatory threshold and V-slope in children. *Pediatric Exercise Science*, *14*, 181–192.

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

Fukuba, Y., Munaka, M., Usui, S., & Sasahara, H. (1988). Comparison of objective methods for determining ventilatory threshold. *Japanese Journal of Physiology*, *38*, 133–144.

Garrard, C. S., & Das, R. (1987). Sources of error and variability in the determination of anaerobic threshold in healthy humans. *Respiration*, *51*, 137–145.

Gaskill, S. E., Ruby, B. C., Walker, A. J., Sanchez, O. A., Serfass, R. C., & Leon, A. S. (2001). Validity and reliability of combining three methods to determine ventilatory threshold. *Medicine and Science in Sports and Exercise*, *33*, 1841–1848.

Gitt, A. K., Wasserman, K., Kilkowski, C., Kleemann, T., Kilkowski, A., Bangert, M., *et al*. (2002). Exercise anaerobic threshold and ventilatory efficiency identify heart failure patients for high risk of early death. *Circulation*, *106*, 3079–3084.

Gladden, L. B., Yates, J. W., Stremel, R. W., & Stamford, B. A. (1985). Gas exchange and lactate anaerobic thresholds: Inter- and intraevaluator agreement. *Journal of Applied Physiology*, *58*, 2082–2089.

Hebestreit, H., Staschen, B., & Hebestreit, A. (2000). Ventilatory threshold: A useful method to determine aerobic fitness in children? *Medicine and Science in Sports and Exercise*, *32*, 1964–1969.

Jones, R. H., & Molitoris, B. A. (1984). A statistical method for determining the breakpoint of two lines. *Analytical Biochemistry*, *141*, 287–290.

Kara, M., Gökbel, H., & Bediz, C. S. (1999). A combined method for estimating ventilatory threshold. *Journal of Sports Medicine and Physical Fitness*, *39*, 16–19.

Meyer, K., Hajric, R., Westbrook, S., Samek, L., Lehmann, M., Schwaibold, M., *et al.* (1996). Ventilatory and lactate threshold determinations in healthy normals and cardiac patients: Methodological problems. *European Journal of Applied Physiology*, *72*, 387–393.

Meyer, T., Lucia, A., Earnest, C. P., & Kindermann, W. (2005). A conceptual framework for performance diagnosis and training prescription from submaximal gas exchange parameters: Theory and application. *International Journal of Sports Medicine*, *26* (suppl. 1), S38–S48.

Myers, J., & Ashley, E. (1997). Dangerous curves: A perspective on exercise, lactate, and the anaerobic threshold. *Chest*, *111*, 787–795.

Older, P., Hall, A., & Hader, R. (1999). Cardiopulmonary exercise testing as a screening test for perioperative management of major surgery in the elderly. *Chest*, *116*, 355–362.

Orr, G. W., Green, H. J., Hughson, R. L., & Bennett, G. W. (1982). A computer linear regression model to determine ventilatory anaerobic threshold. *Journal of Applied Physiology*, *52*, 1349–1352.

Powers, S. K., Dodd, S., & Garner, D. R. (1984). Precision of ventilatory and gas exchange alterations as a predictor of the anaerobic threshold. *European Journal of Applied Physiology*, *52*, 173–177.

Santos, E. L., & Giannella-Neto, A. (2004). Comparison of computerized methods for detecting the ventilatory thresholds. *European Journal of Applied Physiology*, *93*, 315–324.

Schneider, D. A., Phillips, S. E., & Stoffolano, S. (1993). The simplified V-slope method of detecting the gas exchange threshold. *Medicine and Science in Sports and Exercise*, *25*, 1180–1184.

Shimizu, M., Myers, J., Buchanan, N., Walsh, D., Kraemer, M., McAuley, P., *et al.* (1991). The ventilatory threshold: Method, protocol, and evaluator agreement. *American Heart Journal*, *122*, 509–516.

Simonton, C. A., Higginbotham, M. B., & Cobb, F. R. (1988). The ventilatory threshold: Quantitative analysis of reproducibility and relation to arterial lactate concentration in normal subjects and in patients with congestive heart failure. *American Journal of Cardiology*, *62*, 100–107.

Sue, D. Y., Wasserman, K., Moricca, R. B., & Casaburi, R. (1988). Metabolic acidosis during exercise in patients with chronic obstructive pulmonary disease: Use of the V-slope method for anaerobic threshold determination. *Chest*, *94*, 931–938.

Sullivan, M., Genter, F., Savvides, M., Roberts, M., Myers, J., & Froelicher, V. (1984). The reproducibility of hemodynamic, electrocardiographic, and gas exchange data during treadmill exercise in patients with stable angina pectoris. *Chest*, *86*, 375–382.

Svendahl, K., & MacIntosh, B. R. (2003). Anaerobic threshold: The concept and methods of measurement. *Canadian Journal of Applied Physiology*, *28*, 299–323.

Wasserman, K. (1997). Diagnosing cardiovascular and lung pathophysiology from exercise gas exchange. *Chest*, *112*, 1091–1101.

Willems, J. L., Arnaud, P., van Bemmel, J. H., Bourdillon, P. J., Degani, R., Denis, B., *et al.* (1987). A reference data base for multilead electrocardiographic computer measurement programs. *Journal of the American College of Cardiology*, *10*, 1313–1321.

Willems, J.L., Arnaud, P., van Bemmel, J. H., Degani, R., Macfarlane, P. W., & Zywietz, C. (1990). Common standards for quantitative electrocardiography: Goals and main results. *Methods of Information in Medicine*, *29*, 263–271.

Yeh, M. P., Gardner, R. M., Adams, T. D., Yanowitz, F. G., & Crapo, R. O. (1983). ''Anaerobic threshold'': Problems of determination and validation. *Journal of Applied Physiology*, *55*, 1178–1186.