

The validity of the Moxus Modular metabolic system during incremental exercise tests: impacts on detection of small changes in oxygen consumption

Fernando G. Beltrami · Christian Froyd ·
Asgeir Mamen · Timothy D. Noakes

Received: 9 September 2013 / Accepted: 17 January 2014 / Published online: 30 January 2014
© Springer-Verlag Berlin Heidelberg 2014

Abstract

Purpose We investigated the accuracy of the Moxus Modular Metabolic System (MOXUS) against the Douglas Bag Method (DBM) during high-intensity exercise, and whether the two methods agreed when detecting small changes in $\dot{V}O_2$ between two consecutive workloads ($\Delta\dot{V}O_2$).

Methods Twelve trained male runners performed two maximal incremental running tests while gas exchange was analyzed simultaneously by the two systems using a serial setup for four consecutive intervals of 30 s on each test. Comparisons between methods were performed for $\dot{V}O_2$, \dot{V}_E , fractions of expired O_2 (FeO_2) and CO_2 ($FeCO_2$) and $\Delta\dot{V}O_2$.

Results The MOXUS produced significant higher (mean \pm SD, $n = 54$) readings for $\dot{V}O_2$ ($80 \pm 200 \text{ mL min}^{-1}$, $p = 0.005$) and \dot{V}_E ($2.9 \pm 4.2 \text{ L min}^{-1}$, $p < 0.0001$), but not FeO_2 (-0.01 ± 0.09). Log-transformed 95 % limits of agreement for readings between methods were 94–110 % for $\dot{V}O_2$, 97–108 % for \dot{V}_E and 99–101 % for FeO_2 . $\Delta\dot{V}O_2$ for two consecutive measurements was not different

between systems (120 ± 110 vs. $90 \pm 190 \text{ mL min}^{-1}$ for MOXUS and DBM, respectively, $p = 0.26$), but agreement between methods was very low ($r = 0.25$, $p = 0.12$).

Discussion Although it was tested during high-intensity exercise and short sampling intervals, the MOXUS performed within the acceptable range of accuracy reported for automated analyzers. Most of the differences between equipments were due to differences in \dot{V}_E . Detecting small changes in $\dot{V}O_2$ during an incremental test with small changes in workload, however, might be beyond the equipment's accuracy.

Keywords Criterion validity · Douglas Bag method · Gas exchange · Oxygen uptake · Moxus Modular metabolic system

Abbreviations

$\Delta\dot{V}O_2$	Difference in oxygen uptake between two consecutive 30-s intervals
DBM	Douglas Bag method
$FeCO_2$	Fraction of expired carbon dioxide
FeO_2	Fraction of expired oxygen
RER	Respiratory exchange ratio
$\dot{V}CO_2$	Carbon dioxide output
\dot{V}_E	Minute ventilation
$\dot{V}O_2$	Oxygen uptake
$\dot{V}O_{2\text{max}}$	Maximal oxygen uptake

Communicated by Jean-René Lacour.

F. G. Beltrami (✉) · C. Froyd · T. D. Noakes
UCT/MRC Research Unit for Exercise Science and Sports
Medicine, Department of Human Biology, University of Cape
Town and Sports Science Institute of South Africa, Boundary
Road, Newlands 7700, South Africa
e-mail: fernando.beltrami@uct.ac.za

C. Froyd
Faculty of Teacher Education and Sport, Sogn og Fjordane
University College, Sogndal, Norway

A. Mamen
University College of Health Sciences, Campus Kristiania,
Oslo, Norway

Introduction

The first measurements of oxygen consumption ($\dot{V}O_2$) during exercise date back to the 1910s, and the concept of maximal oxygen consumption ($\dot{V}O_{2\text{max}}$) is one of the traditional cornerstones of exercise physiology (Hodges

et al. 2005). $\dot{V}O_{2\max}$ can be defined as the maximal rate at which the body is able to utilize oxygen during exercise, and is thought to be the best measure of cardiorespiratory fitness (Meyer et al. 2005). It is also believed that $\dot{V}O_{2\max}$ is reached when the $\dot{V}O_2$ “plateaus” despite a continually increasing workload (Taylor et al. 1955).

These earlier studies reporting the occurrence of the plateau phenomenon used discontinuous exercise protocols, often performed over multiple days, and the exercise protocols required large changes in $\dot{V}O_2$ (~300 mL O_2) between subsequent workloads (Taylor et al. 1955). During these trials, expired gas was collected in meteorological balloons (now termed the Douglas Bag) and subsequently analyzed with very laborious and time-consuming methods (Mitchell et al. 1958). Over time, the $\dot{V}O_{2\max}$ test protocol evolved (from a discontinuous to a continuous test with shorter stages) and the Douglas Bag method (DBM) was replaced by automated gas analyzers. However, the DBM is still considered the gold standard for $\dot{V}O_2$ measurement. The use of higher resolution gas analyzers combined with continuous ramp-like protocols has significantly decreased the incidence of the plateau phenomenon during maximal exercise tests (Astorino 2009; Doherty et al. 2003).

The ability to correctly detect a plateau in $\dot{V}O_2$ during a single incremental test with short stages and small increments in workload largely depends on the capacity of the equipment used to measure the different gas exchange components of $\dot{V}O_2$. While numerous studies have validated gas analyzers from different manufacturers (Bassett et al. 2001; Crouter et al. 2006; Duffield et al. 2004; La Mere et al. 1993; Medbo et al. 2002; Rietjens et al. 2001; Rosdahl et al. 2013), these studies are often performed at low to moderate exercise intensities (Bassett et al. 2001; Macfarlane and Wu 2013) when using long averaging intervals (Bassett et al. 2001; Crouter et al. 2006; Macfarlane and Wu 2013). This differs from the way in which the analyzers are used in practice. For example, these systems may be unable to accurately measure pulmonary ventilation (\dot{V}_E) at high rates and to match those high rates with rapidly changing fractions of expired O_2 and CO_2 (Gore et al. 2003). Other problems with validation studies include the lack of consensus as to what is acceptable in terms of measurement error for $\dot{V}O_2$ (Macfarlane 2001) and the use of inadequate statistics to evaluate agreement between systems (Atkinson et al. 2005; Atkinson and Nevill 1998). For example, Bassett et al. (2001) found significant differences in $\dot{V}O_2$ readings between a gas analysis system and the DBM (18 mL \min^{-1}) along with very small 95 % limits of agreement, while others (Auchincloss et al. 1971) found no significant difference in $\dot{V}O_2$ readings between two different sets of equipment when the mean \pm standard deviation between methods was 1 ± 9 %, which would

produce very large limits of agreement (Auchincloss et al. 1971). Although the *t* test results could at first indicate that the latter study showed better agreement, in fact the Parvo Medics equipment used by Bassett et al. (2001) had much better accuracy, as given by the very small systematic bias and narrow limits of agreement. Narrow limits of agreement allow unraveling even very small systematic differences, while large limits of agreement may mask even large systematic differences as “random variation”. When only indicators of systematic bias such as *t* test or analysis of variance (La Mere et al. 1993; Wilmore et al. 1976) are reported, little is known about the range of errors between equipments. Therefore, the accuracy of gas analyzer systems should always be tested including some measurement of error, such as the Technological Error of Measurement or limits of agreement. The latter are now widely reported in validation studies, though the technique still causes some confusion. A method can be considered valid if the range determined by the limits of agreement is of no clinical significance (Bland and Altman 1999), not if 95 % of all the data fall within the limits of agreement, as it has been interpreted (Duffield et al. 2004).

Therefore, part of the inconsistent observations of the plateau phenomenon may result not only from different biological responses (Doherty et al. 2003) and the manner in which the respiratory data are analyzed (Midgley and Carroll 2009), but also as a consequence of the inability of automated gas analyzers to detect small changes in $\dot{V}O_2$ at very high \dot{V}_E rates. To test this possibility, the purpose of this investigation was to compare the agreement of respiratory data values obtained using an automated gas analyzer to those obtained simultaneously using the DBM as a criterion method during a short-stage incremental test to exhaustion and to evaluate the practical implications of those measurement errors in the calculation of stage-to-stage differences in $\dot{V}O_2$.

Methods

Subjects

Twelve trained male runners (age 25.5 ± 8.3 years, height 179.9 ± 6.6 cm, body mass 75.1 ± 5.4 kg) were recruited for this investigation. All participants were injury free at the time of the study and were able to exercise to their personal best on a treadmill. A minimum running training volume of at least 40 km per week during the 3 months prior to the study was considered as an inclusion criterion for participation. The subjects were fully informed, verbally and in print, of the risks, benefits and requirements of participation, and were required to sign an informed consent form before taking part in the study, which was approved by the

Research and Ethics Committee of the Faculty of Health Sciences of the University of Cape Town, in accordance to the Declaration of Helsinki. All participants were also clearly informed that they could withdraw from the study at any time if they so wished.

Study design

Participants visited the laboratory on two different occasions, separated by 2–3 days. On day 1, after signing the consent form, participants underwent anthropometric measurements (weight and height) and subsequently performed a maximal incremental exercise test on the treadmill (see description below). On day 2, participants again performed an incremental exercise test on a treadmill. During both days, gas exchange variables were continuously monitored throughout the test using an automated gas analyzer (MOXUS Modular Metabolic System, AEI Technologies, IL, USA). During a 2-min segment of the test expired gas was also collected into Plastic Bags (Douglas Bags, DBs) for later analysis of gas volume and composition. The DBM system was attached in series at the exit port of the Moxus System mixing chamber, so that simultaneous analysis of gas exchange parameters could be performed. This method/system does not introduce problems with increased dead space, as may be the case for other setups (Medbo et al. 2012).

Exercise test

On both days, the participants performed an incremental exercise test to exhaustion on a motor-driven treadmill (PPM 55, Woodway GmbH, Germany). Prior to the test, participants were requested to warm-up by running three 5-min stages at 8, 10 and 12 km h⁻¹, respectively. After the third 5-min stage, the participants rested for 3 min before beginning the maximal test. For the first stage of the test, subjects ran at 12 km h⁻¹ for 1 min. Thereafter the speed was increased by 0.5 km h⁻¹ every 30 s, until subjects reached volitional exhaustion. The treadmill grade was kept at 5 or 10 % during the entire test, depending on the fitness level of each participant.

Data collection and analysis

The automated gas analyzer, Moxus Modular Metabolic System, consists of a mask incorporating a turbine for determination of ventilation volume. The turbine is attached on the inspiratory side of the valve, thus avoiding problems related to condensation of water vapor and accumulation of saliva (Bassett et al. 2001). The mask covered the subject's nose and mouth so that they inhale the room air. All the expired gas traveled through a tube from

the mask into a 4.2 L mixing chamber. Samples were constantly drawn from the mixing chamber (flow regulated to 250 mL min⁻¹), through a dual-stage nafion dryer and into the O₂ and CO₂ content analyzers (S-3A/I and CD-3A, respectively, AEI Technologies, IL, USA). The analyzers use the zirconia and non-dispersive infrared methods for O₂ and CO₂ analysis. Data from the Moxus system were averaged for each 30-s stage. Prior to the start of each testing session, the automated gas analyzer system was calibrated for gas volume and composition according to the manufacturers' instructions. Volume calibration was performed using a high-precision 3 L calibration syringe (Hans-Rudolph, Kansas City, MO, USA) and gas calibration was made using two-point calibration curve, consisting of room air and a gas of known concentration (15.00 % O₂, 6.00 % CO₂, AGA Norgas, Norway). The DBM system (Cranlea, UK) was attached in series with this equipment. This method has been used previously with good results (Medbo et al. 2012), and does not impose the technical problems of increased resistance to breathing (Crouter et al. 2006; Duffield et al. 2004) and increased dead space (Medbo et al. 2002) associated with connecting multiple breathing valves close to the mouth. All expired air for each 30 s stages was collected into separate bags and stored for subsequent analysis, which was done immediately after completion of the exercise test. Collection time was clocked to the closest 0.1 s by an experienced researcher, and the bags were opened and closed always at the end of expirations, to contain only full breathing cycles. Gas volumes in the DBM were measured with a Digital Dry Gas Meter (Harvard Apparatus, MA, US). When determining the volume of expired gas collected into each bag, the small volume of air diverted to the automated gas analyzer (approximately 125 mL per 30 s) was accounted for. Fractions of O₂ and CO₂ in the expired gas collected into the bags were analyzed using the analyzers from the Moxus system, with the air flowing through the analyzers inlet using a closed-loop method, so that the analyzed air returned into the bags.

While the Moxus system could continuously monitor gas exchange variables, the DBM could only collect four bags at a time. Therefore, on the first day the DBM was used during the initial stages of the exercise test, and on the second day the point at which gas samples collection began was determined by each subject's predicted test duration from day 1, to collect data from the final 2 min of the test.

Statistics

All analyzed data are presented using mean \pm standard deviation, and respiratory variables are expressed in STPD units. Fractions of expired O₂ (FeO₂) and CO₂ (FeCO₂) are expressed as percentage-points L⁻¹ to avoid confusion. Regression analyses between variables measured by

the two different systems and residual analyses were performed using Prism 3.0 (Graphpad Software, La Jolla, CA, USA). Values measured by the two equipments were compared by dependent *t* tests (Statistica 10; Statsoft, OK, USA). The range of recorded values was divided into thirds (lower, mid and higher range), and the differences between equipments for the different thirds compared using a one-way ANOVA with Tukey Post hoc when necessary (Prism 3.0). Statistical significance was accepted when $p < 0.05$.

A variety of statistical procedures was used to facilitate comparison between our data and other studies, as has been recommended (Atkinson and Nevill 1998; Hodges et al. 2005; Macfarlane and Wong 2012; Macfarlane and Wu 2013). The Technical Error of Measurement was defined as $\sqrt{\sum D^2/2N}$, where *D* is the difference in readings between methods and *N* is the number of pair comparisons (Macfarlane and Wong 2012), and expressed as a percent of the mean. Absolute percentage error was calculated as the mean of absolute differences in readings between methods divided by the DBM readings times 100 (Macfarlane and Wu 2013). Mean Individual Coefficient of variation was expressed as the mean of individual standard deviation between measurements divided by their mean times 100 (Macfarlane and Wu 2013). Effect sizes (Cohen's *d*), were calculated as the standard error of estimate from the regression between the DBM and Moxus readings divided by the standard deviation of the DBM readings, and interpreted according to a modified Scale for Cohen's *d* (Hopkins 2012; Macfarlane and Wu 2013). The agreement between the two systems was further evaluated using Bland–Altman Plots (Bland and Altman 1999; Hodges et al. 2005). Since heteroscedastic errors are the norm when analyzing physiological parameters (Nevill and Atkinson 1997), data were also log transformed

for calculation of limits of agreement as a ratio between measurements (Bland and Altman 2007; Nevill and Atkinson 1997; Rietjens et al. 2001). Although adjustments in the Bland–Altman can be made to adjust for multiple observations per subject, this has been shown to be of little or no significance when the number of comparison pairs is much greater than the number of observations per subject, as in our study (Bland and Altman 2007). Calculations were performed using Excel (Microsoft, WA, US) spreadsheets.

Data from three participants tested on the same days produced apparently unreliable values for $\dot{V}O_2$ and respiratory exchange ratio (RER). Upon investigation, it was discovered that the drying chamber from the Moxus system required maintenance, and therefore the data from these three participants were excluded from the study. A separate analysis showed that inclusion of these data would not significantly impact our results (not shown) (Salier et al. 2012), but the data were excluded nonetheless (Hodges et al. 2005).

Results

Comparison between the data measured simultaneously by the Moxus and DBM is presented in Table 1. All variables presented high correlation coefficient between the DBM and the Moxus (Figs. 1, 2, 3, 4, 5), ranging from 0.93 ($\dot{V}O_2$) to 0.99 ($\dot{V}E$). Overall, the Moxus produced significantly higher readings for $\dot{V}E$, $\dot{V}O_2$, $FeCO_2$, CO_2 excretion ($\dot{V}CO_2$) and RER, but not for FeO_2 . FeO_2 also displayed the smaller mean coefficient of variation (Table 2) and mean absolute percentage error between the two methods (0.3 and 0.4 %, respectively), whereas $FeCO_2$ showed the larger variation between system (mean coefficient of variation 2.75 %,

Table 1 Regression equation between two systems simultaneously measuring gas exchange, Pearson's coefficient of correlation (*r*), error of regression ($S_{y|x}$), mean \pm SD physiological values for the Douglas

Bag (DB) and Moxus systems, absolute percentage error (APE) and effect size

Variable	Regression equation	<i>r</i>	$S_{y x}$	Mean \pm SD	Mean diff. \pm SD (APE)	Effect size
$\dot{V}O_2$ (L min ⁻¹)	{Y} = 0.986 {x} + 0.141	0.93	0.20	DB Moxus 4.53 \pm 0.51* 4.60 \pm 0.54	-0.08 \pm 0.20 (3.5)	0.37
$\dot{V}CO_2$ (L min ⁻¹)	{Y} = 1.044 {x} - 0.043	0.98	0.18	DB Moxus 4.65 \pm 0.87 [†] 4.81 \pm 0.93	-0.16 \pm 0.18 (3.9)	0.20
$\dot{V}E$ (L min ⁻¹)	{Y} = 1.061 {x} - 5.202	0.99	3.97	DB Moxus 132.08 \pm 25.13 [†] 134.96 \pm 26.95	-2.88 \pm 4.22 (2.6)	0.15
FeO_2 (%)	{Y} = 0.986 {x} + 0.214	0.98	0.09	DB Moxus 16.64 \pm 0.51 16.64 \pm 0.51	0.01 \pm 0.09 (0.4)	0.17
$FeCO_2$ (%)	{Y} = 1.084 {x} - 0.263	0.98	0.08	DB Moxus 4.31 \pm 0.39* 4.41 \pm 0.38	-0.10 \pm 0.08 (2.5)	0.21
RER	{Y} = 0.936 {x} + 0.082	0.98	0.02	DB Moxus 1.02 \pm 0.10 [†] 1.04 \pm 0.10	-0.02 \pm 0.02 (2.1)	0.22

Significantly different means by paired *t* tests at * $p < 0.01$, [†] $p < 0.0001$

Fig. 1 (Left) Fraction of expired O_2 measured by the Moxus versus the DB System. The solid line represents the regression line, the dotted line represents the 95 % confidence interval of regression and the dashed line represents the identity line. (Right) Bland–Altman plot of the error scores. Dashed line represents the mean bias and dotted lines represent the 95 % limits of agreement

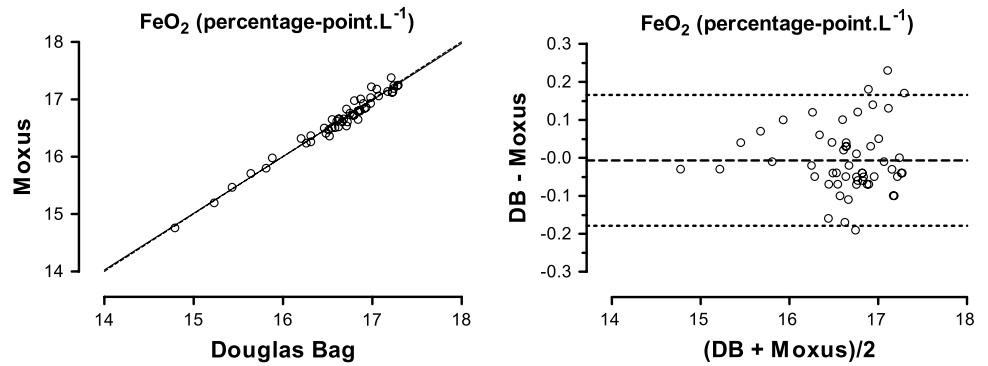


Fig. 2 (Left) Fraction of expired CO_2 measured by the Moxus versus the DB System. The solid line represents the regression line, the dotted line represents the 95 % confidence interval of regression and the dashed line represents the identity line. (Right) Bland–Altman plot of the error scores. Dashed line represents the mean bias and dotted lines represent the 95 % limits of agreement

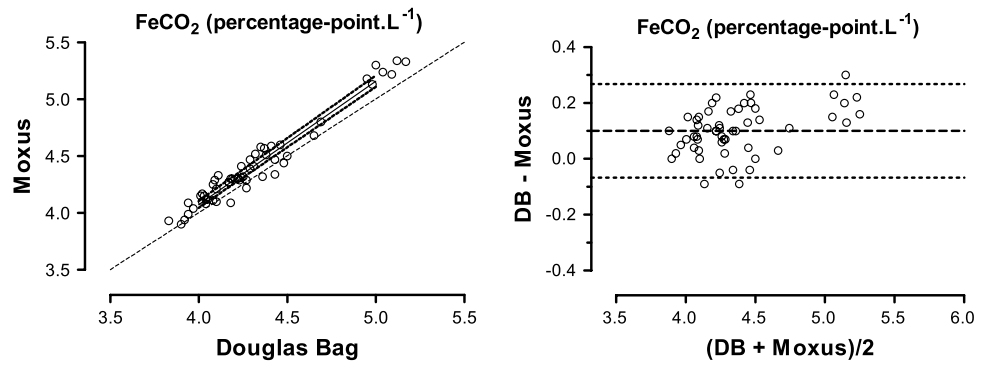


Fig. 3 (Left) Pulmonary ventilation measured by the Moxus versus the DB System. The solid line represents the regression line, the dotted line represents the 95 % confidence interval of regression and the dashed line represents the identity line. (Right) Bland–Altman plot of the error scores. Dashed line represents the mean bias and dotted lines represent the 95 % limits of agreement

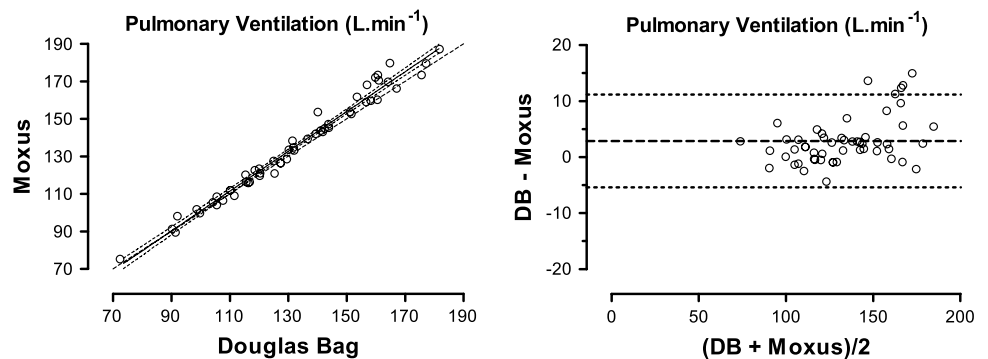
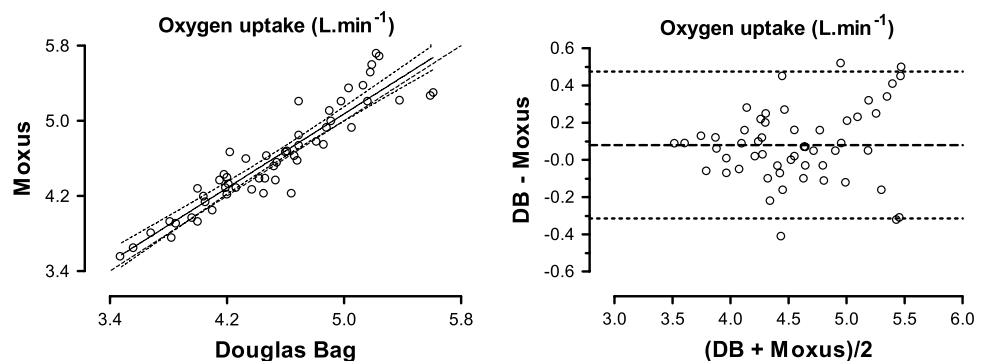


Fig. 4 (Left) Oxygen uptake measured by the Moxus versus the DB System. The solid line represents the regression line, the dotted line represents the 95 % confidence interval of regression and the dashed line represents the identity line. (Right) Bland–Altman plot of the error scores. Dashed line represents the mean bias and dotted lines represent the 95 % limits of agreement



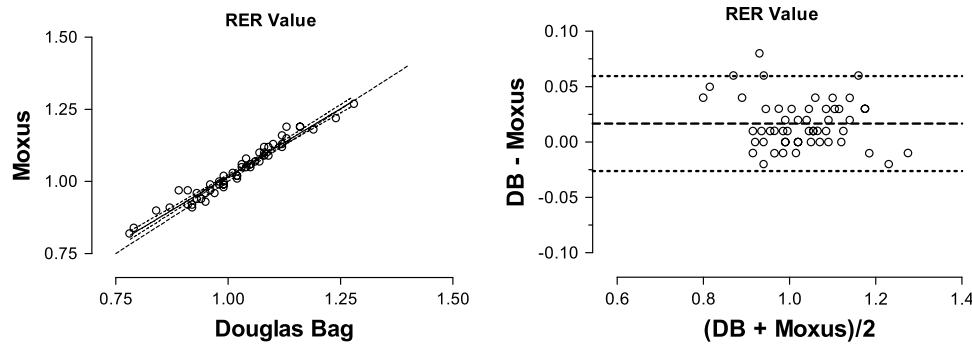


Fig. 5 (Left) R value measured by the Moxus versus the Douglas Bag (DB) System. The solid line represents the regression line, the dotted line represents the 95 % confidence interval of regression and the

dashed line represents the identity line. (Right) Bland–Altman plot of the error scores. Dashed line represents the mean bias and dotted lines represent the 95 % limits of agreement

Table 2 Mean \pm SD bias, 95 % CI of mean bias, 95 % limits of agreement (95 % LOA), technical error of measurement (TEM) and mean individual coefficient of variation (CV) between two systems measuring gas exchange simultaneously

Variable	Bland–Altman bias Mean \pm SD (95 % CI)	95 % LOA (log-transformed LOA)	Inter-device TEM (%)	CV mean (range)
$\dot{V}O_2$ (L min ⁻¹)	0.08 \pm 0.20 (0.02, 0.13)	-0.31, 0.47 (0.94, 1.10)	3.33	2.5 (0, 7.43)
$\dot{V}CO_2$ (L min ⁻¹)	0.16 \pm 0.18 (0.11, 0.21)	-0.20, 0.53 (0.96, 1.11)	3.67	2.75 (0.02, 8.25)
\dot{V}_E (L min ⁻¹)	2.88 \pm 4.22 (1.73, 4.03)	-5.40, 11.16 (0.97, 1.08)	2.69	1.81 (0.07, 6.56)
FeO ₂ *	-0.08 \pm 0.09 (-0.03, 0.02)	-0.18, 0.16 (0.99, 1.01)	0.37	0.30 (0.01, 0.96)
FeCO ₂ *	0.10 \pm 0.08 (0.08, 0.12)	-0.07, 0.27 (0.99, 1.06)	2.11	1.03 (0, 4.05)
RER	0.02 \pm 0.02 (0.01, 0.02)	-0.03, 0.06 (0.97, 1.06)	1.88	1.44 (0, 6.08)

FeO₂ and FeCO₂ are expressed in percentage points

absolute percentage error 3.9 %). Although mean individual coefficient of variation for $\dot{V}O_2$ was 2.5 %, individual values ranged from 0 to 7.43 %. The effect sizes of the standard errors of estimate were deemed “trivial” (0–0.20) according to the Cohen scale, except for $\dot{V}O_2$, RER and FeCO₂, which were deemed “small” (0.21–0.60) (Hopkins, 2012b).

The analysis of variance performed to compare the differences in readings over the lower, middle and higher third of recorded values showed no significant differences for RER, $\dot{V}O_2$ and FeO₂, showing that there was no increase in bias as values increased. Ventilation readings for the higher third of the analyzed range showed higher differences in reading between the two systems than both the lower and middle ranges of values (5.4 \pm 5.4 vs. 1.3 \pm 2.5 vs. 2.0 \pm 3.3 L min⁻¹ higher for Moxus than DBM, respectively, $p < 0.05$). Similar patterns were seen for FeCO₂ and $\dot{V}O_2$ (0.4 \pm 0.3 vs. 0.1 \pm 0.2 vs. 0.1 \pm 0.2 L min⁻¹ for the higher, middle and lower range of values, respectively, $p < 0.01$).

The mean bias and 95 % limits of agreement for the respiratory values are showed in Table 2 and also in Figs. 1, 2, 3, 4, 5. While most variables showed very small systematic bias between systems, some of the respiratory parameters showed considerably large 95 % limits of agreement. The technical error of measurement was larger for $\dot{V}O_2$ than all other variables, which is expected since $\dot{V}O_2$ accumulates the error in measurements from \dot{V}_E , FeCO₂ and FeO₂.

Although there were no significant differences between equipments in the measured $\dot{V}O_2$ difference between two consecutive 30-s intervals (0.12 \pm 0.11 vs. 0.09 \pm 0.19 L min⁻¹ for the Moxus and DBM, respectively, $p = 0.26$), the standard deviation for the difference between the systems was several times greater than the difference itself (0.03 \pm 0.20 L min⁻¹). As shown in Fig. 6a, b, the level of agreement between the two systems when measuring stage-to-stage differences in $\dot{V}O_2$ was very low ($r = 0.25$, $p = 0.12$), and the slope for the regression line between the Moxus and the DBM was not different from zero, further indicating low agreement.

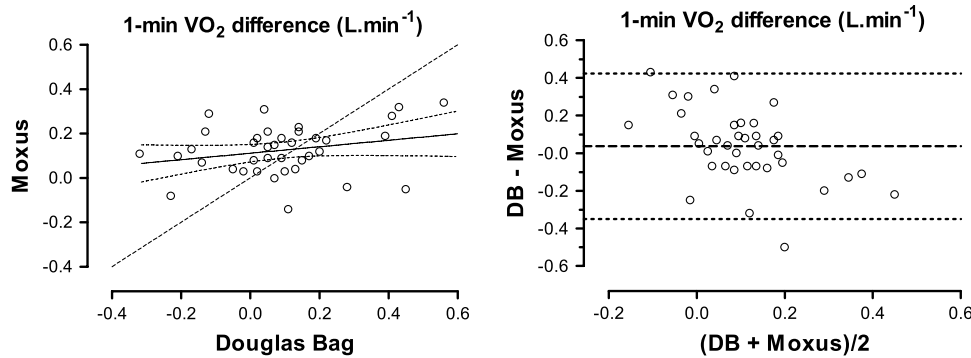


Fig. 6 (Left) Difference in O₂ uptake between two consecutive 30-s intervals measured by the Moxus versus the Douglas Bag (DB) System. The parameters for the regression line are as follows: $\{Y\} = -0.042 \{x\} + 0.072$; $s_b = 0.334$; $S_{Y|x} = 0.109$ L min⁻¹; $r = 0.253$ ($p = 0.124$); $n = 38$. The solid line represents the regres-

sion line, the dotted line represents the 95 % confidence interval of regression and the dashed line represents the identity line. (Right) Bland–Altman plot of the error scores. Dashed line represents the mean bias and dotted lines represent the 95 % limits of agreement

Discussion

To our knowledge, this is one of few studies testing the accuracy of a modern gas analyzer in more practical and challenging conditions (Meyer et al. 2005). The first important finding of this investigation is that the Moxus performed similarly to several other commercially available systems and within recommended standards for validity of gas analysis systems. The second main finding is that even though the Moxus system performed similarly to other commercially available equipments, the agreement between the two methods when comparing $\Delta\dot{V}O_2$ for two consecutive 30-s interval was very low.

Although there is no consensus about what constitutes an acceptable error in measurement for automated gas analyzers, some standards have been proposed (Atkinson et al. 2005; Hodges et al. 2005; Macfarlane 2001). Atkinson et al. (2005) suggest that there is no practical impact on $\dot{V}O_2$ measurements if two methods disagree systematically by no more than ± 2 mL kg⁻¹ min⁻¹, as judged by the 95 % confidence interval (95 % CI) of the mean bias. This was achieved in our study (-1.8 to -0.3 mL kg⁻¹ min⁻¹). MacFarlane (2001) recommends that the technical error of measurement should be smaller than 3 % for $\dot{V}O_2$ and 5 % for \dot{V}_E , which is also supported by Hodges et al. (Atkinson et al. 2005). Our data for \dot{V}_E were within the recommended range (2.69 %), but slightly off for $\dot{V}O_2$ (3.33 %).

When comparing different methods to measure a given physiological variable, the total error in measurement can be split into technological errors and biological variation, which is the normal fluctuation of a variable over the time between measurements (Hopkins 2012). The technological error of measurement can be separated from the biological variations in $\dot{V}O_2$ by testing two systems working simultaneously, and this should be used whenever possible

(Atkinson et al. 2005; Hopkins 2012; Macfarlane 2001). Since the measurement errors should be much smaller when systems work at the same time (Macfarlane and Wu 2013), we will focus our discussion on studies that used similar methodology to ours. More detailed reviews on the performance of gas analyzers tested sequentially or over multiple days can be found elsewhere (Atkinson et al. 2005; Macfarlane 2001).

Validity studies comparing gas analyzers to the DBM have reported results similar to ours. Rietjens et al. (2001) reported wider limits of agreement than those in this study for $\dot{V}O_2$ (ratio of DBM over Oxycon between 0.88 and 1.12 vs. 0.94 and 1.10 in our study) when validating the Oxycon Pro. Jakovljevic et al. (2008) found wider limits of agreement for the Innocor system against the DBM than in our investigation for both $\dot{V}O_2$ and \dot{V}_E (-0.52 to 0.55 L min⁻¹ for $\dot{V}O_2$ and -8.74 and 10.66 L min⁻¹ for \dot{V}_E), though the authors interpreted their values as too wide to be acceptable for clinical testing. Reybrouck et al. (1992) presented limits of agreement for $\dot{V}O_2$ in the range of ± 3.5 and ± 3.0 L min⁻¹ for \dot{V}_E in a study validating an automated system, which are narrower values than those reported in the present investigation. It is important to note, however, that their rate of \dot{V}_E did not exceed 40 L min⁻¹: a very different range of exercise intensities than that in the present investigation.

Crouter et al. reported 95 % limits of agreement for $\dot{V}O_2$ of ± 200 mL min⁻¹ and \dot{V}_E of ± 2.3 L min⁻¹, but like the former study minute ventilation rarely exceeded 80 L min⁻¹, and data were averaged over 2-min intervals. The study from Gore et al. (2003) clearly demonstrates how differences between methods often grow proportionally to the size of the mean of measurements, which further identifies the importance of testing equipment over the full spectrum of physiological values normally measured.

The presence of heteroscedastic errors in physiological variables has already been examined (Atkinson and Nevill 1998) and is further shown by the larger difference between equipments for the higher third of measurements of \dot{V}_E , $\dot{V}CO_2$ FeO₂. In our study, \dot{V}_E values ranged from ~70 to ~190 L min⁻¹, while $\dot{V}O_2$ values ranged between ~3.3 and 5.5 L min⁻¹, values much higher than often reported (Bassett et al. 2001; Macfarlane and Wu 2013; Reybrouck et al. 1992) in validation studies. A much larger study on the Oxycon Pro revealed that almost all data points that fell outside the 95 % limits of agreement range were measured at \dot{V}_E above 120 L min⁻¹ and $\dot{V}O_2$ of 3.8 L min⁻¹, intensities that are seldom tested in validity studies.

Minute ventilation is often considered the main source of error when measuring gas exchange parameters, especially since it directly influences $\dot{V}O_2$ (that is, a 1 % change in \dot{V}_E will change $\dot{V}O_2$ by 1 %) (Carter and Jeukendrup 2002; Hodges et al. 2005; Medbo et al. 2012; Rosdahl et al. 2013). Although the DBM is considered the gold standard for measuring gas exchange parameters, it also has potential for error, especially when measuring \dot{V}_E . Carter and Jeukendrup (2002) used a dry gas meter similar to that used by us, and concluded that the equipment was slightly inaccurate, since the DBM consistently resulted in lower \dot{V}_E values than both an Oxycon Pro, an Oxycon Alpha and more importantly a metabolic simulator. This conclusion is shared by others (Rietjens et al. 2001). Conversely, Macfarlane and Wong (2012) also used the same dry gas meter system and found similar \dot{V}_E values for the Oxycon Pro and the DBM. Another potential problem when measuring $\dot{V}O_2$ and \dot{V}_E with the DBM is the timing for opening and closing the bags (Rietjens et al. 2001; Wilmore et al. 1976). In this regard, automated gas analyzer can measure the start and end of breathing cycles more accurately. This could have been attenuated in our study using longer sampling intervals; however, this would have been against the purpose of this investigation (Bassett et al. 2001), which was to test the accuracy of the analyzer using short intervals at high intensities, a situation that is common in clinical exercise testing but rare in validation studies (Hodges et al. 2005). Another concern when working with the DBM is that the Bags are made of PVC, and can be marginally permeable to air leakage (Carter and Jeukendrup 2002). In our study, the bags were emptied within ~5 min of test completion so that gas leakage was of little consequence (Hodges et al. 2005). Finally, the turbine flowmeter used by the MOXUS to determine \dot{V}_E is no longer commercially available, and the MOXUS system now uses a pneumotachometer sensor (Rosdahl et al. 2013). These two systems, however, have been recently compared and showed very similar degrees of accuracy and day-to-day reliability (Rosdahl et al. 2013).

While errors in ventilation receive significant attention, errors in the measurement of gas fraction are often not

considered in validation studies (Carter and Jeukendrup 2002; Macfarlane and Wong 2012; Reybrouck et al. 1992), but also will have substantial effects on the calculation of $\dot{V}O_2$, particularly FeO₂, and should be reported (Hodges et al. 2005). When measuring FeO₂ in the present investigation, the Moxus performance was better than some analyzers (Crouter et al. 2006; Gore et al. 2003; Jakovljevic et al. 2008), but poorer than others (Bassett et al. 2001; Foss and Hallen 2004). The Moxus readings for FeO₂ showed the best agreement with the DBM measurements from all variables, with a mean absolute percentage error of only 0.4 % and mean bias of 0.01 percent-points L min⁻¹. Still, since each 0.01 percent-point error in FeO₂ amounts to 0.13 mL of O₂ for each liter of \dot{V}_E , and considering our limits of agreement for FeO₂ (± 0.18 percent-points L min⁻¹), FeO₂ errors alone could explain differences of 350 mL in $\dot{V}O_2$ at a \dot{V}_E of 150 L min⁻¹.

Errors in the measurement of gas exchange parameters can substantially affect the capacity to correctly detect small changes in $\dot{V}O_2$, as required in many criteria for determining the occurrence of a plateau during incremental tests. The original plateau criterion (Taylor et al. 1955) was essentially an application of the limits of agreement, where values up to two standard deviation smaller than the mean stage-to-stage difference in $\dot{V}O_2$ were considered normal, so that only lower values would be considered a plateau (Taylor et al. 1955). Nowadays, however, the search for $\dot{V}O_{2max}$ using automated gas analyzers has distorted the original criterion, and multiple plateau criteria have emerged (Noakes and St Clair 2004). While the use of ramp incremental tests with small load increments over time to measure $\dot{V}O_{2max}$ and detect small changes in $\dot{V}O_2$ —a plateau—is widely promoted (Meyer et al. 2005), our results support the idea that this might be beyond the accuracy of modern analyzers (Howley et al. 1995), or perhaps even the accuracy of the DBM system, given that the possible sources of error with the current “gold-standard” method to evaluate oxygen consumption also increase with decrease in the collection interval.

In the present study, comparisons of $\dot{V}O_2$ between the Moxus and the DBM during high-intensity running resulted in similar limits of agreement as the $\Delta\dot{V}O_2$ over two consecutive 30-s intervals measured by both systems (± 402 mL min⁻¹ O₂ vs. ± 386 mL min⁻¹). Thus, in this case, level of uncertainty around the difference between two consecutive measurements was not greater than the uncertainty around each measurement alone (Howley et al. 1995). Nonetheless, these data indicate that if our exercise protocol was designed to test the incidence of the plateau phenomenon, larger increases in workloads should have been used together with longer collection intervals (Howley et al. 1995). Both proposed changes tend to minimize errors in measurements and allow the expected changes

in $\dot{V}O_2$ to be greater than the limits of agreement for the system accuracy in a particular condition. It has been demonstrated, however, that the timeaveraging of gas exchange parameters does not affect the reproducibility of $\dot{V}O_{2\max}$ itself (Midgley et al. 2007). Another possibility is the use of supramaximal testing to confirm the attainment of $\dot{V}O_{2\max}$ (Midgley and Carroll 2009), which has convincingly demonstrated that the occurrence of a plateau is not a pre-requisite for measuring $\dot{V}O_{2\max}$ (Howley 2007).

In our study, \dot{V}_E , $\dot{V}O_2$ and $\dot{V}CO_2$, were lower for the DBM than for the Moxus system. Howley et al. (1995) have already identified this possibility during high-intensity exercise with short sampling intervals, but to our knowledge it had not been evaluated before. Our Moxus system has already been tested against a different set of DBs and Dry Gas Meter and under more stable conditions (Medbo et al. 2012), with closer agreement between methods than the ones we reported here, in spite of equal systematic bias for $\dot{V}O_2$ ($\sim 80 \text{ mL min}^{-1}$). These two studies were performed by the same researchers, and at the same time of the year, the only difference being the DBM equipment used, the averaging interval and the exercise protocols evaluated. Also, our Moxus system has been shown to be extremely reliable when measuring $\dot{V}O_{2\max}$ over different days (Beltrami et al. 2012; Medbo et al. 2012). Taken together, these suggest that our dry gas meter may have been slightly inaccurate (Carter and Jeukendrup 2002), which is a limitation of the present investigation, and points out to the uncertainty of the true value of measurements even when using “gold standard” techniques, as suggested by Bland and Altman (1995), but criticized by others (Medbo et al. 2002). Unfortunately, the equipment (Carter and Jeukendrup 2002) to test the accuracy of the dry gas meter was not available, and testing it by pumping air from the calibration syringe multiple times into the bags is not ideal. However, it must be noted that even metabolic simulators have been suggested to be problematic when used for validating gas analyzers during breath-by-breath and perhaps even time-averaged measurements (Kim and Robergs 2012). This has led to the suggestion that no “true” gold standard method exist for determining oxygen uptake, even though the DBM is still widely used as a criterion method for validating automated gas analyzers (Hodges et al. 2005). Another limitation of the current investigation is that although the Moxus system has a flow velocity controller for the air coming from the mixing chamber, this does not apply to the gas coming from the calibration bottles. O_2 analyzers are known to be highly dependent on pressure, or flow velocity, (Macfarlane 2001) and to counteract this problem all calibrations were done by the same investigator. Finally, both systems used the same gas analyzers. Although this can affect the level of certainty about the values, the S-3A/I and

CD-3A analyzers have been on the market for many years, and are often part of the DBM setup for validation studies (Bassett et al. 2001; Duffield et al. 2004; Macfarlane and Wong 2012).

Conclusion

In summary, our data show that despite good agreement between the DBM and the Moxus system, within recommended ranges of accuracy, there are large discrepancies in $\Delta\dot{V}O_2$ measured over two consecutive 30-s intervals by the two sets of equipment. Researchers should be aware of the accuracy of their equipment before establishing a protocol for an incremental test, and choose load increments and sampling intervals that are large enough to minimize random errors and so to allow small changes in $\dot{V}O_2$ to be correctly detected. Further testing of gas analyzers during high-intensity exercise is also encouraged to determine the true ability of modern equipments to measure the occurrence or not of the plateau phenomenon during maximal incremental exercise tests.

References

- Astorino TA (2009) Alterations in $VO_{2\max}$ and the VO_2 plateau with manipulation of sampling interval. *Clin Physiol Funct Imaging* 29:60–67
- Atkinson G, Nevill AM (1998) Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Med* 26:217–238
- Atkinson G, Davison RC, Nevill AM (2005) Performance characteristics of gas analysis systems: what we know and what we need to know. *Int J Sports Med* 26(Suppl 1):S2–S10
- Auchincloss JH Jr, Gilbert R, Bowman RP, Baule GH (1971) Determination of maximal oxygen uptake with unsteady-state measurements. *J Appl Physiol* 31:191–197
- Bassett DR Jr, Howley ET, Thompson DL, King GA, Strath SJ, McLaughlin JE, Parr BB (2001) Validity of inspiratory and expiratory methods of measuring gas exchange with a computerized system. *J Appl Physiol* 91:218–224
- Beltrami FG, Froyd C, Mauger AR, Metcalfe AJ, Marino F, Noakes TD (2012) Conventional testing methods produce submaximal values of maximum oxygen consumption. *Br J Sports Med* 46(1):23–29
- Bland JM, Altman DG (1995) Comparing methods of measurement: why plotting difference against standard method is misleading. *Lancet* 346:1085–1087
- Bland JM, Altman DG (1999) Measuring agreement in method comparison studies. *Stat Methods Med Res* 8:135–160
- Bland JM, Altman DG (2007) Agreement between methods of measurement with multiple observations per individual. *J Biopharm Stat* 17:571–582
- Carter J, Jeukendrup AE (2002) Validity and reliability of three commercially available breath-by-breath respiratory systems. *Eur J Appl Physiol* 86:435–441
- Crouter S, Antczak A, Hudak J, DellaValle D, Haas J (2006) Accuracy and reliability of the Parvo Medics TrueOne 2400 and

- MedGraphics VO2000 metabolic systems. *Eur J Appl Physiol* 98:139–151
- Doherty M, Nobbs L, Noakes TD (2003) Low frequency of the “plateau phenomenon” during maximal exercise in elite British athletes. *Eur J Appl Physiol* 89:619–623
- Duffield R, Dawson B, Pinnington HC, Wong P (2004) Accuracy and reliability of a Cosmed K4b2 portable gas analysis system. *J Sci Med Sport* 7:11–22
- Foss A, Hallen J (2004) Validity and stability of a computerized metabolic system with mixing chamber. *Int J Sports Med* 26:569–575
- Gore CJ, Clark RJ, Shipp NJ, Van Der Ploeg GE, Withers RT (2003) CPX/D underestimates VO₂ in athletes compared with an automated Douglas bag system. *Med Sci Sports Exerc* 35:1341
- Hodges LD, Brodie DA, Bromley PD (2005) Validity and reliability of selected commercially available metabolic analyzer systems. *Scand J Med Sci Sports* 15:271–279
- Hopkins WG (2013) A new view of statistics. Internet Society for Sports Science. <http://www.sportsci.org/resource/stats/>. Assessed 20 May 2013
- Howley ET (2007) VO_{2max} and the plateau—needed or not? *Med Sci Sports Exerc* 39:101–102
- Howley ET, Bassett DR, Welch HG (1995) Criteria for maximal oxygen uptake: review and commentary. *Med Sci Sports Exerc* 27:1292
- Jakovljevic DG, Nunan D, Donovan G, Hodges LD, Sandercock GRH, Brodie DA (2008) Lack of agreement between gas exchange variables measured by two metabolic systems. *J Sports Sci Med* 7:15–22
- Kim DY, Robergs RA (2012) Validation of a new mixing chamber system for breath-by-breath indirect calorimetry. *Appl Physiol Nutr Metab* 37:157–166
- La Mere VJ, Brown K, Wigglesworth JK, Edwards JE (1993) Reproducibility between three metabolic systems and validation by Douglas Bag method. *Med Sci Sports Exerc* 225:S9
- Macfarlane DJ (2001) Automated metabolic gas analysis systems: a review. *Sports Med* 31:841–861
- Macfarlane DJ, Wong P (2012) Validity, reliability and stability of the portable Cortex Metamax 3B gas analysis system. *Eur J Appl Physiol* 112:2539–2547
- Macfarlane DJ, Wu HL (2013) Inter-unit variability in two Parvo Medics TrueOne 2400 automated metabolic gas analysis systems. *Eur J Appl Physiol* 113:753–762
- Medbo JI, Mamen A, Welde B, von Heimburg E, Stokke R (2002) Examination of the Metamax I and II oxygen analysers during exercise studies in the laboratory. *Scand J Clin Lab Invest* 62:585–598
- Medbo JI, Mamen A, Beltrami FG (2012) Examination of the Moxus Modular Metabolic System by the Douglas-bag technique. *Appl Physiol Nutr Metab* 37:860–871
- Meyer T, Scharhag J, Kindermann W (2005) Peak oxygen uptake. Myth and truth about an internationally accepted reference value. *Z Kardiol* 94:255–264
- Midgley AW, Carroll S (2009) Emergence of the verification phase procedure for confirming ‘true’ VO_{2max}. *Scand J Med Sci Sports* 19(3):313–322
- Midgley AW, McNaughton LR, Carroll S (2007) Effect of the VO₂ time-averaging interval on the reproducibility of VO_{2max} in healthy athletic subjects. *Clin Physiol Funct Imaging* 27:122–125
- Mitchell JH, Sproule BJ, Chapman CB (1958) The physiological meaning of the maximal oxygen intake test. *J Clin Invest* 37:538–547
- Nevill AM, Atkinson G (1997) Assessing agreement between measurements recorded on a ratio scale in sports medicine and sports science. *Br J Sports Med* 31:314–318
- Noakes TD, St Clair GA (2004) Logical limitations to the “catastrophe” models of fatigue during exercise in humans. *Br J Sports Med* 38:648–649
- Reybrouck T, Deroost F, Van der Hauwaert LG (1992) Evaluation of breath-by-breath measurement of respiratory gas exchange in pediatric exercise testing. *Chest* 102:147–152
- Rietjens GJ, Kuipers H, Kester AD, Keizer HA (2001) Validation of a computerized metabolic measurement system (Oxycon-Pro) during low and high intensity exercise. *Int J Sports Med* 22:291–294
- Rosdahl H, Lindberg T, Edin F, Nilsson J (2013) The Moxus Modular metabolic system evaluated with two sensors for ventilation against the Douglas Bag method. *Eur J Appl Physiol* 113(5):1353–1367
- Salier EJ, Rosdahl H, Schantz P (2012) Validity of the Oxycon Mobile metabolic system under field measuring conditions. *Eur J Appl Physiol* 112:345–355
- Taylor HL, Buskirk E, Henschel A (1955) Maximal oxygen intake as an objective measure of cardio-respiratory performance. *J Appl Physiol* 8:73–80
- Wilmore JH, Davis JA, Norton AC (1976) An automated system for assessing metabolic and respiratory function during exercise. *J Appl Physiol* 40:619–624