Dual $VO_{2plateau}$ Method for Confirming VO_{2max} Validity in Trained Female Runners

Original Research

Savanna N. Knight[1,2], Eric M. Scudamore[3], Lynnsey R. Bowling[1,4], Veronika Scudamore[3], Hunter S. Waldman[1], Eric K. O'Neal[1]

[1]*Department of Kinesiology, University of North Alabama, Florence, AL | USA*
[2]*Department of Health & Human Performance, Texas State University, San Marcos, TX | USA*
[3]*Department of Health, Physical Education, and Sport Sciences, Arkansas State University, Jonesboro AR | USA*
[4]*Department of Kinesiology, University of Wisconsin Eau Claire, Eau Claire, WI*

Open Access

Published: February 14, 2026

Research in Strength and Performance: *2026, Volume 6 (Issue 1): 6*

ISSN: 3069-0765

Abstract

Introduction: Breath-by-breath (BxB) data in modern metabolic carts exhibit high variability and are rarely filtered. With sampling intervals (SI) ≤30-s commonly incorporated, not accounting for outlier breaths can greatly inflate maximal oxygen consumption (VO_{2max}) and reduce robustness. Purpose: this study attempted to address these issues by demonstrating that positive BxB outliers from unfiltered 30-s SI (VO_{2max30}) wouldn't differ in a meaningful (d≤0.15) way from 60-s SI (VO_{2max60}), and 15-s SI (VO_{2max15}) would increase in a significant enough fashion (d≥0.20) to be deemed a non-preferred SI. A novel, dual $VO_{2plateau}$ model was then created using ΔVO_{2max60}-VO_{2max30} and ΔVO_{2max60}-$VO_{2non\text{-}max60}$ to form simple and objective sex-specific criterion guidelines for validation of VO_{2max60} assessment.

Methods: Unfiltered BxB averages from the last 2-min of a graded exercise test to exhaustion were collected from female NCAA Division I cross-country runners (n=14).

Results: VO_{2max60} and VO_{2max30} differed statistically but trivially from each other (2.91±0.28 vs 2.94±0.29 L/min; d=0.11). The differences between VO_{2max15} (3.01±0.33 L/min) and VO_{2max30} (d=0.23) and VO_{2max60} (d=0.33) were meaningful, confirming our hypotheses. Furthermore, four VO_{2max15} occurred before the final 2 min. VO_{2max60}-VO_{2max30} (±0.05 L/min; <1%) and VO_{2max60}-$VO_{2non\text{-}max60}$ (±0.08 L/min; 1.60%) exhibited tight Bland-Altman 95% levels of agreement and coefficient of variation. VO_{2max60} validity can be confirmed using criteria of a dual $VO_{2plateau}$ model of Δ ≤0.08 L/min for ΔVO_{2max60}-VO_{2max30} and Δ ≤0.15 L/min for ΔVO_{2max60}-$VO_{2non\text{-}max60}$.

Conclusions: These simple guidelines can replace more subjective secondary confirmation markers to increase confidence in VO_{2max} outcomes and determine need for verification testing without having to filter BxB data.

Key Words: graded exercise test, metabolic data processing, aerobic capacity

Corresponding author: Eric K. O'Neal ekoneal14@outlook.com

Introduction

For over a century,[1] attempts have been made to classify and standardize procedures to define a gold standard in describing highest attainable oxygen consumption capacity (VO_{2max}) while running. In the 1950s, two cornerstone projects[2,3] offered widely adopted suggestions for conducting treadmill-based graded exercise tests (GXT). In both

studies, expired gases were collected in 60-s sampling intervals (SI) via Douglas bags at the end of a single-speed treadmill running stage, followed by a recovery period before the next stage (i.e., a discontinuous protocol). Intensity incrementally increased until a higher work rate elicited a decline or minimal increase in oxygen consumption (i.e., $VO_{2plateau}$). In the 1955 study, Taylor, Buskirk, and Henschel[3] defined $VO_{2plateau}$ as a difference ≤ 0.15 L/min or 2.1 ml/kg/min between stages to determine if the final intensity "elicited a maximal oxygen intake". The rationale for this criterion selection is not directly expressed. Three years later, Mitchell, Sproule, and Chapman[2] defined $VO_{2plateau}$ more stringently as the sample's mean difference between stages minus twice the standard deviation (≤ 0.054 L/min) to indicate a participant "had attained his true maximal intake". These classic studies were laborious and time-intensive for both investigators and participants, with stages of work separated by 10 minutes[2] or entire days.[3] Advances in technology[4] have led to few laboratories using discontinuous protocols and the Douglas bag technique in favor of breath-by-breath (BxB) open-circuit spirometry and 1-3-min stage continuous style GXT. There is still no universally accepted criterion for $VO_{2plateau,}$ much less population-specific guidelines. An intriguing debate has been given to determine if $VO_{2plateau}$ is even a real physiological artifact.[5,6] Secondary confirmation markers such as rate of perceived exertion or respiratory exchange ratio are often used as additional validation parameters to $VO_{2plateau}$ to establish that maximal effort was given by the participant, thus validating the GXT. This approach has also faced criticism to its validity[7] and trained runners may exhibit different respiratory exchange ratio characteristics than the general population.[8]

Verification tests conducted at a supramaximal intensity versus the last stage completed in continuous, traditional style GXT were once believed to be the answer for GXT validation.[9] Verification tests typically occur after a recovery period of 5-15 min from the GXT and consist of a brief warm-up phase that leads to the supramaximal intensity (multi-stage test) or with a single supramaximal intensity known as a square wave design.[10] Exhaustion is typically reached in <5 min, and the verification tests outcomes are then compared to the last stage of the initial GXT for a confirmation of $VO_{2plateau}$. Instrumentation and terminology have changed, but the latest attempt of using supramaximal verification tests to validate a GXT is essentially a replication of the designs proposed over 70 years ago[2,3] with the exception that all but the last stage occur continuously.

Regardless of the protocol selected, one major difference in the assessment of aerobic capacity since the advent of the modern metabolic cart is the ability to instantly view VO_2 data in multiple SI. To our knowledge, nobody has taken advantage of this capacity to use multiple aspects of $VO_{2plateau}$ to confirm GXT validity. VO_{2max} SI ranges of 10-30 s[11,12,13,14] are often selected in place of the original 60-s SI. When SI is decreased VO_{2max} increases. However, BxB data have high variability, and the capacity of a few outlier breaths to skew VO_{2max} in shorter SI is of concern. Fewer data points and high variability with short SI lead to reduced robustness, creating uncertainty in the validity and repeatability GXT-based VO_{2max} results. Filtering BxB data could reduce this concern, but is rarely reported as being performed, particularly when VO_{2max} is only used for descriptive purposes. Repeated attempts to establish standardization of $VO_{2plateau}$ confirmation criterion and use of a standardized SI have also been unsuccessful.[15,16,17] Contemporary suggestions to remedy these issues and use shorter SI transformed via linear-log relationship modeling have also been promoted but not adopted in the scientific community.[18]

The first purpose of this study was to compare the agreement of the original 60-s SI used by Taylor, Buskirk, and Henschel[3] and Mitchell, Sproule, and Chapman[2] against two shorter and common reporting SI of 15-s and 30-s over the last 2 min of a GXT without manually removing outlier breath data or using rolling averages. We hypothesized that the magnitude of difference between VO_{2max60} and $VO2_{max30}$ would be marginal ($d \leq 0.15$), but BxB outliers for VO_{2max15} would be significant enough ($d \geq 0.20$) to eliminate VO_{2peak} assessed via 15-s SI ($VO_{2peak15}$) as an ideal SI. If confirmed, the second aim was to develop a novel and simple, dual $VO_{2plateau}$ model (DPM) for GXT-based VO_{2max60} validation specific to trained, female runners using a combination of (a) ΔVO_{2max60}-VO_{2max30} and (b) VO_{2max60} and the penultimate 60-s VO_2 of the final 2-min of the GXT (i.e. ΔVO_{2max60}-$VO_{2non\text{-}max60}$).

Scientific Methods

Participants

GXT data from the current study were taken from previously published work in our laboratories investigating models to predict collegiate cross-country performance and compare physiological profiles of men's and women's NCAA Division I competitors at race pace (publication details have been hidden to allow blinded review procedures) using a sample of convenience. No *a priori* power analysis for this study was conducted as such. A methodological goal of the study was to ensure the participant pool would be representative of non-elite, but trained female runners accustomed

to exercise to volitional exhaustion (i.e., collegiate cross-country races and training) with a sample size typical of running-focused investigations. National Collegiate Athletics Association Division I female collegiate cross-country runners (n = 14) completed all study requirements. Personal bests from the previous season for 5-km competition distance were 19.04 ± 1.01 min. Height (162 ± 9 cm) and weight (54.6 ± 3.8 kg) were obtained using a stadiometer (Invicta Plastics Limited, Leicester, England) and digital scale (BWB-800; Tanita, Inc., Tokyo, Japan), respectively. This study was approved by the local institutional review boards at the participating universities.

Protocol

In the day preceding testing, participants were instructed to refrain from strenuous exercise as well as caffeine and alcohol consumption. Additionally, runners were asked to arrive at the laboratory for testing in a fasted state of at least 3 hours. Participants were permitted to follow individual warm-up protocols before initiation of the GXT. Oxygen uptake was assessed throughout the trial via indirect calorimetry using a metabolic cart (TrueOne 2400, Parvo Medics Inc., Sandy, UT), which has been validated against the Douglas bag method in multiple laboratories.[4,19] Calibration was conducted following the manufacturer's instructions, prior to the initiation of the GXT, which was completed on a slat-belt treadmill (4Front, Woodway, Waukesha, WI).

Early[20,21] and later[22,23] incremental GXT treadmill protocols consisting of 1-3-min stages have consistently displayed high repeatability in VO_2 outcomes, with more recent protocols often opting to increase intensity using treadmill speed versus severe grade in trained running populations.[24,25] With these considerations, the GXT protocol used in the current study was developed in our laboratory years ago specifically for trained collegiate cross-country runners with treadmill speed versus grade as the primary mechanism to increase intensity and produce volitional exhaustion in 8-12 min. Initial speed was determined by subtracting 2.4 km/h from each participant's recent 5-km pace. Grade was maintained at 1% throughout the test, and speed was increased by 0.8 km/h every 2-min until the participant reached volitional exhaustion. Trained runner specific RER values were used as a secondary, non-$VO_{2plateau}$ confirmation factor.[8] All participants met this threshold. Upon completion of the GXT, each participant's metabolic VO_2 data were saved and printed in SI of 15, 30, and 60-s. Data from the final 2-min of the GXT were manually searched individually by two investigators. Outcomes were then compared for confirmation, and the highest values were extracted for data analyses respective to each SI.

Statistical Analysis

Statistical significance was considered at $p \leq 0.05$. SPSS V27 (IBM, Chicago, IL) and Microsoft Excel were used to analyze data and prepare figures, respectively.

Determination of VO_{2max60} as preferred SI

Multiple data analysis approaches were undertaken to determine if the first phase of this study, justification of selection of VO_{2max60} as an optimal SI, could be established. First, repeated measures analysis of variance was conducted to detect potential differences among SI. It should be noted that statistical significance was essentially a guaranteed outcome as selecting the highest VO_2 for each SI meant that lower SI could only produce individual VO_2 equal to or greater than a longer SI. If sphericity was violated during Mauchly's test, Greenhouse-Geisser adjustments were implemented to adjust degrees of freedom. Bonferroni corrected *post-hoc* tests were conducted in the case of main effects for SI.

The next step included calculation of Cohen's d effect sizes for all three SI. Effect sizes were calculated using the mean difference divided by the pooled standard deviations between each SI. The hypothesis that no meaningful differences between VO_{2max60} and VO_{2max30} would be exhibited was based on $d \leq 0.15$ indicating no meaningful difference. VO_{2max15} was deemed meaningfully different versus VO_{2max60} and VO_{2max30} if $d \geq 0.20$. Additional concern for detecting unfiltered BxB outliers influencing VO_{2max} was based on incidents of SI peaks occurring outside of expected times. The data outputs provided for the three SI were assessed for the number of instances in which the highest value occurred (a) outside of the final 2 min of the GXT or (b) outside of the whole minute window in which the highest 60-s SI VO_2 occurred during the final 2 min of the GXT.

Modified Bland-Altman plots (BA) were then created to visually represent the agreement between VO_{2max30} and VO_{2max15} to VO_{2max60} to further confirm VO_{2max60} as an optimal SI. Traditional BA plots[26] depict the difference between two measurements within the same participant against the mean of the two scores. However, when a new or different measurement is being compared to an established criterion measurement, Krouwer[27] suggested the difference in

agreement be expressed as the difference of the new measurement versus the reference measurement. This methodology was implemented in the current study based on VO_{2max60} being operationally defined as the optimal reporting SI (i.e., criterion). Furthermore, within-subject coefficient of variation ((CV%; standard deviation of individual values/mean of individual values) x 100) was calculated and expressed as a percentage using VO_{2max60}, VO_{2max30}, and VO_{2max15} data.

Development of Dual $VO_{2plateau}$ Model

DPM guidelines were qualitatively established based on visual inspection of the BA 95% level of agreement ceiling for individual data points. The creation of the BA plot used for the first step of the model ΔVO_{2max60}-VO_{2max30} has been described above. An additional BA plot depicting the agreement of VO_{2max60} and ΔVO_{2max60}-$VO_{2non\text{-}max60}$ was created to establish guidelines for the second step of the DPM. Paired *t* tests were also performed for the (a) next to last and last 60-s SI and (b) VO_{2max60} and $VO_{2non\text{-}max60}$.

Results

Establishment of VO_{2max60} as Preferred SI

Significant main effects for SI were found for absolute and relative VO_{2max} ($p < 0.01$). Pairwise comparisons of SI for absolute and relative VO_{2max} via Bonferroni corrections indicated that all three SI were found to be significantly different from each other ($p < 0.01$; Table 1). A decrease in SI from VO_{2max60} to VO_{2max30} increased VO_{2max} by ~1% but over 3% versus VO_{2max15}. Confirming our hypotheses, Cohen's *d* effect sizes were 0.11 (VO_{2max30} vs VO_{2max60}), 0.23 (VO_{2max15} vs VO_{2max30}), and 0.33 (VO_{2max15} vs VO_{2max60}).

VO_{2max30} displayed tight agreement with VO_{2max60} in both absolute (Figure 1A; mean difference ± 95% upper and lower levels of agreement = 0.04 ± 0.05 L/min) and relative units (Figure 1B; mean difference ± 95% upper and lower levels of agreement = 0.71 ± 0.86 ml/kg/min). In contrast, the mean difference in agreement of VO_{2max15} versus VO_{2max60} exceeded the 95% level of agreement of VO_{2max30} and VO_{2max60} (Figure 1A & B). Absolute (Figure 1A; mean difference ± 95% upper and lower levels of agreement = 0.10 ± 0.13 L/min) and relative (Figure 1B; mean difference ± 95% upper and lower levels of agreement = 1.89 ± 2.17 ml/kg/min) VO_{2max} agreement increased by ≥250% versus VO_{2max60} and VO_{2max30}. No runner's absolute VO_2 difference in VO_{2max30} and VO_{2max60} met or exceeded 0.10 L/min, but nearly half of participants' data exceeded this limit with VO_{2max15}. In terms of relative VO_{2max}, half of the runners' VO_{2max15} was 1.4 ml/kg/min or higher versus VO_{2max60}, whereas no participants exceeded this threshold for VO_{2max30}. Three participants' VO_{2max15} exceeded VO_{2max60} by > 3.0 ml/kg/min (5.6% of the average VO_{2max60}).

Coefficient of variation fell below 1% for VO_{2max30} and VO_{2max60} compared to values approaching 2.5% for VO_{2max15} (Table 1). Instances where VO_{2max} values occurred before the final 2 min of the GXT or outside the min in which VO_{2max60} occurred can be found in Table 2. Approximately 30% of VO_{2max15} occurred before the final 2 minutes of the GXT and 50% occurred outside the VO_{2max60} during the final 2 min.

VO_{2max60} and $VO_{2non\text{-}max60}$

VO_{2max60} (2.91 ± 0.28 L/min) exceeded ($p < 0.01$) $VO_{2non\text{-}max60}$ (2.84 ± 0.29 L/min) in the final 2 min of the GXT as expected but displayed promising contextual agreement for verification of $VO_{2plateau}$ with a CV ≤ 1.6% (Table 1). Nine of the highest $\dot{V}O_2$ scores occurred in the final min of the GXT (Figure 2A), with all participants differing ≤ 0.15 L/min from the next to last min. Penultimate (2.87 ± 0.29 L/min) and last (2.88 ± 0.29 L/min) min absolute VO_2 did not differ ($p = 0.50$). VO_{2max60} and $VO_{2non\text{-}max60}$ exhibited strong agreement with only 2 runners differing by > 0.10 L/min (Figure 2B).

Dual plateau model guidelines

Visual inspection of Figures 1A and 2B was used to develop DPM guidelines. These suggestions were made with three assumptions. (1) All participants ran to true volitional exhaustion. (2) Sample agreement ΔVO_{2max60}-VO_{2max30} and ΔVO_{2max60}-$VO_{2non\text{-}max60}$ represented population-based characteristics of trained female runners running to true volitional exhaustion. (3) Outcomes outside developed guidelines represent potential that a true VO_{2max} was not reached because true volitional exhaustion was not achieved. Results concerning other non-$VO_{2plateau}$ based VO_{2max} validity criteria (e.g., respiratory exchange ratio) have been published elsewhere[8] in detail for this sample. For trained female runners, it was decided that $\Delta \leq 0.08$ L/min and $\Delta \leq 0.15$ ml/kg/min would be suitable levels of agreement to

confirm ΔVO_{2max60}-VO_{2max30} $VO_{2plateau}$. A $\Delta \leq 0.15$ L/min for ΔVO_{2max60}-$VO_{2non\text{-}max60}$ was established as the second $VO_{2plateau}$ criterion to confirm GXT validity.

Table 1. VO_{2max} data from the final 2 min of a graded exercise test by sampling interval duration (n = 14) and within-subject coefficient of variation (CV) versus VO_{2max60} or *$VO_{2nonmax60}$.

	VO_{2max15}		**VO_{2max30}**		**VO_{2max60}**	
	(mean ± SD)	CV	(mean ± SD)	CV	(mean ± SD)	CV*
VO_{2max} (L/min)	3.01 ± 0.33[a,b]	2.42%	2.94 ± 0.29[b,c]	0.88%	2.91 ± 0.28[a,c]	1.60%
VO_{2max} (ml/kg/min)	55.1 ± 3.1[a,b]	2.45%	53.9 ± 2.8[b,c]	0.93%	53.2 ± 2.7[a,c]	1.56%

[a] = $p < 0.01$ versus VO_{2max30}; [b] = $p < 0.01$ versus VO_{2max60}; [c] = $p < 0.01$ versus VO_{2max15}.

Table 2. Instances in which VO_{2max} and peak respiratory factors occurred before the final 2 min of GXT or outside of the VO_{2max60} SI during the final 2 min of GXT (n = 14).

	Before final 2 min	**Outside VO_{2max60} during final 2 min**		
	VO_{2max}	**VO_{2max}**	**RER**	**RR (breath/min)**
VO_{2max60}	1 (7.1)	--	6 (42.9)	3 (21.4)
VO_{2max30}	--	4 (28.6)	9 (64.3)	6 (42.9)
VO_{2max15}	4 (28.6)	7 (50.0)	10 (71.4)	9 (64.3)

Values are presented as number of cases (% of cases).
RER = respiratory exchange ratio; RR = respiratory rate; GXT = graded exercise test; SI = sampling interval.

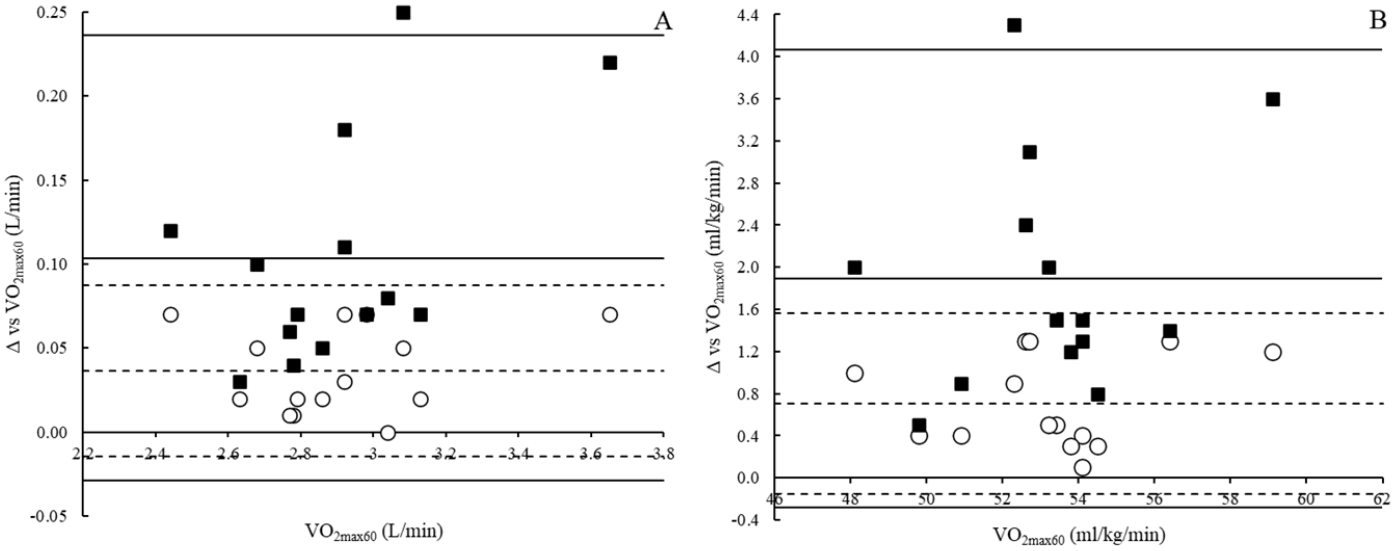

Figure 1. Agreement for (A) absolute and (B) relative VO_{2max} between VO_{2max30} (circle markers) and VO_{2max15} (square markers) versus VO_{2max60} (n = 14). Dashed lines represent mean and 95% upper and lower levels of agreement between VO_{2max30} and VO_{2max60}. Solid lines represent mean and 95% upper and lower levels of agreement between VO_{2max15} and VO_{2max60}.

Discussion

The methodology and instrumentation for the formative investigations[2,3] establishing running VO_{2max} validity based on $VO_{2plateau}$ criterion were ideally fashioned for their time. A GXT with discontinuous stages and 60-s SI produced robust data with a high likelihood to find operationally defined $VO_{2plateau}$ between stages. It is important to recognize that the 60-s SI was not used to reduce BxB measurement noise that is inherent in modern automated systems, but the impracticality of short SI use with the Douglas bag technique. Decades later and with advanced instrumentation, there is still no formal consensus on the most appropriate $VO_{2plateau}$ criteria or SI to use when assessing aerobic capacity. This exploratory study re-examined these issues using a novel, dual $VO_{2plateau}$ model based on multiple SI and developed sex-specific guidelines for trained, female runners. This study had three primary findings. (1) VO_{2max60} and

VO_{2max30} displayed strong enough levels of agreement with each other that using the longer SI did not meaningfully diminish the higher VO_{2max} produced with the shorter 30-s SI. (2) The high variability in unfiltered BxB data for VO_{2max15} eliminated its consideration as an optimal SI. (3) With the assumption that all participants completed the GXT to true volitional exhaustion, agreement between model parameters supports the DPM and population-specific guidelines could be used to confirm GXT validity in trained female runners or if a verification test is needed.

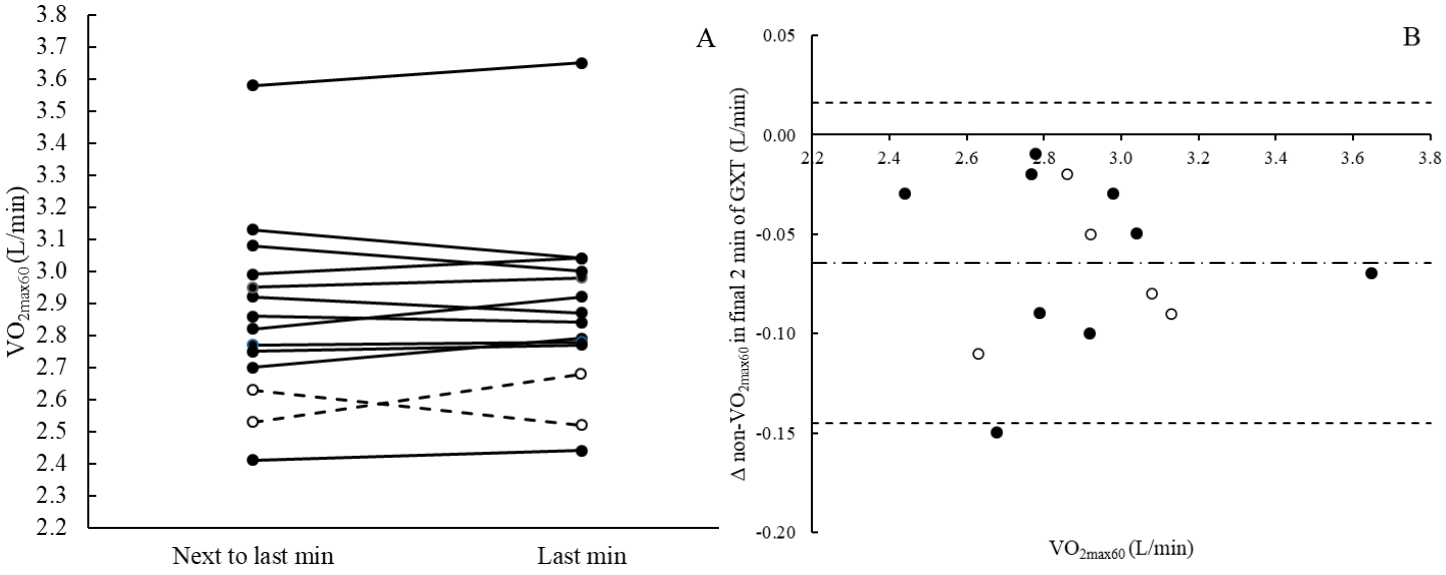

Figure 2. Individual participant assessment of final two, 60 SI (n = 14). A) Individual differences in absolute VO_2 for first and second 60 SI. Dashed lines and open markers indicate an inter-min difference of VO_2 > 0.10 L/min. B) Bland-Altman plot of difference in absolute $VO_{2non\text{-}max60}$ and VO_{2max60}; mean difference = -0.06 L/min; 95% levels of agreement = ± 0.08 L/min. Open markers indicate highest VO_2 value occurring in the first of the final 2 min.

The primary SI selection issue is a debate of precision versus robustness. Robergs, Dwyer, and Astorino[17] reported up to 70% of the variance in VO_2 between single breaths can be attributed to variation in expiratory volume. Single breaths during the last minute of a GXT can differ by as much as ~18 and 24 ml/kg/min for trained female and male endurance athletes, respectively.[28] In the current study, nearly two-thirds of the highest 15-s SI respiratory rates occurred outside the 60-s period in which VO_{2max60} was expressed (Table 2). It is likely BxB expiratory volume differences, not a true max followed by a decline in VO_2, explain why ~30% of runners experienced a VO_{2max15} prior to the final 2 min of the GXT. In contrast, VO_{2max60} and VO_{2max30} shared tighter agreement than even $VO_{2peak60}$ and $VO2_{non\text{-}peak60}$ (Figures 1 & 2B). Our hypothesis that for unfiltered data, VO_{2max60} is the optimal SI is intuitively justifiable by the group outcomes in Table 1, but for GXT validity assessment of individual runners, agreement is the most critical quantitative perspective. Tables 1 and 2 and Figure 1 offer supplementary, if not stronger support for VO_{2max60}. The "V" in VO_{2max} indicates volume. The dot above the "V" can be traced back to the notation style of Sir Isaac Newton. In this case, it indicates that volume is represented as a unit of time, which is unambiguously interpreted as per min. While aerobic capacity is universally reported as a unit per min, it is more often than not extrapolated from a much shorter SI. This detail is important from both research and coaching interpretation aspects. Many studies assign intensity, assess running economy, or examine fractional utilization using 60-s SI for sub-maximal variables, but may base these outcomes or intensity on the highest attained VO_2 during GXTs that utilize shorter SI durations.[7,29,30,31] If the reported aerobic capacity outcome is not reliable, it cannot be considered valid. VO_{2max60} is both. VO_{2max60} can be robustly used to assess outcomes in running research or prescribe training and assess VO_{2max} changes for individual runners by coaches. The higher VO_{2max15} values are either evidence that (a) *true* VO_{2max} is only attainable and sustainable in SI < 30-s, (b) represent noise of single BxB ventilatory outliers, or (c) a combination of the two during a GXT. Regardless, it is counterintuitive to prioritize these shorter and more transitory values versus the highest VO_2 that can be maintained for not only 60-s but also demonstrate high agreement across the final two 60-s SI of a GXT (Figure 2).

The second and primary aim of this study was to establish a dual $VO_{2plateau}$ model criterion specifically for confirming GXT validity in trained female runners using unfiltered data from the last 2-min of testing. It is beyond the scope of the current paper to exhaustively review the concept of $VO_{2plateau}$ for confirming VO_{2max}, but it is imperative to briefly address aspects of this topic before offering novel DPM criterion guidelines for establishing VO_{2max60}. Both Midgley, McNaughton, and Carroll[32] and Poole and Jones[9] called for dismissal of an alternate term (VO_{2peak}) to describe aerobic capacity. VO_{2peak} is often incorporated without $VO_{2plateau}$ confirmation. The authors suggested subsequent $VO_{2plateau}$ confirmation from the initial GXT and supramaximal verification tests could be used to authenticate VO_{2max}. Meta-analytics confirm verification tests broadly produce equivocal outcomes,[33] and trained runners have often demonstrated higher VO_{2max} outcomes during GXT versus verification protocols.[34,35,36]

In trained populations, exercise to volitional exhaustion non-compliance is uncommon. Verification tests have not become standard practice. Participants should not have to endure a second exhaustive bout of treadmill running while also extending time voluntarily given to laboratory session unless there is objective evidence for its need. Our first suggestion and evidence of indication that VO_{2max} was achieved involves a simple examination of $VO_{2plateau}$ characteristics of ΔVO_{2max60}-VO_{2max30}. Returning to Table 1, current participants experienced only a marginal decrease in VO_{2max60} versus VO_{2max30} of 0.03 L/min (0.70 ml/kg/min). If runners' VO_2 was continuing to climb in coincidence with a premature volitional exhaustion incident (i.e., an invalid GXT), these shorter 30-s SI should have exhibited evidence of ascending VO_2. This was not the case.

Again, returning to the importance of agreement, comparison of mean data can obscure within-subject agreement for VO_{2max} assessment interpretation. Day, Rossier, Coats, Skasick, and Whipp[36] deemphasized $VO_{2plateau}$ attainment importance during a ramp test versus a single intensity verification test because there was no statistical difference between VO_{2max} of testing modalities. However, when Day and colleagues[36] depicted individual data of the two test types in a scatterplot, a contingent of participants showed significant disagreement between test types. Error occurred in both directions (i.e. some participants scored higher on a ramp style versus supramaximal test and vis versa), essentially nullifying any statistical difference. Similarly, Dideriksen and Mikkelsen[37] reported 15-s SI VO_{2max} did not differ across the 3 trials for 13 recreationally competitive triathletes. When the data were presented in BA plots, 11 of the 39 comparisons made between 3 trials approached or exceeded differences of 0.3 L/min.[37] These differences would be highly impactful whether the data were used for research or training prescription purposes and illustrate the concern of short SI. In contrast to Day et al.,[36] Neimeyer, Bergmann, and Beneke[5] reported near-unanimous $VO_{2plateau}$ confirmation during a cycling ramp test. However, the slope calculation-based plateau technique used by Niemeyer, Bergmann, and Beneke[5] was only deemed acceptable if a 100-s SI (change of 50 W during ramp test) was used. This is essentially the same approach taken in Figure 2B (ΔVO_{2max60}-$VO_{2nonmax60}$) minus 10 s per SI. Furthermore, this process was only used to determine if $VO_{2plateau}$ occurred. Niemeyer, Bergmann, and Beneke[5] chose to use the highest 30-s SI for reporting VO_{2max} despite the requirement of two, 50-s SI comparisons to confirm $VO_{2plateau}$. It is unclear if the reported values occurred during the windows $VO_{2plateau}$ was confirmed. Our position is not that the findings of Niemeyer et al.[5] concerning $VO_{2plateau}$ artifacts during a cycling ramp test are without merit, but that evidence from the current study refutes that such great interpretation lengths are needed in trained runners. Figures 2 A & B provide secondary quantitative support evidence that a $VO_{2plateau}$ was reached by participants. Five participants exhibited VO_{2max60} in the next to last 60-s SI, suggesting validation of the highest standard of VO_{2max} confirmation, an apex then fall in VO_2. While 9 runners' VO_{2max60} occurred in the last min of the GXT, 5 of these values differed by ≤ 0.05 L/min from the previous min. The excellent agreement (Figure 2) between the last two, 60 SI of the GXT does not suggest a meaningful increase in VO_2 was likely for any individual runner and can be calculated quickly using simple arithmetic.

Conclusions

In the past, aerobic capacity testing via indirect calorimetry was highly limited. This is no longer the case. Running coaches and competitors may desire to use VO_{2max} data for monitoring progress or modifying training activities and programming. BxB data is rarely reported as being filtered in research scenarios. It is difficult to imagine coaches would take this additional step. Secondary confirmation markers such as rate of perceived exertion or percentage of maximal heart rate are limited in their capacity to confirm GXT validity. No general population $VO_{2plateau}$ criteria have been established, much less criteria specific to runners based on ability or sex. Trained female runner data is highly underrepresented in this area, and one reason we elected to develop this model in females first. The simple, dual $VO_{2plateau}$ model developed in this study can give confidence to researchers or coaches in the robustness of assessed aerobic capacity from a GXT and quickly allow for judgement if a verification test is needed on an individual basis. For trained female runners with VO_{2max60} between ~50-60 ml/kg/min, agreement of ≤ 0.08 L/min or 1.5 ml/kg/min

between VO_{2max30} and VO_{2max60} and ≤ 0.15 L/min difference for ΔVO_{2max60}-$VO_{2non\text{-}max60}$ can be used to validate GXT-based VO_{2max} outcomes.

Acknowledgements
The authors thank all of the runners who participated in the current study and the coaches who assisted with laboratory and athletic collaboration. The authors received no external funding for this project and have no conflicts of interest in regard to the data presented.

References

1. Hill A, Lupton H. Muscular exercise, lactic acid, and the supply and utilization of oxygen. *QJM (Quarterly J Med)*. 1923;16:135-171. doi:10.1093/qjmed/os-16.62.135 OUP Academic+2journals.humankinetics.com+2
2. Mitchell JH, Sproule BJ, Chapman CB. The physiological meaning of the maximal oxygen intake test. *J Clin Invest*. 1958;37:538-547. doi:10.1172/JCI103636
3. Taylor HL, Buskirk E, Henschel A. Maximal oxygen intake as an objective measure of cardio-respiratory performance. *J Appl Physiol*. 1955;8:73-80. doi:10.1152/jappl.1955.8.1.73
4. Bassett DR Jr, Howley ET, Thompson DL, King GA, Strath SJ, McLaughlin JE, Parr BB. Validity of inspiratory and expiratory methods of measuring gas exchange with a computerized system. *J Appl Physiol*. 2001;91:218-224. doi:10.1152/jappl.2001.91.1.218
5. Beltz MB, Gibson AL, Janot JM, Kravitz L, Mermier CM, Dalleck LC. Graded exercise testing protocols for the determination of VO_2max: historical perspectives, progress, and future considerations. *J Sports Med (Hindawi Publ Corp)*. 2016;2016:3968393.
6. Howley ET, Bassett DR Jr, Welch HG. Criteria for maximal oxygen uptake: review and commentary. *Med Sci Sports Exerc*. 1995;27:1292-1301.
7. Carder MJ, Scudamore EM, Savanna KN, Pribyslavska V, Bowling LR, O'Neal EK. Retrospective and contemporary predictors of National Collegiate Athletic Association Division I cross-country performance are sex specific. *J Strength Cond Res*. 2023;37:2267-2272.
8. Bowling LR, Knight SN, Scudamore EM, Waldman HS, Scudamore V, O'Neal EK. Trained Runners Need Lower Respiratory Exchange Ratio Criterion During Graded Exercise Tests. *Res in Strength and Perf*. 2025; 5(1).
9. Poole DC, Wilkerson DP, Jones AM. Validity of criteria for establishing maximal O_2 uptake during ramp exercise tests. *Eur J Appl Physiol*. 2008;102:403-410.
10. Poole DC, Jones AM. Measurement of the maximum oxygen uptake VO_2max: VO_2peak is no longer acceptable. *J Appl Physiol*. 2017;122:997-1002.
11. Schaun GZ. The maximal oxygen uptake verification phase: a light at the end of the tunnel? *Sports Med Open*. 2017;3:44.
12. Dexheimer JD, Brinson SJ, Pettitt RW, Schroeder ET, Sawyer BJ, Jo E. Predicting maximal oxygen uptake using the 3-minute all-out test in high-intensity functional training athletes. *Sports (Basel)*. 2020;8.
13. Farrell JW 3rd, Dunn A, Cantrell GS, Lantis DJ, Larson DJ, Larson RD. Effects of group running on the training intensity distribution of collegiate cross-country runners. *J Strength Cond Res*. 2021;35:2862-2869.
14. Hebert-Losier K, Finlayson SJ, Driller MW, Dubois B, Esculier JF, Beaven CM. Metabolic and performance responses of male runners wearing three types of footwear: Nike Vaporfly 4%, Saucony Endorphin racing flats, and their own shoes. *J Sport Health Sci*. 2022;11:275-284.
15. Pace MT, Green JM, Killen LG, Swain JC, Chander H, Simpson JD, O'Neal EK. Minimalist-style boot improves running but not walking economy in trained men. *Ergonomics*. 2020;63:1329-1335.
16. Niemeyer M, Bergmann TGJ, Beneke R. Oxygen uptake plateau: calculation artifact or physiological reality? *Eur J Appl Physiol*. 2020;120:231-242.
17. Taylor K, Seegmiller J, Vella CA. The decremental protocol as an alternative protocol to measure maximal oxygen consumption in athletes. *Int J Sports Physiol Perform*. 2016;11:1094-1099.
18. Martin-Rincon M, González-Henríquez JJ, Losa-Reyna J, Pérez-Suárez I, Ponce-González JG, de La Calle-Herrero J, Pérez-Valera M, Pérez-López A, Curtelin D, Cherouveim ED, Morales-Alamo D, Calbet JAL. Impact of data averaging strategies on VO_2max assessment: mathematical modeling and reliability. *Scand J Med Sci Sports*. 2019;29:1473-1488.
19. Crouter SE, Antczak A, Hudak JR, DellaValle DM, Haas JD. Accuracy and reliability of the ParvoMedics TrueOne 2400 and MedGraphics VO2000 metabolic systems. *Eur J Appl Physiol*. 2006;98:139-151.
20. Balke B, Ware RW. An experimental study of physical fitness of Air Force personnel. *US Armed Forces Med J*. 1959;10:675-688.

21. Bruce RA, Blackmon JR, Jones JW, Strait G. Exercise testing in adult normal subjects and cardiac patients. *Pediatrics*. 1963;32(suppl):742-756.
22. Myers J, Buchanan N, Walsh D, Kraemer M, McAuley P, Hamilton-Wessler M, Froelicher VF. Comparison of the ramp versus standard exercise protocols. *J Am Coll Cardiol*. 1991;17:1334-1342.
23. St Clair Gibson A, Lambert MI, Hawley JA, Broomhead SA, Noakes TD. Measurement of maximal oxygen uptake from two different laboratory protocols in runners and squash players. *Med Sci Sports Exerc*. 1999;31:1226-1229.
24. Gaddie JW, Kennedy EP, Green M, Killen LG, Linder BA, Heinkel AA, O'Neal EK. Effects of three modest levels of proximal loading on marathon-pace running economy. *Int J Exerc Sci*. 2020;13:1120-1131.
25. Joubert DP, Guerra NA, Jones EJ, Knowles EG, Piper AD. Ground contact time imbalances strongly related to impaired running economy. *Int J Exerc Sci*. 2020;13:427-?
26. Altman DG, Bland JM. Measurement in medicine: the analysis of method comparison studies. *J R Stat Soc Ser D (The Statistician)*. 1983;32:307-317.
27. Krouwer JS. Why Bland–Altman plots should use X, not (Y+X)/2 when X is a reference method. *Stat Med*. 2008;27:778-780.
28. Astorino TA. Alterations in VO_2max and the VO_2 plateau with manipulation of sampling interval. *Clin Physiol Funct Imaging*. 2009;29:60-67.
29. Conley DL, Krahenbuhl GS. Running economy and distance-running performance of highly trained athletes. *Med Sci Sports Exerc*. 1980;12:357-360.
30. Turner AM, Owings M, Schwane JA. Improvement in running economy after 6 weeks of plyometric training. *J Strength Cond Res*. 2003;17:60-67.
31. Weston AR, Mbambo Z, Myburgh KH. Running economy of African and Caucasian distance runners. *Med Sci Sports Exerc*. 2000;32:1130-1134.
32. Midgley AW, McNaughton LR, Carroll S. Effect of the VO_2 time-averaging interval on the reproducibility of VO_2max in healthy athletic subjects. *Clin Physiol Funct Imaging*. 2007;27:122-125.
33. Costa VAB, Midgley AW, Baumgart JK, Carroll S, Astorino TA, Schaun GZ, Fonseca GF, Cunha FA. Confirming the attainment of maximal oxygen uptake within special and clinical groups: a systematic review and meta-analysis of cardiopulmonary exercise test and verification phase protocols. *PLoS One*. 2024;19:e0299563.
34. Sánchez-Otero T, Iglesias-Soler E, Boullosa DA, Tuimil JL. Verification criteria for the determination of VO_2max in the field. *J Strength Cond Res*. 2014;28:3544-3551.
35. Succi PJ, Benitez B, Kwak M, Bergstrom HC. Methodological considerations for the determination of VO_2max in healthy men. *Eur J Appl Physiol*. 2023;123:191-199.
36. Day J, Rossiter H, Coats E, Skasick A, Whipp B. The maximally attainable VO_2 during exercise in humans: the peak vs. maximum issue. *J Appl Physiol*. 2003;95:1901-1907.
37. Dideriksen K, Mikkelsen UR. Reproducibility of incremental maximal cycle-ergometer tests in healthy recreationally active subjects. *Clin Physiol Funct Imaging*. 2017;37:173-182.