Received: 18 May 2023 | Revised: 7 August 2023 | Accepted: 25 August 2023

DOI: 10.1111/sms.14490

ORIGINAL ARTICLE

WILEY

Accuracy of respiratory gas variables, substrate, and energy use from 15 CPET systems during simulated and human exercise

Bas Van Hooren[1] | **Tjeu Souren**[2] | **Bart C. Bongers**[1,3]

[1]Department of Nutrition and Movement Sciences, NUTRIM School of Nutrition and Translational Research in Metabolism, Maastricht University, Maastricht, The Netherlands

[2]Independent Consultant, Utrecht, The Netherlands

[3]Department of Surgery, NUTRIM School of Nutrition and Translational Research in Metabolism, Maastricht University, Maastricht, The Netherlands

Correspondence
Bas Van Hooren, Department of Nutrition and Movement Sciences, NUTRIM School of Nutrition and Translational Research in Metabolism, Maastricht University, Universiteitssingel 50, 6229 ER, Maastricht, The Netherlands.
Email: basvanhooren@hotmail.com

Abstract

Purpose: Various systems are available for cardiopulmonary exercise testing (CPET), but their accuracy remains largely unexplored. We evaluate the accuracy of 15 popular CPET systems to assess respiratory variables, substrate use, and energy expenditure during simulated exercise. Cross-comparisons were also performed during human cycling experiments (i.e., verification of simulation findings), and between-session reliability was assessed for a subset of systems.

Methods: A metabolic simulator was used to simulate breath-by-breath gas exchange, and the values measured by each system (minute ventilation [$\dot{V}E$], breathing frequency [BF], oxygen uptake [$\dot{V}O_2$], carbon dioxide production [$\dot{V}CO_2$], respiratory exchange ratio [RER], energy from carbs and fats, and total energy expenditure) were compared to the simulated values to assess the accuracy. The following manufacturers (system) were assessed: COSMED (Quark CPET, K5), Cortex (MetaLyzer 3B, MetaMax 3B), Vyaire (Vyntus CPX, Oxycon Pro), Maastricht Instruments (Omnical), MGC Diagnostics (Ergocard Clinical, Ergocard Pro, Ultima), Ganshorn/Schiller (PowerCube Ergo), Geratherm (Ergostik), VO2master (VO2masterPro), PNOĒ (PNOĒ), and Calibre Biometrics (Calibre).

Results: Absolute percentage errors during the simulations ranged from 1.15%–44.3% for $\dot{V}E$, 1.05–3.79% for BF, 1.10%–13.3% for $\dot{V}O_2$, 1.07%–18.3% for $\dot{V}CO_2$, 0.62%–14.8% for RER, 5.52%–99.0% for Kcal from carbs, 5.13%–133% for Kcal from fats, and 0.59%–12.1% for total energy expenditure. Between-session variation ranged from 0.86%–21.0% for $\dot{V}O_2$ and 1.14%–20.2% for $\dot{V}CO_2$, respectively.

Conclusion: The error of respiratory gas variables, substrate, and energy use differed substantially between systems, with only a few systems demonstrating a consistent acceptable error. We extensively discuss the implications of our findings for clinicians, researchers and other CPET users.

Section editor: J. Calbet.

KEYWORDS
graded exercise testing, metabolic cart, precision, reliability, simulation, validity

1 | INTRODUCTION

Cardiopulmonary exercise testing (CPET) is commonly used to assess physiological variables and indices, such as the first and second ventilatory thresholds,[1–3] maximal oxygen uptake ($\dot{V}O_{2max}$),[4–6] oxygen uptake kinetics,[7] substrate utilization,[8,9] and total energy expenditure.[10] Accurate determination of these physiological variables is important since CPET outcomes are often used in clinical decision-making, for training prescription, and as gold-standard device for measuring cardiorespiratory fitness and exercise-limiting factors. For example, firemen that do not meet a predefined $\dot{V}O_{2max}$ value may not be allowed to continue their profession[11] and patients that do not meet a predefined $\dot{V}O_{2max}$ value may be advised not to undergo major surgery[12] or to delay treatment.[13] Similarly, accurate measurements are also of critical importance for (professional) athletes as the outcomes are used for decisions to adjust or continue training (e.g., with RED-s syndrome[14]). Furthermore, the outcomes of CPET are often used to determine training zones, which in turn are used to prescribe training intensity.[1] As small errors in the intensity can lead to exacerbated fatigue,[15] accurate training zone determination is important. Finally, CPET is also often used as the gold-standard method, for example, to determine the validity of other methods for estimating physiological thresholds,[16,17] to examine the accuracy of prediction equations,[5] or to assess the accuracy of wearable technology for estimating $\dot{V}O_2$ or energy expenditure.[18,19]

Physiological variables such as the rate of oxygen consumption ($\dot{V}O_2$), carbon dioxide production ($\dot{V}CO_2$), and minute ventilation ($\dot{V}E$) can be measured using different techniques during CPET's. For example, the volume of expired gasses can be measured using volume-sensing or flow-sensing devices, with multiple types available for each device (e.g., hot-wire anemometers [mass-flow controllers] or turbine pitot tubes to measure gas flows). Similarly, the respiratory gas concentrations can be analyzed in different ways (e.g., paramagnetic analyzers or Zirconia fuel cells for O_2 and infrared or thermal conductivity for CO_2). Importantly, the method used to determine flow/volume and gas composition can affect the validity of the measured physiological variables.[20–22] Since commercially available metabolic gas analyses devices employ various methods to measure physiological variables (Table 1), their validity likely also differs.

To validate the physiological variables measured using metabolic gas analyzer systems, some studies have used combustion tests with methanol, ethanol, or propane.[23–25] Since alcohol combustion has a well-defined theoretical value of $\dot{V}O_2$ and $\dot{V}CO_2$, this can be used to determine the accuracy of the CPET system. However, a major limitation of this approach is that it provides only limited information on the accuracy of the CPET system during high intensity exercise, as the combustion flow of gasses is low relative to (progressive) exercise testing. Moreover, the respiratory exchange ratio (RER) and energy expenditure will also be low relative to a human exercise test. Finally, this method allows only the accuracy of $\dot{V}O_2$ and $\dot{V}CO_2$ to be evaluated, but not the accuracy of variables derived from flow and volume measurements such as tidal volume and minute ventilation ($\dot{V}E$). To circumvent these limitations, several studies have compared different CPET systems to each other during exercise,[24,26–29] or to the Douglas bag method.[30–33] However, the true error remains unknown in CPET comparison studies, as even the gold-standard device has some inherent technical measurement error. Additionally, the accuracy of both CPET comparison studies and Douglas bag studies is influenced by biological variability, such that only a small part of the variability between systems reflects measurement error.[34] Finally, the Douglas bag method requires specific skills to ensure valid and reliable results,[35] and this requirement introduces potential for error.

More recently, studies have compared CPET systems to a metabolic simulator, whereby gas flows of known composition and volume mimic the metabolic state during exercise.[34,36–43] Such a setup can provide helpful information on the accuracy of the CPET systems in conditions relevant to high-intensity exercise and may overcome some of the limitations of CPET comparison and Douglas bag studies. However, most simulation studies limited their analysis to one specific CPET system. Yet numerous other systems are routinely used for CPET tests, and their accuracy during (simulated) exercise has yet to be investigated. Therefore, the primary purpose of this study was to investigate and compare the accuracy of 15 popular and commercially available metabolic cart (CPET) systems during simulated exercise. To this purpose, a state-of-the-art metabolic simulator consisting of a breathing simulator combined with a gas-infusion system (Relitech Systems BV; Figure 1) was used to simulate exercise across a range of intensities in continuous breath-by-breath simulation. This system has been shown to be reliable and produces highly accurate breath-by-breath variables.[37] The between-day reliability (i.e., variability in the error) was quantified for a subset of the CPET devices as a secondary aim.

A metabolic simulator does not fully mimic human exercise; for example, it uses dry gasses, while expired human breaths contain ~75% relative humidity during exercise in typical room conditions.[44] Similarly, the temperature of the simulator gasses is lower (typically room temperature of ~21°C vs. ~28–30°C in expired human gas during exercise in typical laboratory room conditions[44,45]), and the simulated breathing pattern is different (stable sinusoidal vs. individual human breathing patterns, with its natural fluctuations in volume, pressure and breathing frequency).[46] A tertiary aim was, therefore, to verify the results obtained during the simulation experiments by comparing all systems against each other during a steady-state cycling test in well-trained individuals.

2 | METHODS

2.1 | General study design

This study comprises of two parts: (1) validation of metabolic analyzers during simulated exercise testing and (2) verification/comparison during steady-state cycling on trained human participants. All measurements were performed over a total of four separate measurement days. This was necessary as not all manufacturers could attend the experiments on the same day.

2.2 | Equipment

CPET data was collected using 15 popular CPET systems (Table 1). To this purpose, all manufacturers were contacted and invited to provide a system for participation in the experiments. We also invited all manufacturers to have their staff present to ensure calibration and handling of the system in line with the manufacturer's guidelines. The following manufacturers were invited but did not participate in the experiments: Dynostics (Dynostics), ParvoMedics (ParvoMedics Inc.), and KORR (KORR Medical Technologies). Reasons for no participation were (a) unwillingness to provide a license to assess the accuracy of the system, despite the availability of the system at the testing facility (Dynostics), (b) cost and time investment (KORR), (c) unclear (ParvoMedics). Finally, PNOĒ did not respond to multiple invitations for participation, but a system was nevertheless acquired from a local athletics coach.

The manufacturers of the CPET systems or the metabolic simulator had no role in the study design, data analysis, interpretation of the data collected, in the report's writing, nor in the decision to submit the paper for publication.

2.3 | Metabolic simulator

The human gas exchange response during exercise was mimicked using a state-of-the-art metabolic simulator consisting of a breathing simulator combined with a gas-infusion system (Relitech Systems BV; Figure 1). This system is reliable and produces highly accurate breath-by-breath variables.[37] The breathing simulator uses a motorized syringe (piston) to simulate breathing variables by adjusting the tidal volume and breath frequency (BF). The tidal volume can range from 1 to 3 L, in steps of 0.5 L, while the BF can be set between 5 and 80 breaths·min^{-1}. This results in a minute ventilation ($\dot{V}E$) range of 10 L·min^{-1} up to 240 L·min^{-1}. The maximum tidal volume is slightly lower than the maximum tidal volume reported in the literature for well-trained athletes (3 vs ~3.8 L·min^{-1}), the BF is higher (80 vs ~65 breaths·min^{-1}), and the resulting $\dot{V}E$ is slightly lower (240 vs. ~250 L·min^{-1}) as reported in literature.[47–50]

The metabolic simulator can also simulate different gas concentrations by using room air pumped back and forth and injecting amounts of pure CO_2 and N_2 (purity ≥99.99%; Linde Gas, Netherlands). The injection of 100% CO_2 creates a gas that simulates a precise amount of $\dot{V}CO_2$ at different breathing frequencies, while 100% N_2 dilutes the ambient air O_2 to a specific O_2 concentration to simulate $\dot{V}O_2$ rates. The simulated $\dot{V}O_2$ and $\dot{V}CO_2$ are automatically calculated using the following equations:

$$\dot{V}CO_2\left(mL \bullet min^{-1}\right) = \dot{V}injCO_2 - FiCO_2 \times \frac{\dot{V}injN_2}{1 - FiO_2 - FiCO_2} \quad (1)$$

$$\dot{V}O_2\left(mL \bullet min^{-1}\right) = FiO_2 \times \frac{\dot{V}injN_2}{1 - FiO_2 - FiCO_2} \quad (2)$$

Where $\dot{V}injCO_2$ and $\dot{V}injN_2$ are the injected amounts of CO_2 and N_2 from the mass-flow controllers in standard temperature pressure dry, respectively, FiO_2 is the fraction of ambient O_2 concentration, and $FiCO_2$ is the ambient CO_2 concentration (0.2093 and 0.0004, respectively).

The ratio between $\dot{V}CO_2$ and $\dot{V}O_2$ (i.e., RER) can also be set to vary between 0.75 and 1.05. The amount of injected CO_2 and N_2 during each breath exhaled by the metabolic simulator is regulated by high-precision mass flow controllers, resulting in a precision of <0.2% for the simulated $\dot{V}O_2$ and $\dot{V}CO_2$. Combined with the simulator's volume stroke accuracy, the metabolic simulator creates $\dot{V}O_2$ and $\dot{V}CO_2$ with an accuracy of <0.5%, even at the high VE ranges. The simulator was certified 1.5 years prior to the first test day and certified again 2 weeks before the last testing day. The system is routinely used at Maastricht

TABLE 1 Software and hardware specifications for CPET system.

	Vyntus CPX	Oxycon Pro	Omnical V6	Ergostik	Metalyzer 3B	MetaMax 3B	VO2 masterPro
Manufacturer	Vyaire Medical, Mettawa, IL, USA	Vyaire Medical, Mettawa, IL, USA	Maastricht Instruments, Maastricht, The Netherlands	Geratherm Respiratory GmbH, Bad Kissingen, Germany	Cortex Biophysik, Leipzig, Germany	Cortex Biophysik, Leipzig, Germany	VO2master Health Sensors Inc., Vernon, BC, Canada
Type	Breath-by-breath	Mixing-chamber & breath-by-breath	Mixing-chamber/ diluted flow	Breath-by-breath	Breath-by-breath	Breath-by-breath	Breath-by-breath
Volume measurement	Turbine (Vyaire, Mettawa, IL, USA)	Turbine (Vyaire, Mettawa, IL, USA)	Balgengasmeter (Itron G16, Liberty Lake, WA, USA)	Differential pressure (Geratherm)	Turbine (Cortex)	Turbine (Cortex)	Differential pressure (VO2master)
O_2 measurement	Chemical fuel cell (Teledyne, CA, USA)	Chemical fuel cell (Teledyne, CA, USA)	Paramagnetic (ABB Magnos206, Frankfurt, Germany)	Chemical fuel cell (Envitec NJ, USA)	Chemical fuel cell (Teledyne, CA, USA)	Chemical fuel cell (Teledyne, CA, USA)	Chemical fuel cell (Envitec NJ, USA)
CO_2 measurement	Non-Dispersive Infrared (Vyaire, Mettawa, IL, USA)	Non-Dispersive Infrared (Vyaire, Mettawa, IL, USA)	Infrared Photometer analyzer (ABB Uras26, Frankfurt, Germany)	Non-Dispersive Infrared (Treymed, NJ, USA)	Non-Dispersive Infrared (Treymed, NJ, USA)	Non-Dispersive Infrared (Treymed, NJ, USA)	N/A
Accuracy for volume, $\dot{V}O_2$, $\dot{V}CO_2$	±3% (50 mL) for all outcomes	±3% (50 mL) for all outcomes	Not stated	±3% (50 mL) for all outcomes	±3% for all outcomes	±3% for all outcomes	Not stated
Approximate system cost[a]	€ 30.700[b]	Not applicable	€ 70.000*	€ 16.000[b]	€ 15.000[b]	€ 24.000[b]	€ 6.100

[a]Cost for a system in The Netherlands in 2022–2023, exclusive of shipping costs. Note that the cost for most systems is dependent on the configurations (e.g., with or without ECG add-on). We assumed one dollar corresponded to one euro.

[b]Exclusive of local taxes.

[c]Cost of base system without calibration gasses and regulators, facemasks, et cetera.

*The cost for a newer version of the system will be substantially reduced.

University Medical Center+ for the quality control program of clinically used metabolic carts.

2.4 | Simulation protocol

The CPET systems were connected directly to the outlet of the metabolic stimulator, as shown in Figure 1. Custom-made adaptors were used to connect the systems when required (see supplementary Figure SI for an example). We attempted to use the same dead space for all systems, and to minimize turbulations introduced by the custom-made adaptors. Each CPET system underwent a standardized protocol to assess $\dot{V}E$, BF, $\dot{V}O_2$, $\dot{V}CO_2$, and RER as primary outcomes. Additional data assessed included FiO_2, $FiCO_2$ (the percentage of oxygen and carbon dioxide in inspired air, respectively), and FeO_2, $FeCO_2$ (percentage of oxygen and carbon dioxide in expired air, respectively). Note that not all systems measured or provided all this additional data. The mixing chamber methodology applied in Omnical V6 and Oxycon Pro does not measure continuously FiO_2 and $FiCO_2$ (but rather at the start of a measurement), Calibre does not provide these parameters in the time and breath table output, and VO2masterPro determines only mixed FeO_2 values.

The "Std" mode on the simulator was used first, with the tidal volume set at 2 L, and RER at 1.00 ($\dot{V}O_2$, $\dot{V}CO_2$ equal). During the experiments, BF changed from $20 \cdot min^{-1}$, to $40 \cdot min^{-1}$, $60 \cdot min^{-1}$, and $80 \cdot min^{-1}$. $\dot{V}O_2$ and $\dot{V}CO_2$ at each BF were 1, 2, 3, and $4\,L \cdot min^{-1}$. The BF's and tidal volume used mimic physiological values reported during human physical activity and exercise testing.[31,47–50] A second protocol was performed in "CPX" mode to simulate different combinations of RERs with increasing BFs and $\dot{V}E$. The RER variations were performed to mimic the increased oxidation of carbohydrates with increasing exercise intensity and to mimic buffering of ion concentrations $[H^+]$ by bicarbonate $[HCO_3^-]$ at very high exercise intensities.[51] The simulated RER values were 0.75, 0.85,

PowerCube Ergo	Quark CPET	K5	Ultima CPX	Ergocard CPX clinical	Ergocard CPX Pro	PNOĒ	Calibre
Ganshorn, Medizin Electronic GmbH, Niederlauer, Germany	COSMED, Rome, Italy	COSMED, Rome, Italy	MGC Diagnostics, Dinant, Belgium	MGC Diagnostics, Dinant, Belgium	MGC Diagnostics, Dinant, Belgium	ENDO Medical, Palo Alto, CA, USA	Calibre Biometrics, Wellesley, MA, USA
Breath-by-breath	Mixing-chamber & breath-by-breath	Mixing-chamber & breath-by-breath	Breath-by-breath	Breath-by-breath	Breath-by-breath	Breath-by-breath	Breath-by-breath
Differential pressure (Ganshorn)	Turbine (COSMED)	Turbine (COSMED)	Pitot tube (MGC Diagnostics)	Pitot tube (MGC Diagnostics)	Pitot tube (MGC Diagnostics)	Thermal sensor, (Sensirion, Stäfa, Switzerland)	Thermal sensor, (Sensirion, Stäfa, Switzerland)
Chemical fuel cell (Envitec NJ, USA)	Paramagnetic (Servomex, Ltd., Sussex, UK)	Galvanic fuel cell (City Technology, NC, USA)	Galvanic fuel cell (Teledyne, CA, USA)	Galvanic fuel cell (Teledyne Ls-10, CA, USA)	Laser Spectrometer (Oxigraf CA, USA)	Chemical fuel cell (Teledyne, CA, USA)	Electro chemical (Angst Pfister, Switzerland)
Ultrasound (Ganshorn, Germany)	Non-Dispersive Infrared (COSMED, Italy)	Non-Dispersive Infrared (COSMED, Italy)	Non-Dispersive Infrared (MGC Diagnostics, Belgium)	Non-Dispersive Infrared (Treymed Comet II, NJ, USA)	Non-Dispersive Infrared (Treymed Comet II, NJ, USA)	Thermal conductivity (Sensirion, (Switzerland)	Thermal conductivity (Sensirion, (Switzerland)
±3% for all outcomes	±3% (50 mL) for all outcomes	±3% (50 mL) for all outcomes	<4% for all outcomes	<4% for all outcomes	<4% for all outcomes	Not stated	Not stated
€ 24.000	€ 20.500	€ 30.000[b]	€ 29.000[b, c]	€ 13.000[c]	€ 24.900[c]	€ 15.600	€ 399

0.95, and 1.05, with $\dot{V}O_2$ being 1, 2, 3, and 4 L·min^{-1} at each RER, corresponding to a $\dot{V}CO_2$ of 0.75, 1.7, 2.85, and 4.2 L·min^{-1}. Note that the lowest step of the CPX procedure (i.e., with BF of 10·min^{-1}, RER 0.75, and $\dot{V}O_2$ of 1 L·min^{-1}) required a separate setting for Omnical V6 (that used a lower active flow), and a separate mask set-on for VO2masterPro. These stages were therefore simulated separately while using these different configurations.

Each stage lasted at least 2 min for breath-by-breath systems to ensure sufficient time for a stable breath collection, and the graphical user interface for each system was checked to ensure a steady state (Figure 1). Each stage lasted ~5 min for mixing chamber systems to ensure sufficient time to flush the mixing chamber, which was again confirmed by visual inspection of the graphical user interface. For mixing chamber systems, we also quantified the time required for each system to reach a steady state in gas exchange variables. To this purpose, the simulation and data collection were started simultaneously and the delay was quantified as the time difference between the first sample at which the steady state was reached (determined using visual inspection) and the start of the simulation (see supplementary file I, Figure S3 for more details).

Finally, we quantified the between-day reliability for all systems that were available at the lab for at least two experimental sessions, by repeating the same simulation experiments (see section 2.8). Between-day reliability was not assessed for all systems because most manufacturers were only present for one day at the testing facility with their system.

2.5 | Human validation protocol

Human exercise was used to verify the results obtained during the simulation tests and are further detailed in supplementary file I, section 2. Briefly, a total of three well-trained healthy individuals cycled at the highest intensity at which physiological variables remained stable (i.e., ~25 Watts below their gas exchange/first ventilatory

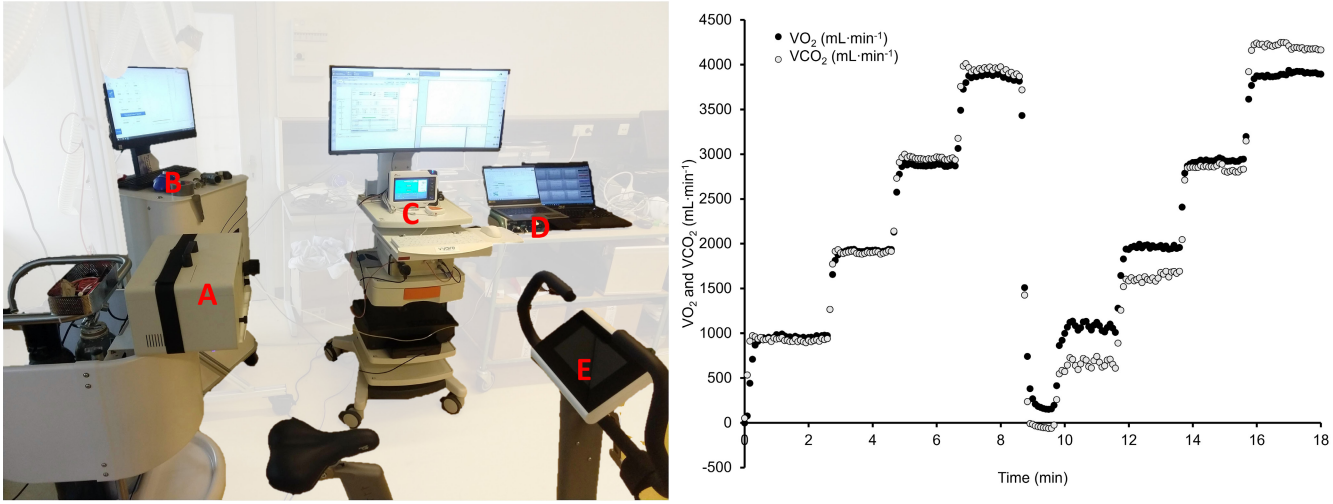

FIGURE 1 Left: Experimental set-up with the metabolic simulator (A), three of the CPET systems (B = Omnical V6; C = Vyntus CPX; D = MetaLyzer 3B), and the bike used for the human tests (E = Lode Corival CPET). The CPET systems were connected to the outlet of the metabolic simulator as shown in the image (in this case for the Vyntus CPX). Right: example recording of the simulation protocol by one of the CPET systems (Omnical v6). The first stepwise increase represents the "Std" mode with a constant RER of 1.00, and the second stepwise increase the "CPX" mode with an increase in RER for each stage.

threshold) while gas exchange data were collected two times per system for three (breath-by-breath) or five (mixing chamber) minutes in a randomized and counterbalanced order.

2.6 | Data collection settings for each CPET system

The metabolic simulator mimics human breathing and creates artificial, highly accurate known breaths. From their design, the mass-flow controllers used in the metabolic simulator for CO_2 and N_2 have a temperature-controlled output normalized to absolute volume output in standard temperature and pressure dry (STPD) (SLN, normalized standard liters), as detailed in equations 1 and 2. $\dot{V}E$, the volume strokes from the piston pump of the metabolic simulator, uses room air, and is thus at ambient conditions (ambient temperature and pressure; ATP).

CPET systems are typically used for human testing and because human expired volumes have a higher temperature and humidity than ambient air, the expired volumes are expressed in saturated body temperature, and pressure conditions (BTPS). By measuring or assuming a specific humidity, temperature, and pressure of the expired air, the CPET systems convert the values measured in BTPS to STPD to allow comparison between different measurement conditions. For example, CPET systems typically assume the expired gas is 100% humid and has a temperature of 31.5°C. Since this assumption is incorrect during the metabolic simulation experiments, the gas volumes in STPD require correction to allow comparison with the simulated values. The manufacturers were therefore asked to turn off the BTPS correction within the software application when possible. Specifically, the Quark CPET, K5, MetaLyzer, MetaMax, Vyntus CPX, Oxycon Pro, Ergocard Clinical and Pro, Ultima, PowerCube, Ergostik and Calibre applications used a setting that stopped the conversion from ATP to BTPS for $\dot{V}E$, to allow direct comparison with the simulated values. Omnical already expressed $\dot{V}O_2$ and $\dot{V}CO_2$ in STPD by measuring the humidity and temperature of the gas, and no correction was therefore required for the simulation tests. VO2masterPro and PNOĒ expressed $\dot{V}O_2$ and $\dot{V}CO_2$ in STPD, assuming the measured exhaled air is 100% humid at ambient pressure and with an exhaled air temperature of 34, and 31.5°C, respectively. Using these values, the $\dot{V}O_2$ and $\dot{V}CO_2$ were corrected from ATP to STPD, and $\dot{V}E$ was corrected from ATP to BTPS.

Room temperature and relative humidity ranged between 19 and 21°Celsius, and 45%–57%, respectively during all simulation and cycling measurements. During all experiments, the lab was ventilated by opening windows and doors, and all individuals present during testing were asked to maintain >5 m distance from the measurement area.

2.7 | CPET calibration

Each CPET system was calibrated according to the manufacturer guidelines prior to the "Std" simulation, before the "CPX" simulation, and again prior to the human experiments. All manufacturers used their own gas for

calibration to best reflect typical system calibration. The only exception was PNOĒ, which states that only room air calibration is required for routine purposes. The use of certified calibration gas is optional and not standard. We, therefore, used PNOĒ after room air calibration and in a second measurement after certified gas calibration mode for the simulation experiments, whereby CO_2/O_2 mix (5% CO_2/16%O_2) calibration gas was used to calibrate the CO_2/O_2 sensor. The volume for all systems was calibrated using a certified 3 L syringe from each respective manufacturer, except for ErgoCard CPX Clinical, Ergocard CPX Pro, and COSMED Quark where the manufacturer preferred to calibrate their system using the motorized 3 L piston syringe pump of the metabolic simulator. The potential impact of this is discussed later.

2.8 | Data processing

During the simulation tests, the mean value of the last minute of each stage was used for analyses to ensure adequate flushing of the gas-filled dead space of the simulator. The period selected for analyses was also confirmed by visual inspection of a steady state.

Data processing for the human cycling experiments is detailed in supplementary file I, section 3. Briefly, data were analyzed over the final minute of each period and subsequently averaged over the two counterbalanced 1-min periods to make comparisons between systems. Reference values for session, two, three and four were calculated based on the average $\dot{V}O_2$ and $\dot{V}CO_2$ values recorded by Vyntus CPX and Oxycon Pro (B×B) while correcting their measured values for the respective errors in $\dot{V}O_2$ and $\dot{V}CO_2$ from the simulation experiments. Vyntus CPX and Oxycon Pro were used to calculate the reference value because these systems (a) were present at the research facility during all human experiments, (b) showed generally high accuracy during the simulation experiments, and (c) showed good to acceptable between-day reliability. For the first test, the average $\dot{V}O_2$ and $\dot{V}CO_2$ values for Vyntus CPX, Omnical V6, and Ergostik were used as reference (with correction) as Oxycon Pro was not available during these experiments.

2.9 | Statistical analysis

The accuracy of the CPET systems were assessed for the main ventilatory and gas exchange variables: $\dot{V}E$ ($L \cdot min^{-1}$), BF ($breaths \cdot min^{-1}$), $\dot{V}O_2$ ($mL \cdot min^{-1}$), $\dot{V}CO_2$ ($mL \cdot min^{-1}$), and RER. For the trials with RER <1.00 (metabolic simulator in "CPX" mode), we also computed the energy expenditure derived from fats and carbohydrates and total energy expenditure from the simulated and measured $\dot{V}O_2$ and $\dot{V}CO_2$ using Jeukendrup's equation for moderate- to high-intensity exercise.[51] This was done to determine the impact of errors in the measured $\dot{V}O_2$ and $\dot{V}CO_2$ values on substrate and energy expenditure estimation.

Agreement between the CPET systems and metabolic simulator was assessed in several ways. First, the measurement error was calculated for the simulation test by subtracting the expected value (i.e., simulated) from the measured value (i.e., converted CPET readouts). We expressed this error as a percentage of the expected value (i.e., [(measured − expected)/expected] × 100) and computed the average relative percentage error and average absolute percentage error (AAPE) for all simulation steps for each system to indicate the overall measurement error.

To objectively assess the agreement between the simulator and CPET systems, we used a statistical approach proposed by Shieh[52] with the percentage difference as the unit for comparison. In this method the mean difference and variability of the difference between the simulator and CPET system is assessed in relation to an a priori determined threshold, whereby a specified proportion of the data should fall within the threshold to declare agreement. Errors for the main ventilatory and gas exchange variables were considered: good, when the errors were <3%, acceptable, <5%, and poor ≥5%. This classification is in line with the error of 3% specified by most manufacturers for these outcomes (Table 1), and approximately in line with an error of <3% being acceptable for volume measurements according to the 2019 American Thoracic and European respiratory societies.[53] We used slightly higher ranges for substrate use and rated errors of <5% as good, <10% as acceptable, and ≥10% as poor. For energy expenditure, previous studies defined a 2% error as acceptable in resting metabolic rate measurements,[23,25] and we considered a slightly higher error acceptable during exercise testing. The error for energy expenditure was therefore interpreted similarly to the main ventilatory and gas exchange variables. The central null-proportion (reflecting the fraction of datapoints that should fall within this threshold) was set to 0.95 in line with the widely used 95% limits of agreement, and the alpha level to 0.05. Therefore, if the 95% confidence intervals of the limits of agreement between the simulator and CPET system for the assessed outcome, fell within the specified threshold, the null-hypothesis that there is no agreement between the systems was rejected.

To assess if the relative (i.e., non-absolute) error changed with higher simulated values, we assessed if the slope of the regression line fitted on the error differed significantly from zero.

Between-day reliability was quantified by calculating the standard deviation over all repeated measurements per system. This reliability measure represents the typical variation in the measured value from day to day. The reliability was also expressed as a percentage by dividing the standard deviation by the mean of the measurements multiplied by 100 (i.e., coefficient of variation). This approach was used as we typically only had two repeated measures on each system, thus not allowing us to calculate a standard error of measurement or intraclass correlation coefficient.

3 | RESULTS

3.1 | Metabolic simulation

All data, and errors in original units and both relative and absolute percentage errors for all individual simulation steps are available from online supplementary file II.

Relative percentage errors averaged over all simulated volumes are reported in Tables 2 and 3, as well as depicted in Figure 2. The absolute percentage error for $\dot{V}E$, BF, $\dot{V}O_2$, $\dot{V}CO_2$, RER, and the overall error for each device averaged over all simulated volumes are reported in Table S1 and illustrated in Figure 3. Table S2 reports the absolute percentage errors for energy derived from fats, carbohydrates, and total energy expenditure averaged over all simulated steps, while Figure S2 visualizes these errors.

The relative percentage error significantly increased with higher simulated volumes for some devices, while it remained constant or decreased for others (Figure 4 and Table S3).

Between-day reliability for a subset of the tested devices is reported in Table S4 (in original units) and S5 (in percentage units/coefficient of variation), while Table S6 and Figure S3 depict the time to reach a steady-state gas concentration in the three mixing-chamber devices assessed. Table S8 shows the overall mean absolute percentage error (combined over gas exchange and substrate/energy use) for each system.

3.2 | Human validation

The measured gas exchange variables, substrate use, and energy expenditure measured during the cycling experiments is reported in Table S7. Figure 5 also shows the $\dot{V}O_2$ and $\dot{V}CO_2$ measured by each system during the cycling experiments in the four sessions.

TABLE 2 Mean ± SD relative percentage errors (%e) for respiratory parameters, averaged over all simulated steps.

System	%e $\dot{V}E$	%e BF	%e $\dot{V}O_2$	%e $\dot{V}CO_2$	%e RER	Overall %e
Vyntus CPX	−4.15 ± 1.92	−2.90 ± 1.30	1.15 ± 1.07 *	2.14 ± 0.89 *	0.97 ± 1.45 *	−0.56
Oxycon Pro B × B	−1.84 ± 1.17 *	−3.79 ± 2.67	0.24 ± 1.33 *	−0.31 ± 1.14 *	−0.25 ± 2.11	−1.19
Oxycon Pro MC	−7.86 ± 1.59	−3.75 ± 2.69	−2.28 ± 1.23	−2.12 ± 1.39	0.15 ± 0.69 **	−3.13
Omnical V6	−6.52 ± 4.57	−2.78 ± 1.58	−2.24 ± 1.77	−2.06 ± 2.46	0.59 ± 2.69	−2.57
Ergostik	−3.93 ± 1.16	−3.75 ± 2.89	−0.91 ± 3.14	0.72 ± 2.13	1.72 ± 2.08	−1.23
MetaLyzer 3B	1.24 ± 1.81	−2.50 ± 1.34	2.85 ± 2.22	5.40 ± 1.86	2.23 ± 2.77	1.84
MetaMax 3B	0.89 ± 1.35	−2.79 ± 1.28	1.64 ± 1.87	1.67 ± 2.73	0.04 ± 2.65	0.29
VO2masterPro[a]	−3.84	−2.17	−11.68	–	–	−5.82[a]
PowerCube Ergo	−3.34 ± 3.81	−3.70 ± 2.93	2.90 ± 7.80	18.3 ± 10.2	14.8 ± 2.75	5.78
Quark CPET	0.24 ± 2.00	−2.77 ± 1.25	0.60 ± 1.18 *	−4.15 ± 2.13	−4.69 ± 2.53	−2.15
Ultima CPX	−8.91 ± 1.29	−2.84 ± 1.28	−8.97 ± 1.17	−5.41 ± 1.95	3.89 ± 1.07	−4.45
Ergocard CPX Clinical	5.80 ± 2.24	−2.78 ± 1.24	−3.10 ± 1.54	0.22 ± 4.03	3.47 ± 4.53	0.72
Ergocard CPX Pro	6.47 ± 2.03	−2.56 ± 1.45	−2.52 ± 2.14	−2.82 ± 4.12	−0.30 ± 2.88	−0.34
K5	−0.80 ± 1.03 *	−2.85 ± 1.31	−7.80 ± 2.93	−5.95 ± 0.88	2.12 ± 2.88	−3.05
PNOĒ	44.3 ± 8.22	−1.22 ± 1.30	8.25 ± 5.72	3.39 ± 4.33	−4.36 ± 1.92	10.1
Calibre	−2.33 ± 2.84	−1.69 ± 1.63	0.23 ± 1.41 *	0.68 ± 1.38 *	−0.02 ± 1.85 *	−0.63

Note: ** good agreement (<3% error); * acceptable agreement (<5% error); No star indicates that we were unable to establish good or acceptable agreement (≥5% error).

Abbreviations: BF, breathing frequency; MC, mixing chamber; RER, respiratory exchange ratio; $\dot{V}E$, minute ventilation; $\dot{V}CO_2$, carbon dioxide production; $\dot{V}O_2$, oxygen consumption.

[a]Note that the overall error for this device does not include $\dot{V}CO_2$ or RER.

TABLE 3 Mean ± standard deviation relative percentage errors (%e) for substrate use and total energy expenditure, averaged over all simulated steps.

System	%e Energy from carbs	%e Energy from fats	%e Total energy expenditure
Vyntus CPX	−1.90 ± 6.65	8.53 ± 7.76	1.96 ± 0.78 *
Oxycon Pro B×B	5.38 ± 8.10	0.48 ± 7.09	0.04 ± 0.79 **
Oxycon Pro MC	4.66 ± 10.6	−2.31 ± 7.76	−2.52 ± 0.56 *
Omnical V6	−8.04 ± 13.1	0.06 ± 17.9	−2.53 ± 1.58
Ergostik	32.3 ± 38.4	−28.1 ± 10.5	−3.24 ± 2.47
MetaLyzer 3B	52.9 ± 60.0	−24.8 ± 9.83	3.75 ± 1.09
MetaMax 3B	36.3 ± 57.3	2.87 ± 26.1	2.29 ± 0.19
VO2masterPro	–	–	–
PowerCube Ergo	99.0 ± 72.3	−133 ± 134	−2.16 ± 6.16
Quark CPET	−16.2 ± 4.31	43.2 ± 61.4	−0.68 ± 0.99 *
Ultima CPX	26.3 ± 31.5	−39.5 ± 20.8	−8.35 ± 0.26
Ergocard CPX Clinical	50.2 ± 48.3	−42.7 ± 22.4	−1.21 ± 2.03
Ergocard CPX Pro	8.90 ± 20.0	−3.12 ± 5.66	−0.44 ± 2.28
K5	1.79 ± 11.3	−8.01 ± 3.94	−6.27 ± 0.19
PNOĒ	−39.5 ± 29.2	76.1 ± 74.8	8.16 ± 5.13
Calibre	0.23 ± 13.4	0.68 ± 13.8	−0.02 ± 1.03 **

Note: ** good agreement (<3% error); * acceptable agreement (<5% error); No star indicates that we were unable to establish good or acceptable agreement (≥5% error).

Abbreviation: MC, mixing chamber.

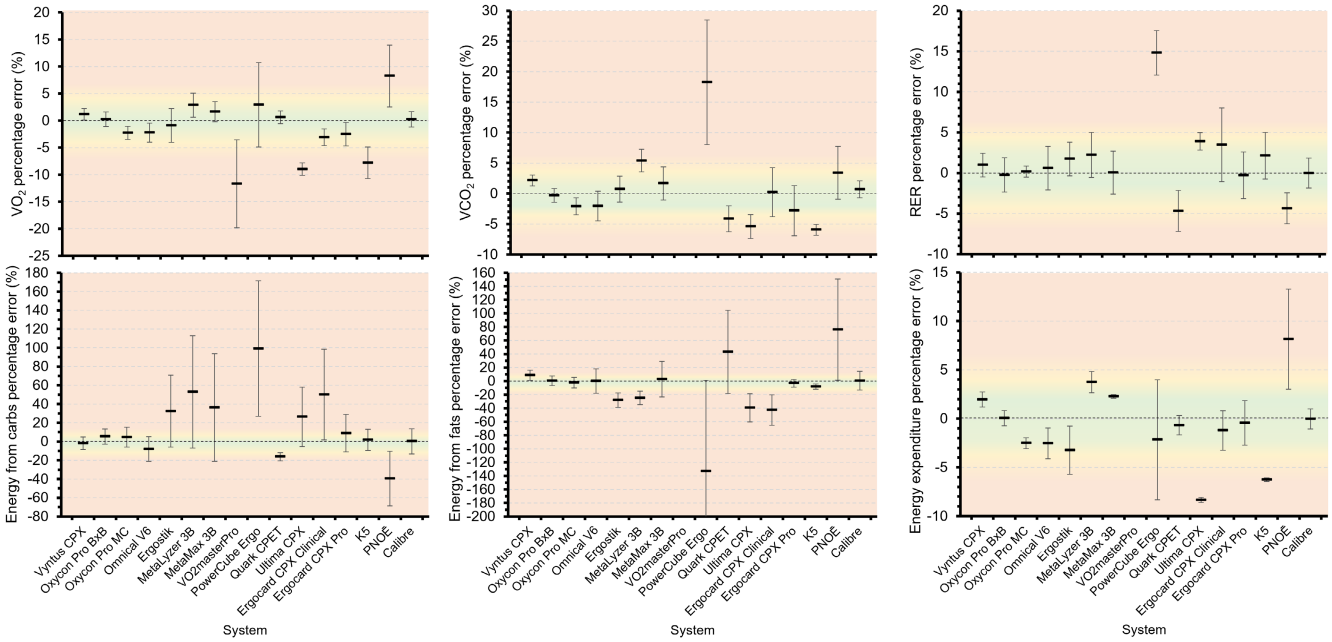

FIGURE 2 Mean relative percentage errors for each device for $\dot{V}O_2$, $\dot{V}CO_2$, RER, energy derived from fats, energy derived from carbohydrates, and total energy expenditure. Dashed lines represent the average error over all simulated steps, while error bars represent the standard deviation of the error over all simulated steps. Wider error bars indicate a lower precision of the measured variable. Note that in the middle bottom figure, the relative percentage error ranges from 1% to −267% for PowerCube Ergo, but only part of the error bar is shown to maintain readable scaling. No error for substrate usage or total energy expenditure is available for VO2masterPro as this device measures only $\dot{V}O_2$.

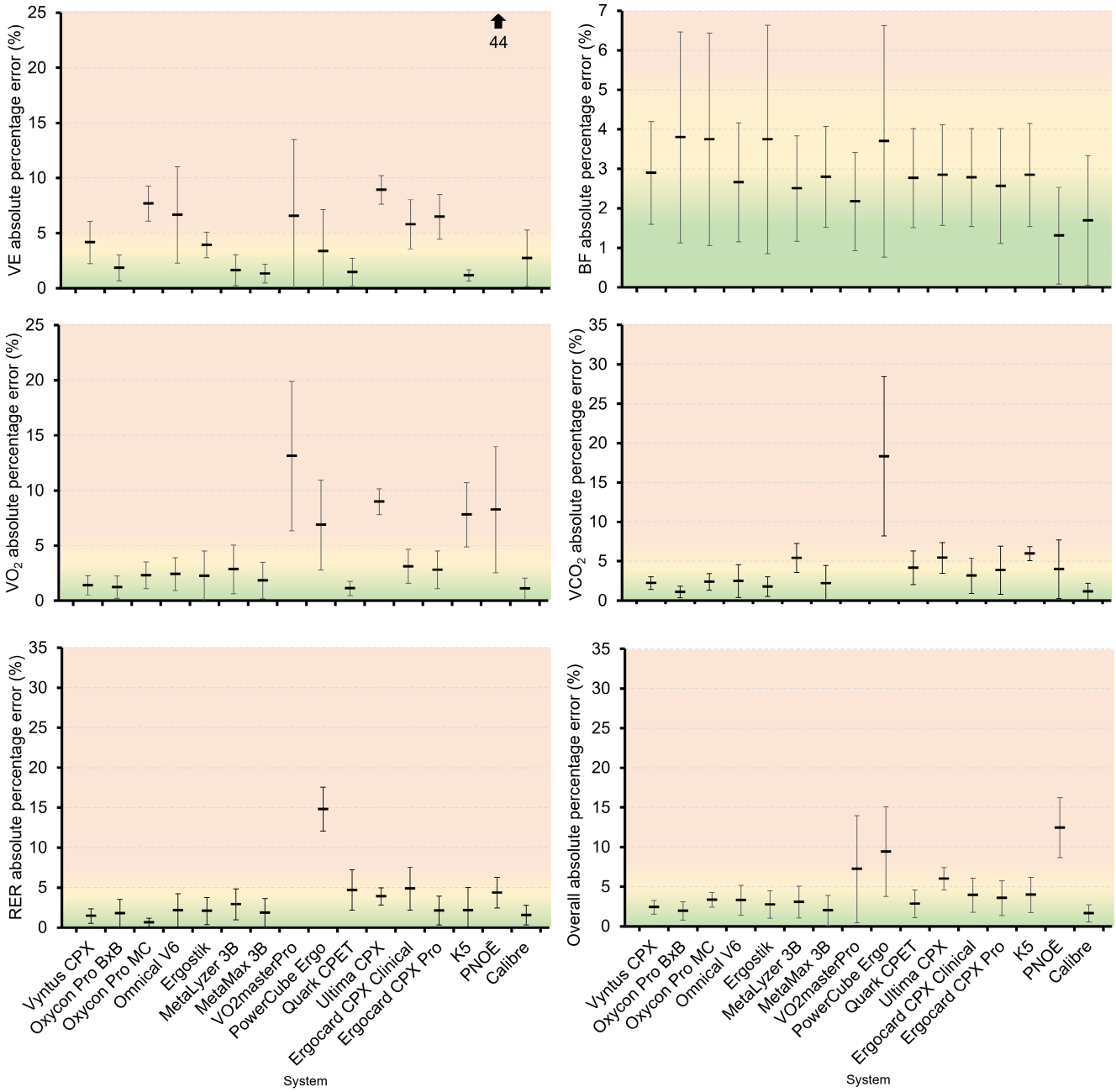

FIGURE 3 Mean ± standard deviation of absolute percentage errors for gas exchange variables per device. Dashed lines depict the mean error over all simulated steps while error bars represent the standard deviation of the error. Note that in the left top figure the mean error for $\dot{V}E$ for PNOÊ was 44%. No error for $\dot{V}CO_2$ or RER is available for VO2masterPro as this device measures only $\dot{V}O_2$. The overall percentage error is computed over all gas exchange variables in the figure.

4 | DISCUSSION

The primary aim of this study was to assess the accuracy by which commonly used CPET systems can assess respiratory gas exchange variables and substrate and energy use during simulated exercise. The following sections discuss the observed relative and absolute errors, prior to explaining potential causes for the observed errors. Finally, we comment briefly on the verification of these errors during the human tests and end with practical implications for CPET users.

4.1 | Summary of the relative and absolute errors

When averaged over all simulated volumes and over all systems, $\dot{V}O_2$ was underestimated by an average of −1.35% (median 0.34%; Figure 2). However, there were substantial differences in the accuracy between systems. Eleven out of the 16 systems assessed, under- or overestimated $\dot{V}O_2$ by less than 3% (Figure 2, Table 2), but the within-device variability in this accuracy resulted in none of the systems

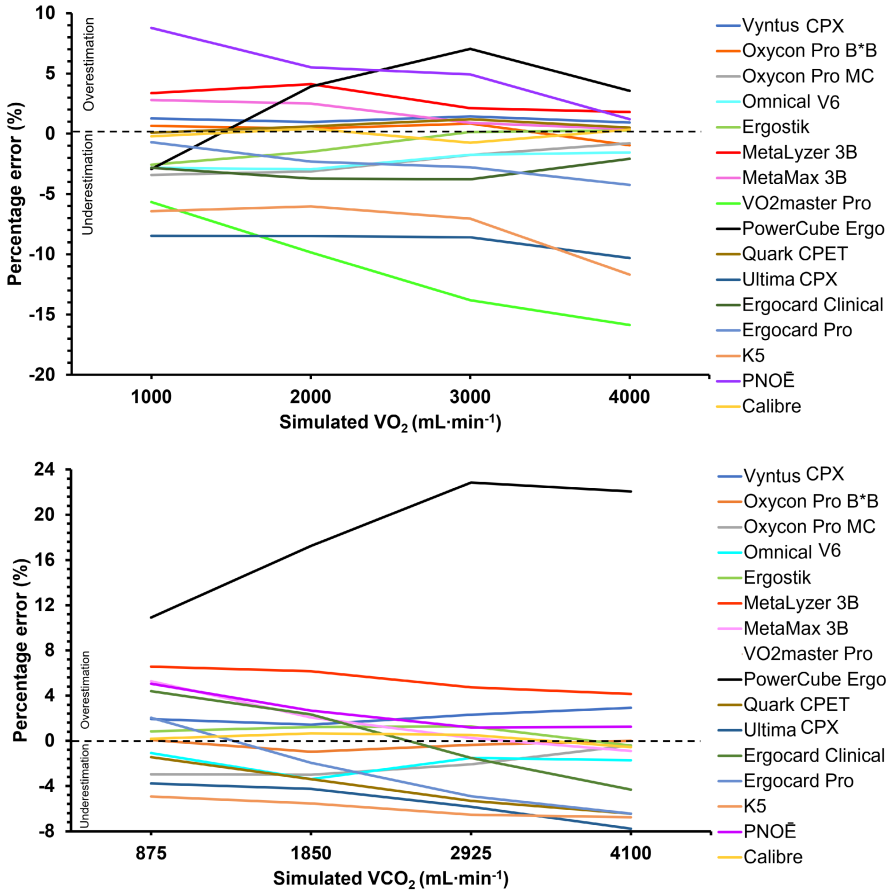

FIGURE 4 Relative percentage error for $\dot{V}O_2$ (top) and $\dot{V}CO_2$ (bottom), as a function of the simulated $\dot{V}O_2$ and $\dot{V}CO_2$ for each device. Errors are averaged over each step of the "Std" (i.e., RER = 1.00) and "CPX" (i.e., RER increases with increased $\dot{V}O_2$) protocols. Because the simulated $\dot{V}CO_2$ differed between the "Std" and "CPX" protocols, the average simulated value is depicted on the x-axis in the figure. MC, mixing chamber; RER, respiratory exchange ratio; $\dot{V}CO_2$, rate of carbon dioxide production; $\dot{V}O_2$, rate of oxygen uptake.

achieving statistical agreement at a 3% error level (Table 2). Nevertheless, four systems had sufficiently low variability in this accuracy to achieve acceptable statistical agreement. One system showed a mean relative error within 5%, while the remaining four systems all had mean errors >5% and thus showed poor accuracy. The relative error in $\dot{V}O_2$ remained constant for the majority (10/16) of systems with higher simulated $\dot{V}E$, thus demonstrating no proportional bias (Table S3; Figure 4), although further research is required on their accuracy at higher volumes seen in elite athletes. Conversely, some systems overestimated $\dot{V}O_2$ at low simulated $\dot{V}E$, but the error in measured $\dot{V}O_2$ decreased with higher simulated $\dot{V}E$. While this demonstrates better accuracy in the range investigated, it could lead to underestimation of $\dot{V}O_{2max}$ at higher volumes seen in elite athletes. One system (VO2masterPro) consistently underestimated $\dot{V}O_2$ and this underestimation increased with higher $\dot{V}E$. Similarly, two other systems (K5, Ultima) also consistently underestimated $\dot{V}O_2$ and although the underestimation increased with higher volumes, the slope did not reach statistical significance. Nevertheless, care should therefore be taken when these systems are used, in particular in $\dot{V}O_{2max}$ testing as it will lead to increasingly larger underestimations with increasing absolute $\dot{V}O_2$ levels.

The average relative error for $\dot{V}CO_2$ was 0.64% (median 0.22%), although there were again notable differences in accuracy between systems, with nine systems demonstrating <3% error, two systems showing 3%–5% error, and four systems showing a mean error >5% (Figure 2). Only three systems exhibited sufficiently low variability in the error to achieve statistical agreement at the 5% level. Although the relative error also remained constant for most (10/16) systems with higher simulated $\dot{V}E$, all other systems showed a negative slope (Table S3, Figure 4). Similar to $\dot{V}O_2$, some systems therefore underestimated $\dot{V}CO_2$ by an increasingly larger magnitude with higher simulated $\dot{V}CO_2$. The over- or underestimation for $\dot{V}O_2$ and $\dot{V}CO_2$ can lead to significant errors in RER when the direction of over- or underestimation differs between the two variables. However, most systems consistently under- or overestimated both $\dot{V}O_2$ and $\dot{V}CO_2$ such that 10 systems had an RER error <3%, four 3%–5%, and only one system >5% (Figure 2, Table S3).

Estimation of the energy derived from different substrates, as well as total energy expenditure, requires accurate measurement of $\dot{V}O_2$, $\dot{V}CO_2$, and RER. For example, while an equivalent underestimation of $\dot{V}O_2$ and $\dot{V}CO_2$ may yield a highly accurate RER, it will lead to an underestimation in the energy derived from fats and carbohydrates, and thus total energy expenditure (e.g., Oxycon Pro mixing chamber in Figure 2). Due to the sensitivity of substrate use for accurate $\dot{V}O_2$, $\dot{V}CO_2$, and RER measures,

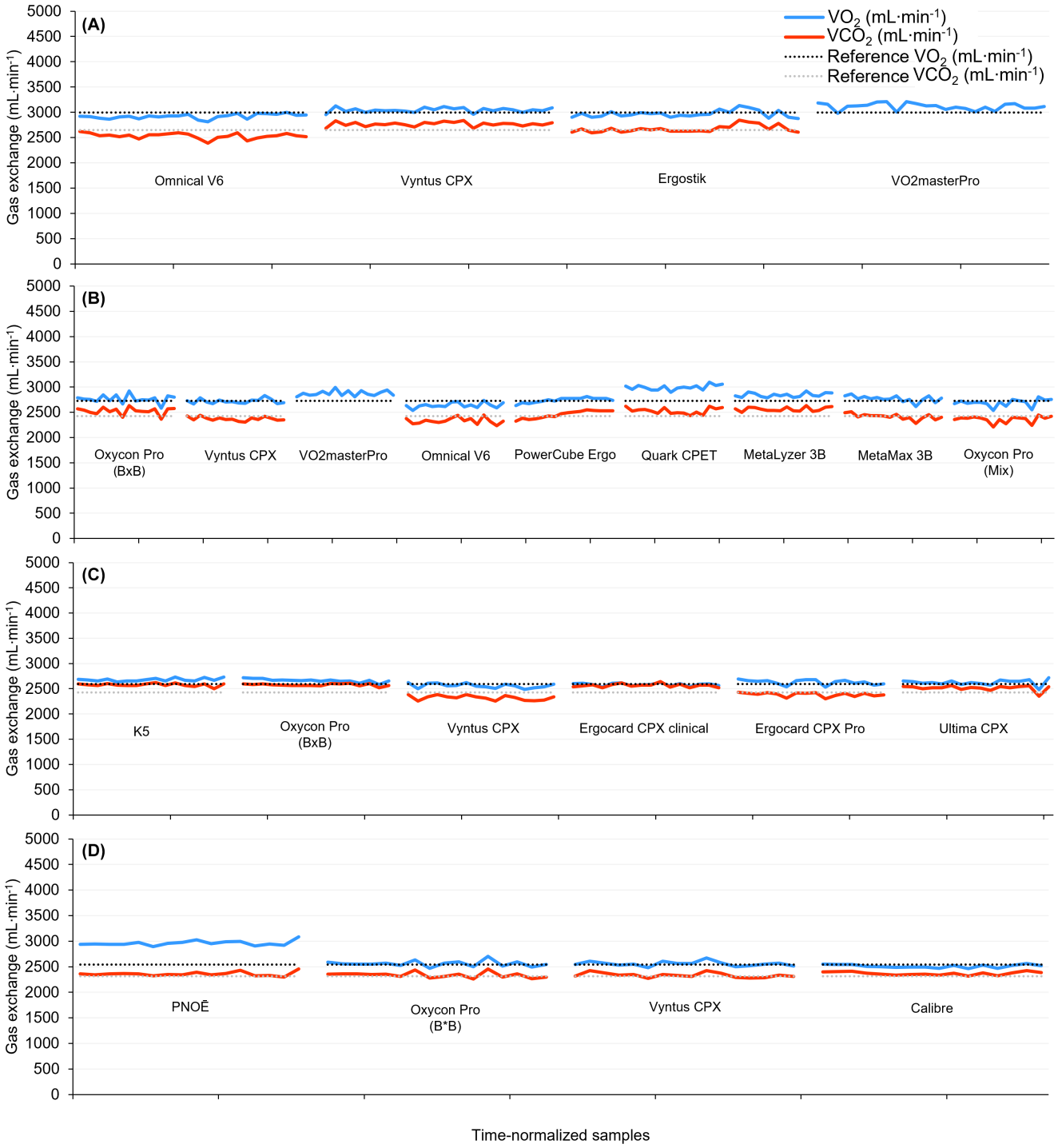

FIGURE 5 Measured $\dot{V}O_2$ and $\dot{V}CO_2$ during the cycling experiments in sessions 1 (A), 2 (B), 3 (C), and 4 (D). All $\dot{V}O_2$ and $\dot{V}CO_2$ values were first averaged over the two counterbalanced trials within each subject and then averaged between subjects. For all tests, reference values were calculated as specified in supplementary file I, section 3. $\dot{V}CO_2$, rate of carbon dioxide production; $\dot{V}O_2$, rate of oxygen uptake.

only three systems achieved an error <5% for the amount of energy derived from carbs, while five systems achieved an error <5% for energy derived from fats. Yet, 12 systems achieved an error <5% for total energy expenditure (Figure 2; Table 3).

When considering absolute errors, all but six systems exhibited an absolute percentage error <3% for assessing total energy expenditure during simulated exercise (Table S2, Figure S2). In contrast, none of the assessed systems showed an absolute percentage error of <5% for assessing the amount of energy derived from carbohydrates or fats. MetaMax 3B, for instance, showed a relatively small absolute percentage error of 1.9% in RER, but absolute percentage errors of ~39% and ~19% for energy

derived from carbohydrates and fats, respectively. Similarly, the absolute percentage error for RER was ~4.7% for Quark CPET, but this resulted in absolute percentage errors of 16% and 43% for energy derived from carbohydrates and fats, respectively. These findings suggest that substrate use at an individual level derived from most CPET systems should be interpreted with (great) caution. Moreover, even at a group level substrate use should be interpreted with caution, as some devices systematically under- or overestimated energy derived from carbohydrates and fats (Figure 2).

4.2 | Potential causes of observed errors

The largely comparable accuracy for most systems for assessing gas exchange variables during the simulated exercise (Figure 2) was achieved despite various methods used to measure volume, or O_2 and CO_2 gas concentrations (Table 1). However, some devices that used similar methods differed substantially in accuracy (e.g., Ultima CPX vs. Ergocard CPX Clinical, both from the same manufacturer, or Ergostik vs. VO2masterPro). This indicates that the different calibration methods, and the way the different measurement methodologies are integrated within the device's proprietary algorithms are important to the overall accuracy of the results, and accuracy can therefore not simply be inferred from the technical (hardware) specifications.

By examining the $\dot{V}E$, and fractions of O_2 and CO_2 in inspired and expired air, more insight can be gained into the potential causes of the errors in the measured respiratory gas variables. For example, PowerCube Ergo showed a rather large overestimation of $\dot{V}CO_2$ by 18% (Figure 2), but not $\dot{V}O_2$ or $\dot{V}E$ (both <3%). Therefore we can assume that the CO_2 sensor response was not accurate, despite duplicate gas calibration procedures. In support of this, the $FeCO_2$ value was 34% higher than the median value measured by other systems, which therefore leads to a higher $\dot{V}CO_2$ for a given flow and $FiCO_2$. As a result, the system yielded extremely large errors in the energy derived from carbohydrates and fats (Figure 2; Table 3). Similar inaccuracies in measured $\dot{V}CO_2$ were observed in pilot experiments for other manufacturers, suggesting CO_2 sensors in particular, require regular checks for accuracy to ensure accurate CPET results.

VO2masterPro underestimated $\dot{V}O_2$ by an average of 12%, with the underestimation also increasing with higher simulated $\dot{V}E$ (Figure 4). This increasing underestimation of $\dot{V}E$ suggests that the differential pressure sensor for measuring flow was primarily causing this error. Note that another manufacturer (Ergostik) showed only a small underestimation in $\dot{V}E$ despite also using a differential pressure sensor for measuring flow. This indicates that the method per se is not inaccurate. Inaccurate volume corrections might cause errors in $\dot{V}E$ measurement with the differential pressure sensor in VO2masterPro due to the differences int calibration procedures or algorithms.

Our findings also show how the calibration method might introduce errors. Specifically, the volumes of Ergocard CPX Clinical and CPX Pro both were calibrated using the 3 L volume stroke of the metabolic simulator, whereas the Ultima CPX was calibrated using the manufactures 3 L calibration syringe. The Ultima system underestimated $\dot{V}E$ by ~9%, whereas both other systems overestimated $\dot{V}E$ by ~6%, with this difference potentially being caused by the different calibration methods as all systems use a similar method for $\dot{V}E$ measurement and likely very similar proprietary algorithms for data processing.

4.3 | Wearable versus stationary, and breath-by-breath versus mixing chamber

Stationary devices such as Quark CPET, MetaLyzer 3B, and Vyntus CPX are often preferred in a lab setting over wearable (portable) devices because of the general perception that stationary devices exhibit a higher accuracy.[33] Our findings do however not necessarily support this notion, because some wearable devices showed similar or even better accuracy than the stationary devices. For example, the wearable COSMED K5 showed a ~1% point larger absolute percentage error compared to the stationary Quark CPET for assessing respiratory gas exchange variables (Table S1, Figure 3). Similarly, the overall absolute percentage error for the wearable MetaMax 3B from Cortex was 1% point smaller than the Cortex stationary MetaLyzer 3B. For both manufacturers, such differences likely fall within the technical standard error of measurement of repeated measures (Table S4 and S5), and thus suggests equivalent performance of these systems, in line with the similar methods employed for measuring volume and O_2 and CO_2 concentrations. This finding is in agreement with studies on older versions of these devices that suggested equivalent performance.[54] In contrast, other wearable systems (VO2masterPro and PNOĒ) showed lower accuracy than most stationary devices. VO2masterPro underestimated $\dot{V}O_2$ by an average of ~12%, while PNOĒ overestimated $\dot{V}O_2$ by an average of ~8.3% (Table 2, Figure 2). The (absolute) percentage error also increased with higher VE rates for VO2masterPro, indicating larger underestimation with higher volumes (Figure 2). While the absolute percentage error decreased for PNOĒ with higher $\dot{V}E$, the device did not measure any data when BF exceeded 60 breaths·min^{-1}, which may limit its

application to submaximal exercise testing. Furthermore, the PNOĒ manufacture guidelines state that the device requires only ambient air calibration. Yet, the errors were considerably larger when we assessed the device with only ambient air calibration (i.e., 4.9% overestimation of $\dot{V}O_2$, 16.3% underestimation of $\dot{V}CO_2$, and 17% underestimation of RER [supplementary file I, Figure S4]). These errors became smaller when we used a standard approach for calibration with CO_2/O_2 mix calibration gas, thus strongly suggesting calibration with certified calibration gasses is required when using this system. Nevertheless, even with the slight improvements as a result of this calibration, the errors for most outcomes remained (very) high (Figure 2).

Another portable device, Calibre, showed overall a very low (absolute) percentage error (~ −0.63%; Table 2, Figure 2). To the best of our knowledge, this is the only CPET device to employ machine learning to predict gas exchange variables from the measured values, which allowed it to achieve high accuracy, at a substantially lower cost than other (wearable) devices (Table 1). Moreover, in contrast to most other wearable devices, Calibre does not require the user to wear a data collection unit, which is beneficial for activities such as running, cycling, or and daily life activities where extra mass or restraints may influence performance and limit the ability to obtain valid measures.

While previous studies report mixing chamber systems to be more accurate at high volumes (i.e., $\dot{V}O_{2max}$ test),[24,36,55] we observed no apparent differences between OxyconPro in the mixing chamber mode or breath-by-breath mode. These conflicting findings may reflect the use of different systems in previous studies (all COSMED), and the volume at which devices were compared (up to 4.9 L·min^{-1} in[55] vs 4 L·min^{-1} in the present study). Note that one of the previous studies also used a metabolic simulator and found mixing chambers to be more accurate,[36] suggesting differences between the simulated and real breathing pattern are not the primary cause of these differences. Although some mixing chamber systems might thus be more accurate, they have a lower temporal resolution and need a longer time to achieve a steady state in gas exchange variables. This longer time required to reach a steady state may reduce the appearance of a plateau in $\dot{V}O_{2max}$.[55] We quantified the time to achieve steady state for the mixing chamber devices assessed in our study, with this being up to 3 min for Calibre, up to 90 s for Oxycon Pro mixing chamber and 140 s for Omnical V6. As some individuals may need a shorter time to achieve metabolic steady state (e.g., 60–90 s[56]), these findings suggests longer measurements may be required before this steady-state is also accurately reflected in the mixing chamber systems.

4.4 | Between-session reliability

While high accuracy of the measured gas exchange variables is important in many situations, a high reliability (i.e., low variability in repeated measures of the same simulated value) is important for repeated measurements. We quantified between-day reliability for a subset of devices that were available in the lab for >1 day by re-performing the same simulation experiments and computing the standard deviation of the recorded values between the days. Overall, the typical variation of the measured $\dot{V}O_2$ and $\dot{V}CO_2$ was <1.6% (Table S4 and S5) for all devices except for VO2masterPro and PNOĒ. Both these devices showed a rather substantial variation of >12% in the measured $\dot{V}O_2$ and/or $\dot{V}CO_2$ from day-to-day. These errors arose primarily as a result of variability in the accuracy of $\dot{V}E$ (CV of ~7%–8%, supplementary file II), and to a smaller extend variability in the measured O_2 fractions. However, for PNOĒ there also was a large (up to 37%) variability in CO_2 fractions. This suggests caution needs to be taken when using these devices as they were neither highly accurate (Figure 2), nor very reliable from day-to-day. Between-day variation for the other devices were relatively small for total energy expenditure (~0.8%), but larger for substrate use, ranging from 3.07%–68.5% for energy derived from carbohydrate and 2.8%–12.5% for energy derived from fats. Caution is therefore warranted when using CPET devices to estimate changes in substrate use and using these outcomes for guidance in for example weight management plans or nutritional optimization for athletes or patients. A considerable proportion in the changes of carbohydrate or fat metabolism may simply reflect technical measurement errors. These findings may explain the poor between-session reliability for peak fat oxidation observed previously.[57]

4.5 | Verification during human exercise

A metabolic simulator does not fully mimic human exercise; thus, we also compared all systems against each other during a steady-state human cycling test in well-trained individuals. The relative differences between systems in these cycling experiments did mostly, but not always match the relative differences in the metabolic simulator experiments. Quark CPET, for instance, showed a very low mean relative percentage error for assessing $\dot{V}O_2$ in the simulation experiments (overestimation by 0.60%; Figure 2, Table 2). Yet, it recorded ~10% higher $\dot{V}O_2$ values compared to reference value during the cycling experiments (Figure 5, Table S7). Similarly, VO2masterPro underestimated $\dot{V}O_2$ by an average of ~12% in the simulation

experiments, but overestimated $\dot{V}O_2$ by a magnitude of ~4%–5.5% during cycling test 1 and 2.

One reason for the discrepancy between the simulation and human exercise results is that the accuracy during the cycling experiments is influenced by biological variability, so that only a small part of the variability between systems reflects measurement error.[34] Our findings indirectly support this finding and suggest that care should be taken when comparing devices to assess their accuracy. However, the observed differences may also have some technical basis because the relative difference for the majority of devices was overall in line with the simulation experiments. A potential reason for differences is that some devices exhibit a different breathing resistance, which increases $\dot{V}O_2$ during the human tests, but it does not affect the measured value during simulation experiments.[21] While the participants subjectively noticed differences in breathing resistance between some devices, the effect of higher breathing resistance on $\dot{V}O_2$ is expected to be negligible in contemporary devices,[21,58] making this an unlikely explanation. Another reason for the discrepancy is that the exhaled human air temperature for systems like Quark CPET and VO2masterPro is assumed to be higher than the temperature of the expired air assumed by other devices. This may cause the gas volume to be overestimated in the human tests for these devices because the volume of a gas is directly proportional to its temperature. However, Quark CPET assumed the temperature of the exhaled air to be 31°C, while VO2masterPro assumed an exhaled temperature of 34°C and these assumptions are largely similar to most other devices (e.g., 31°C for the Vyaire and Cortex systems), and thus unlikely to (fully) explain the relatively higher values in the human tests as opposed to the simulation tests. Indeed, a 3°C increase in assumed temperature would explain only a ~ 2% higher $\dot{V}E$ and thus $\dot{V}O_2$ for VO2masterPro. A final reason is that humidity inside the volume, O_2 or CO_2 sensors may have interfered with the human measurements, which in turn caused up to a ~ 10%–18% increase in the recorded $\dot{V}O_2$ and $\dot{V}CO_2$ for some devices. For example, in non-dispersive infrared sensors typically used for assessing CO_2 concentrations (Table 1), H_2O molecules may lead to absorption of infrared light in addition to CO_2 molecules, which could lead to an overestimation of the CO_2 concentration. Similarly, H_2O molecules are also paramagnetic and could thus affect the accuracy of paramagnetic fuel cells for measuring O_2 concentrations. The difference between devices in the potential effect of humidity during the human tests may reflect the design-specific ways that different systems use to control for the effect of humidity in the measured air. Yet even the same method may lead to different accuracies over time. For example, some systems use a PermaPure nafion sample line in the gas sampling circuit to control for humidity on the sensor output signal. This membrane selectively removes water vapor from the measured gas, while allowing other gasses to pass through. The membranes can however become saturated with water vapor over time, which can decrease its effectiveness in removing water vapor from the gas stream and lead to inaccurate measurements. These findings therefore also highlight the importance of human verification in addition to simulation testing with dry gas.

4.6 | Comparison with other studies

A small number of other studies used a metabolic simulator to assess the accuracy of CPET devices, with most of these studies assessing solely COSMED (K4/5 and Quark) devices.[34,36–43,59] For example, Beijst and colleagues[36] reported relative percentage errors of 9%–12% and 5%–7% for $\dot{V}O_2$ and $\dot{V}CO_2$, respectively in the Quark device in breath-by-breath mode over a similar simulated range as in our study. These errors are larger than found in our study, with relative percentage errors ranging from −1.6% to 1.7% for $\dot{V}O_2$ and − 7.1 to −0.7% for $\dot{V}CO_2$ in our study. The smaller errors observed in the present study may primarily reflect differences in the device calibration procedures with the volume sensor of Quark being calibrated against the simulator in the present study, and potentially in gas analysis sensor sensitivity (e.g., new device as provided by the manufacturer in the present study vs a potentially older device in the prior study). In contrast, while the K5 device in our study showed a largely comparable mean relative percentage error for $\dot{V}E$ as compared to a previous study (−0.8% vs. −0.5% in[34]), mean errors for $\dot{V}O_2$ and $\dot{V}CO_2$ were larger in the present study (−7.8% vs. −0.04% and − 6.0% vs. −1.03%, respectively). These differences may in part also be attributed to sensor sensitivity, as well as differences in the simulation protocol (e.g., $\dot{V}O_2$ range), and simple between-day variability (see also Tabe S4 and S5). In support of sensor sensitivity and calibration procedures as being the primary determinants of differences, one other study assessed the Vyntus device against a Relitech and Vacumed simulator and showed errors below 3% for all gas exchange levels up to 80 breath/min, which is comparable to our findings.[37] In this context, the PowerCube Ergo also showed relatively large errors in a previous simulation study,[59] thus suggesting the large errors observed in our study do not reflect an incidentally poorly performing device.

Most devices have been assessed for accuracy by comparing them with other devices during real (human) exercise. Among these studies, a large relative percentage error has also been reported for PNOĒ when compared

to the Quark device (34% overestimation of $\dot{V}O_2$, 57% overestimation of $\dot{V}CO_2$),[29] which is approximately in line with our findings during the simulation experiments (Figures 2 and 4). The error observed in our study was however smaller, potentially due to the use of calibration gas as opposed to ambient air calibration as recommended by the manufacturer. For VO2masterPro, a previous study showed this device to underestimate $\dot{V}O_2$ during low-intensity cycling experiments, but overestimate $\dot{V}O_2$ at high intensities when compared to the Parvomedics metabolic cart.[26] Such findings are in partial agreement with our findings as we found a consistent underestimation during the simulation experiments, with this difference becoming larger at higher simulated values. However, these findings do not agree with the cycling experiments, where $\dot{V}O_2$ was slightly overestimated.

A different comparison can be made between the error of devices as measured during (simulated) exercise (present study) and (methanol) combustion studies. In one such study,[23] the Omnical, Quark and Parvomedics devices were shown to exhibit an absolute error of <2% for all assessed outcomes ($\dot{V}O_2$, $\dot{V}CO_2$, RER), while the Oxycon Pro showed relatively large errors. These findings partially contrast our study where the Oxycon Pro showed a very high accuracy on these outcomes (1.36 to 1.76% absolute error for B*B and mixing chamber respectively), with both Omnical and Quark showing intermediate accuracy (2.44% and 3.32%, Table S1). Another study simulating basal metabolic rate also found the Omnical to exhibit the highest accuracy among the investigated devices.[25] The discrepancy between these previous and our findings may primarily be related to the higher flow rate during (simulated) exercise as opposed to combustion experiments or simulated basal metabolic rates. In exercise experiments, the accuracy of volume measurements may also become more critical, whereas combustion experiments primarily assess the accuracy of the sensors that assess gas concentrations.

Overall, these findings indicate that the results of the present study, with all devices undergoing the same protocol and test procedures enables a fair comparison between devices.

4.7 | Limitations

A first limitation is that while the range in simulated $\dot{V}O_2$ corresponds to the range in $\dot{V}O_2$ observed in the literature for recreational and well-trained individuals,[60–62] it is lower than reported for samples of elite athletes.[4] For example, a $\dot{V}O_2$ of 5500 mL·min^{-1} would be required to mimic a $\dot{V}O_{2max}$ of 79 mL·kg^{-1}·min^{-1} for a 70 kg individual. However, a high BF may arguably be the most challenging component for sensors, and this did approach peak values reported in the literature. Although we attempted to extrapolate the error at higher than simulated volumes, the change in error with volume increases was highly variable for some systems (Figure 4), which therefore did not allow us to accurately extrapolate the error to higher than simulated values (e.g., $\dot{V}O_2$ 5000 or 6000 mL·min^{-1}). Nevertheless, a strength is that the cycling experiments in our study were performed at a higher intensity than most prior studies, which adds more relevance to exercise situations in trained individuals. The average $\dot{V}O_2$ during cycling in a previous study was ~1400 mL·min^{-1}[163] and was on average ~ 2600–3000 mL·min^{-1} in our study (Figure 5, Table S7). This submaximal $\dot{V}O_2$ for the participants in the present study corresponds to a maximum intensity for lesser trained individuals. A second limitation is that the time required to reach a steady state was determined visually (Figure S3). The exact time period at which a steady state is achieved is, therefore, arbitrary and may vary between observers. Nevertheless, we used a conservative approach to maximize the chance of achieving a steady state when using these values in practice. A third limitation is that we assessed only one device from each manufacturer, and it remains unknown if the devices assessed reflect the accuracy of the devices in-field. We are currently undertaking a follow-up field study to get more insights on this. Related, the relatively small number of datapoints also reduced the power of the statistical test used to objectively assess agreement. Some devices that did not achieve good or acceptable statistical agreement may therefore still achieve this with a larger dataset.

4.8 | Perspective

Whether the magnitude of under- or overestimation in $\dot{V}O_2$, $\dot{V}CO_2$, substrate use, and energy expenditure is relevant for practical applications depends on the context. A first consideration in this regard is related to whether a single individual or multiple individuals are being measured. When a single individual is measured once, there is a larger potential for error as underestimation in one test and overestimation in another cannot rule each other out. In such situations, the absolute percentage errors would best reflect the potential error (Figure 3 and S2, supplementary file I, tables S1 and S2). Depending on the outcome considered and the device used, the error in such situations could influence clinical decision-making. An absolute percentage error of 10% for $\dot{V}O_2$ could for instance result in a fireman not meeting a predefined $\dot{V}O_{2max}$ value required to continue their profession[11] and patients not meeting a predefined $\dot{V}O_{2max}$ value advised to undergo major surgery[12] or delay medical treatment.[13] Conversely,

it could also lead to these individuals falsely meeting the criteria, which increases subsequent risks during the profession in the case of the fireman, or during surgery for patients. For world-class athletes, even small differences in $\dot{V}O_{2max}$ (e.g., <1.5%) could lead to relevant inaccuracies in performance predictions (e.g.,[64]), or talent identification.[65] Similarly, the typically large absolute percentage errors for substrate use suggest particular caution when assessing substrate use of a single individual. This caution is also warranted when doing repeated measurements as the measured values differed substantially between different days (Table S4, S5). Even the generally highly accurate Oxycon Pro, for instance, showed an absolute percentage difference of ~9% in the energy derived from fats between two repeated measurements, which would therefore require substantial alterations in substrate oxidation at an individual level to be detected, in particular when combined with biological variability. We therefore strongly recommend CPET users to perform multiple repeated measurements to reduce the impact of both technical and biological measurement error.

When assessing multiple individuals or performing multiple assessments of the same individual, underestimation in one test and overestimation in another *can* rule each other out, resulting in a lower overall error (Tables 2 and 3). The relative percentage errors may be most relevant in this situation. When considering these errors, some devices systematically under- or overestimate $\dot{V}O_2$ and $\dot{V}CO_2$ (Figures 2 and 4). This is important to consider when comparing these results to those measured in other studies obtained with a different device, such as when comparing running economy, cycling efficiency or $\dot{V}O_{2max}$ between different populations measured in different studies with different brand devices. As an example, K5 is expected to underestimate the oxygen cost of exercise by an average of ~8%, which could lead to overly optimistic values for cycling efficiency or running economy, but overly pessimistic value for $\dot{V}O_{2max}$. Similarly, the MetaLyzer 3B on average overestimated the energy derived from carbs by ~53%, and underestimated the energy derived from fats by ~25%, which could have important consequences for studies interested in quantifying substrate use during exercise and subsequent nutritional recommendations.

It is important to note that differences in substrate use and total energy expenditure may be even larger when using the estimated energy derived from carbohydrates and fats or total energy expenditure determined by the manufacturers due to different equations being available to estimate these.[66] For that reason the same equation[51] was used in the current study to calculate energy expenditure and substrate utilization from $\dot{V}O_2$ and $\dot{V}CO_2$ for all manufacturers. The equation used is considered the most accurate to estimate substrate use during exercise as compared to the ^{13}C:^{12}C ratio technique.[67] Notably, while most devices exhibited an absolute percentage error for total energy expenditure of <6% (Table S2), three devices (i.e., Ultima, K5, and PNOĒ) exhibited an error of 6%–9%. Although this may be regarded as relatively large, all devices were still more accurate in estimating energy expenditure than even the best-performing wearable-inertial-measurement-unit-based system (13% error), and in particular when compared to smartwatches (42% error) or heart rate-based estimates.[19] This therefore suggests energy expenditure derived from even lower accuracy (portable) systems has some utility over wearable-based estimates of energy expenditure.

Another implication is related to threshold determination during exercise. Errors in either $\dot{V}O_2$ or $\dot{V}CO_2$ can impact the determination of threshold inflection points used to demarcate training zones, with the magnitude of the error depending on the method used, and the amplitude and direction of the error in respiratory gas exchange variables. For example, when we modeled a proportionally larger underestimation of $\dot{V}CO_2$ with higher $\dot{V}E$ as observed in some devices (Figure 4), the gas exchange threshold as determined using the 'V-slope' method occurred at a lower workload/$\dot{V}O_2$ (see supplementary file I, Figure S5). Errors in threshold inflection points may particularly impact patient populations that require strict control of exercise intensity (e.g., ischemic heart disease or congestive heart failure), but also athletes that may as a result be performing a large volume of training at an inappropriate intensity.

The findings of this study may be used by clinicians, researchers, medical performance staff, sports practitioners, and coaches as guidance on which device to buy for metabolic exercise testing. Here we therefore provide some considerations when using these findings to this purpose. Two important factors to consider when purchasing a device often include its price and accuracy. Interestingly, our findings show only a small correlation of $r = -0.13$ between the approximate price (Table 1) and overall accuracy (Table S8) of CPET devices, highlighting that more expensive devices are not necessarily more accurate (supplementary File I, Figure S6). This discrepancy between price and accuracy may at least partly be related to additional software and hardware functionalities among devices, that notably also need to be considered within a purchase decision. For example, some devices (e.g., Vyntus) include an automatic volume and gas calibration option, while this must be performed manually for other devices. Similarly, some devices include an automated determination of physiological outcomes such as the first and second ventilatory thresholds, or VO_{2peak}, while this needs to be manually determined for others. While automated determination of physiological outcomes always

needs to be confirmed by an experienced individual, the automated determination may save time. Moreover, some devices are wearable and thus allow for measurements in-field. While these devices are typically more expensive when compared to the stationary device from the same manufacturer, they may be useful for individuals that are working with athletes. Another important consideration in this context is the choice between breath-by-breath and mixing chamber devices. While breath-by-breath devices exhibit a higher temporal resolution, some findings[36,55] and anecdotal observations suggest that their accuracy is compromised at very high exercise intensities seen in world-class athletes, thus potentially necessitating mixing chamber devices for accurate measurement in these situations. Finally, some devices allow integration of other measurement tools such as electrocardiogram, blood pressure, and oxygen saturation, and this may also be an important consideration for some purposes. Given all data and additional considerations discussed in this paper, we cannot recommend one device as best to use for all purposes. Which device to choose needs to be decided in the context of its intended use, required precision and accuracy in the context of the application, the skills of the staff, availability of internal/external support, durability, and financial budget possibilities. Nevertheless, when solely considering accuracy, the devices that perform relatively well (i.e., <5% average absolute percentage error over both gas exchange and substrate/energy outcomes; Table S8) include Oxycon Pro, Vyntus CPX, Calibre and Ergocard Pro. Devices with slightly lower but still acceptable accuracy (5%–6% average overall absolute percentage error) include Omnical V6 and K5. In contrast, devices that show low relative accuracy (absolute percentage errors >20%) and/or reliability include VO2masterPro, PNOĒ, and PowerCube Ergo.

5 | CONCLUSION

The error of $\dot{V}E$, BF, $\dot{V}O_2$, $\dot{V}CO_2$, and RER during simulated exercise is generally <5% but differs substantially between systems. A large variability in accuracy was also observed for substrate utilization, suggesting substrate utilization derived from indirect calorimetry during exercise should be particularly interpreted with caution. The observed errors may impact outcomes derived from CPET measurements such as $\dot{V}O_{2max}$, exercise economy, and thresholds inflection points used for zone demarcation.

Our findings also indicate substantial variability in between-day accuracy for some devices. This impacts the validity of repeated testing of one individual, and it may also affect the accuracy of comparisons between small subject groups.

Another notable finding is that the performance of mixing chamber devices did not substantially differ from breath-by-breath devices in the investigated range, and some wearable devices yielded similar accuracy as state-of-the-art stationary devices.

Moreover, devices with similar technical specifications could still show substantial differences in their accuracy. This overall highlights the need to assess the accuracy of each individual device as the accuracy is likely not only dependent on the hardware, but also on proprietary software algorithms.

Finally, the findings from the human experiments highlight the importance of human verification in addition to simulation testing with dry gas for a comprehensive assessment of accuracy.

ACKNOWLEDGEMENTS

No funding was received. The authors would like to thank Manon Broekhuijsen from Maastricht University Medical Center+ for allowing us to borrow the metabolic simulator. Additionally, we would like to thank Aimee Boersen, Remy Queisen, and Skip Veugen for their assistance during the experiments and/or data analysis. Finally, we would like to thank all manufacturers that provided equipment and send staff for their cooperation.

CONFLICT OF INTEREST STATEMENT

The authors do not report any conflicts of interest or relationships with any of the companies. The results of the study are presented clearly, honestly, and without fabrication, falsification, or inappropriate data manipulation.

DATA AVAILABILITY STATEMENT

The data that supports the findings of this study are available in the supplementary material of this article

ORCID

Bas Van Hooren https://orcid.org/0000-0001-8163-693X
Bart C. Bongers https://orcid.org/0000-0002-1948-9788

REFERENCES

1. Iannetta D, Inglis EC, Mattu AT, et al. A critical evaluation of current methods for exercise prescription in women and men. *Med Sci Sports Exerc*. 2020;52(2):466-473. doi:10.1249/MSS.0000000000002147
2. Keir DA, Iannetta D, Mattioni Maturana F, Kowalchuk JM, Murias JM. Identification of non-invasive exercise thresholds: methods, strategies, and an online app. *Sports Med*. 2022;52(2):237-255. doi:10.1007/s40279-021-01581-z
3. Vainshelboim B, Arena R, Kaminsky LA, Myers J. Reference standards for ventilatory threshold measured with cardiopulmonary exercise testing: the fitness registry and the importance

of exercise: a National Database. *Chest*. 2020;157(6):1531-1537. doi:10.1016/j.chest.2019.11.022
4. Jones AM, Kirby BS, Clark IE, et al. Physiological demands of running at 2-hour marathon race pace. *J Appl Physiol*. 2021;130(2):369-379. doi:10.1152/japplphysiol.00647.2020
5. Koutlianos N, Dimitros E, Metaxas T, Cansiz M, Deligiannis A, Kouidi E. Indirect estimation of VO2max in athletes by ACSM's equation: valid or not? *Hippokratia*. 2013;17(2):136-140.
6. Ross R, Blair SN, Arena R, et al. Importance of assessing cardiorespiratory fitness in clinical practice: a case for fitness as a clinical vital sign: a scientific statement from the American Heart Association. *Circulation*. 2016;134(24):e653-e699.
7. Burnley M, Jones AM. Oxygen uptake kinetics as a determinant of sports performance. *Eur J Sport Sci*. 2007;7(2):63-79. doi:10.1080/17461390701456148
8. Burke LM, Ross ML, Garvican-Lewis LA, et al. Low carbohydrate, high fat diet impairs exercise economy and negates the performance benefit from intensified training in elite race walkers. *J Physiol*. 2017;595(9):2785-2807. doi:10.1113/JP273230
9. Cano A, Ventura L, Martinez G, et al. Analysis of sex-based differences in energy substrate utilization during moderate-intensity aerobic exercise. *Eur J Appl Physiol*. 2022;122(7):29-70. doi:10.1007/s00421-022-04961-z
10. Van Hooren B, Cox M, Rietjens G, Plasqui G. Determination of energy expenditure in professional cyclists using power data: validation against doubly-labelled water. *Scand J Med Sci Sports*. 2022;33(4):407-419.
11. Peate WF, Lundergan L, Johnson JJ. Fitness self-perception and $\dot{V}o_2$max in firefighters. *J Occup Environ Med*. 2002;44(6):546-550.
12. Levett DZH, Jack S, Swart M, et al. Perioperative cardiopulmonary exercise testing (CPET): consensus clinical guidelines on indications, organization, conduct, and physiological interpretation. *Br J Anaesth*. 2018;120(3):484-500. doi:10.1016/j.bja.2017.10.020
13. Groen WG, Naaktgeboren WR, van Harten WH, et al. Physical fitness and chemotherapy tolerance in patients with early-stage breast cancer. *Med Sci Sports Exerc*. 2022;54(4):537-542. doi:10.1249/MSS.0000000000002828
14. Stellingwerff T, Heikura IA, Meeusen R, et al. Overtraining syndrome (OTS) and relative energy deficiency in sport (RED-S): shared pathways, symptoms and complexities. *Sports Med*. 2021;51(11):1-30. doi:10.1007/s40279-021-01491-0
15. Brownstein CG, Pastor FS, Mira J, Murias JM, Millet GY. Power output manipulation from below to above the gas exchange threshold results in exacerbated performance fatigability. *Med Sci Sports Exerc*. 2022;54(11):1947-1960. doi:10.1249/MSS.0000000000002976
16. Nikooie R, Gharakhanlo R, Rajabi H, Bahraminegad M, Ghafari A. Noninvasive determination of anaerobic threshold by monitoring the% SpO2 changes and respiratory gas exchange. *J Strength Cond Res*. 2009;23(7):2107-2113.
17. Kang SK, Kim J, Kwon M, Eom H. Objectivity and validity of EMG method in estimating anaerobic threshold. *Int J Sports Med*. 2014;35(9):737-742. doi:10.1055/s-0033-1361182
18. Düking P, Van Hooren B, sperlich B. Assessment of peak oxygen uptake with a smartwatch and its usefulness for running training. *Int J Sports Med*. 2022;43(7):642-647.
19. Slade P, Kochenderfer MJ, Delp SL, Collins SH. Sensing leg movement enhances wearable monitoring of energy expenditure. *Nat Commun*. 2021;12(1):4312. doi:10.1038/s41467-021-24173-x
20. Macfarlane DJ. Automated metabolic gas analysis systems: a review. *Sports Med*. 2001;31(12):841-861. doi:10.2165/00007256-200131120-00002
21. Ward SA. Open-circuit respirometry: real-time, laboratory-based systems. *Eur J Appl Physiol*. 2018;118(5):875-898.
22. Macfarlane DJ. Open-circuit respirometry: a historical review of portable gas analysis systems. *Eur J Appl Physiol*. 2017;117(12):2369-2386. doi:10.1007/s00421-017-3716-8
23. Kaviani S, Schoeller DA, Ravussin E, et al. Determining the accuracy and reliability of indirect calorimeters utilizing the methanol combustion technique. *Nutr Clin Pract*. 2018;33(2):206-216. doi:10.1002/ncp.10070
24. Perez-Suarez I, Martin-Rincon M, Gonzalez-Henriquez JJ, et al. Accuracy and precision of the COSMED K5 portable Analyser. *Front Physiol*. 2018;9:1764. doi:10.3389/fphys.2018.01764
25. Alcantara JMA, Galgani JE, Jurado-Fasoli L, et al. Validity of four commercially available metabolic carts for assessing resting metabolic rate and respiratory exchange ratio in non-ventilated humans. *Clin Nutr*. 2022;41(3):746-754. doi:10.1016/j.clnu.2022.01.031
26. Montoye AHK, Vondrasek JD, Hancock JB 2nd. Validity and reliability of the VO2 master pro for oxygen consumption and ventilation assessment. *Int J Exerc Sci*. 2020;13(4):1382-1401.
27. Howe CC, Matzko RO, Piaser F, Pitsiladis YP, Easton C. Stability of the K4b(2) portable metabolic analyser during rest, walking and running. *J Sports Sci*. 2014;32(2):157-163. doi:10.1080/02640414.2013.812231
28. Dieli-Conwright CM, Jensky NE, Battaglia GM, McCauley SA, Schroeder ET. Validation of the CardioCoachCO2 for submaximal and maximal metabolic exercise testing. *J Strength Cond Res*. 2009;23(4):1316-1320. doi:10.1519/JSC.0b013e3181a3c5e8
29. Tsekouras YE, Tambalis KD, Sarras SE, Antoniou AK, Kokkinos P, Sidossis LS. Validity and reliability of the new portable metabolic analyzer PNOE. *Front Sports Act Living*. 2019;1:24. doi:10.3389/fspor.2019.00024
30. Crouter SE, LaMunion SR, Hibbing PR, Kaplan AS, Bassett DR Jr. Accuracy of the cosmed K5 portable calorimeter. *PloS One*. 2019;14(12):e0226290. doi:10.1371/journal.pone.0226290
31. Rosdahl H, Gullstrand L, Salier-Eriksson J, Johansson P, Schantz P. Evaluation of the Oxycon Mobile metabolic system against the Douglas bag method. *Eur J Appl Physiol*. 2010;109(2):159-171. doi:10.1007/s00421-009-1326-9
32. Nieman DC, Austin MD, Dew D, Utter AC. Validity of COSMED's quark CPET mixing chamber system in evaluating energy metabolism during aerobic exercise in healthy male adults. *Res Sports Med*. 2013;21(2):136-145.
33. Brehm MA, Harlaar J, Groepenhof H. Validation of the portable VmaxST system for oxygen-uptake measurement. *Gait Posture*. 2004;20(1):67-73. doi:10.1016/S0966-6362(03)00097-3
34. Guidetti L, Meucci M, Bolletta F, Emerenziani GP, Gallotta MC, Baldari C. Validity, reliability and minimum detectable change of COSMED K5 portable gas exchange system in breath-by-breath mode. *PloS One*. 2018;13(12):e0209925. doi:10.1371/journal.pone.0209925
35. Shephard RJ. Open-circuit respirometry: a brief historical review of the use of Douglas bags and chemical analyzers. *Eur J Appl Physiol*. 2017;117(3):381-387. doi:10.1007/s00421-017-3556-6
36. Beijst C, Schep G, Breda E, Wijn PF, Pul C. Accuracy and precision of CPET equipment: a comparison of breath-by-breath and

mixing chamber systems. *J Med Eng Technol*. 2013;37(1):35-42. doi:10.3109/03091902.2012.733057
37. Souren T, Rose E, Groepenhoff H. Comparison of two metabolic simulators used for gas exchange verification in cardiopulmonary exercise test carts. *Front Physiol*. 2021;12:667386. doi:10.3389/fphys.2021.667386
38. Ballal MA, Macdonald IA. An evaluation of the oxylog as a portable device with which to measure oxygen consumption. *Clin Phys Physiol Meas*. 1982;3(1):57-65. doi:10.1088/0143-0815/3/1/005
39. Rodriguez FA, Keskinen KL, Kusch M, Hoffmann U. Validity of a swimming snorkel for metabolic testing. *Int J Sports Med*. 2008;29(2):120-128. doi:10.1055/s-2007-964973
40. Carter J, Jeukendrup AE. Validity and reliability of three commercially available breath-by-breath respiratory systems. *Eur J Appl Physiol*. 2002;86(5):435-441. doi:10.1007/s00421-001-0572-2
41. Baldari C, Meucci M, Bolletta F, Gallotta M, Emerenziani G, Guidetti L. Accuracy and reliability of COSMED K5 portable metabolic device versus simulating system. *Sport Sci Health*. 2015;11(1):58.
42. Macfarlane DJ, Wong P. Validity, reliability and stability of the portable cortex Metamax 3B gas analysis system. *Eur J Appl Physiol*. 2012;112(7):2539-2547. doi:10.1007/s00421-011-2230-7
43. Prieur F, Castells J, Denis C. A methodology to assess the accuracy of a portable metabolic system (VmaxST). *Med Sci Sports Exerc*. 2003;35(5):879-885. doi:10.1249/01.MSS.0000065003.82941.B0
44. Turner N, Parker J, Hudnall J. The effect of dry and humid hot air inhalation on expired relative humidity during exercise. *Am Ind Hyg Assoc J*. 1992;53(4):256-260. doi:10.1080/15298669291359618
45. Atkins KJ, Thompson MW, Ward JJ, Kelly PT. Expired air temperature during prolonged exercise in cool- and hot-humid environments. *Eur J Appl Physiol Occup Physiol*. 1997;76(4):352-355. doi:10.1007/s004210050260
46. Younes M, Kivinen G. Respiratory mechanics and breathing pattern during and following maximal exercise. *J Appl Physiol Respir Environ Exerc Physiol*. 1984;57(6):1773-1782. doi:10.1152/jappl.1984.57.6.1773
47. Carey D, Pliego G, Raymond R. How endurance athletes breathe during incremental exercise to fatigue interaction of tidal volume and frequency. *J Exerc Physiol Online*. 2008;11(4):44-51.
48. Bongers BC, Hulzebos EH, Van Brussel M, Takken T. *Pediatric norms for cardiopulmonary exercise testing: in relation to sex and age (second edition)*. Uitgeverij BOXPress; 2014.
49. Blackie SP, Fairbarn MS, McElvaney NG, Wilcox PG, Morrison NJ, Pardy RL. Normal values and ranges for ventilation and breathing pattern at maximal exercise. *Chest*. 1991;100(1):136-142. doi:10.1378/chest.100.1.136
50. Lucia A, Carvajal A, Calderon FJ, Alfonso A, Chicharro JL. Breathing pattern in highly competitive cyclists during incremental exercise. *Eur J Appl Physiol Occup Physiol*. 1999;79(6):512-521. doi:10.1007/s004210050546
51. Jeukendrup AE, Wallis GA. Measurement of substrate oxidation during exercise by means of gas exchange measurements. *Int J Sports Med*. 2005;26 Suppl 1(S 1):S28-S37. doi:10.1055/s-2004-830512
52. Shieh G. Assessing agreement between two methods of quantitative measurements: exact test procedure and sample size calculation. *Statis Biopharm Res*. 2020;12(3):352-359. doi:10.1080/19466315.2019.1677495
53. Graham BL, Steenbruggen I, Miller MR, et al. Standardization of spirometry 2019 update. An official American Thoracic Society and European Respiratory Society technical statement. *Am J Respir Crit Care Med*. 2019;200(8):e70-e88. doi:10.1164/rccm.201908-1590ST
54. Meyer T, Davison RC, Kindermann W. Ambulatory gas exchange measurements—current status and future options. *Int J Sports Med*. 2005;26 Suppl 1(S 1):S19-S27. doi:10.1055/s-2004-830507
55. Winkert K, Kirsten J, Kamnig R, Steinacker JM, Treff G. Differences in V O2max measurements between breath-by-breath and mixing-chamber mode in the COSMED K5. *Int J Sports Physiol Perform*. 2021;16(9):1335-1340.
56. Whipp BJ. Physiological mechanisms dissociating pulmonary CO2 and O2 exchange dynamics during exercise in humans. *Exp Physiol*. 2007;92(2):347-355. doi:10.1113/expphysiol.2006.034363
57. Chrzanowski-Smith OJ, Edinburgh RM, Thomas MP, et al. The day-to-day reliability of peak fat oxidation and FAT(MAX). *Eur J Appl Physiol*. 2020;120(8):1745-1759. doi:10.1007/s00421-020-04397-3
58. Dressendorfer RH, Wade CE, Bernauer EM. Combined effects of breathing resistance and hyperoxia on aerobic work tolerance. *J Appl Physiol Respir Environ Exerc Physiol*. 1977;42(3):444-448. doi:10.1152/jappl.1977.42.3.444
59. Souren T, De Bliek E, Roeykens J, De Soomer K, Oostveen E. Quality control of cardiopulmonary exercise equipment. *Eur Respir J*. 2020;56(suppl 64):3786. doi:10.1183/13993003.congress-2020.3786
60. Riboli A, Coratella G, Rampichini S, Limonta E, Esposito F. Testing protocol affects the velocity at VO(2max) in semi-professional soccer players. *Res Sports Med*. 2022;30(2):182-192. doi:10.1080/15438627.2021.1878460
61. Gaskill SE, Ruby BC, Walker AJ, Sanchez OA, Serfass RC, Leon AS. Validity and reliability of combining three methods to determine ventilatory threshold. *Med Sci Sports Exerc*. 2001;33(11):1841-1848.doi:10.1097/00005768-200111000-00007
62. Nixon RJ, Kranen SH, Vanhatalo A, Jones AM. Steady-state above MLSS: evidence that critical speed better represents maximal metabolic steady state in well-trained runners. *Eur J Appl Physiol*. 2021;121(11):3133-3144. doi:10.1007/s00421-021-04780-8
63. Akkermans MA, Sillen MJ, Wouters EF, Spruit MA. Validation of the oxycon mobile metabolic system in healthy subjects. *J Sports Sci Med*. 2012;11(1):182-183.
64. Van Hooren B, Lepers R. A physiological comparison of the new-over 70 years of age-marathon record holder and his predecessor: a case report. *Front Physiol*. 2023;14:1122315. doi:10.3389/fphys.2023.1122315
65. IJzerman J, Damen T, Koens G, Collée T. Improving talent identification and development in young distance runners. *New Stud Athlet*. 2008;23(3):35-48.
66. Kipp S, Byrnes WC, Kram R. Calculating metabolic energy expenditure across a wide range of exercise intensities: the equation matters. *Appl Physiol Nutr Metab*. 2018;43(6):639-642. doi:10.1139/apnm-2017-0781

67. González-Haro C. Concordance between 13C: 12C ratio technique respect to indirect calorimetry to estimate carbohydrate and fat oxidation rates by means stoichiometric equations during exercise. A reliability and agreement study. *Physiol Rep*. 2019;7(8):e14053.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Van Hooren B, Souren T, Bongers BC. Accuracy of respiratory gas variables, substrate, and energy use from 15 CPET systems during simulated and human exercise. *Scand J Med Sci Sports*. 2024;34:e14490. doi:10.1111/sms.14490