



## Popularity and prevalence of gas exchange data processing methods: a scoping review

Journal:	<i>International Journal of Sports Medicine</i>
Manuscript ID	IJSM-07-2024-10699-re
Manuscript Type:	Review
Key word:	Averaging, Outliers, Interpolation, Cardiopulmonary exercise testing, Breath-by-breath, Reproducibility
Abstract:	<p>Cardiopulmonary exercise testing involves collecting variable breath-by-breath data, sometimes requiring data processing of outlier removal, interpolation, and averaging before later analysis. These data processing choices, such as averaging duration, are known to affect calculated values such as <math>\text{VO}_2\text{max}</math>. However, assessing the effects of data processing without knowing popular methods worth comparing is difficult. In addition, such details aid study reproduction. We conducted a scoping review of articles with exercise testing that collected data breath-by-breath from three databases. Of the 8,351 articles, 376 (<math>4.5 \pm 0.4\%</math>) and 581 (<math>7.0 \pm 0.5\%</math>) described outlier removal and interpolation, respectively. An estimated <math>66.8 \pm 2.8\%</math> reported averaging methods (<math>n = 1078</math>). Commonly documented outlier cutoffs were <math>\pm 3</math> or <math>4</math> SD (<math>39.1\%</math> and <math>51.6\%</math>, respectively). The dominating interpolation duration and procedure were one second (<math>93.9\%</math>) and linear interpolation (<math>92.5\%</math>). Averaging methods commonly described were 30 (<math>30.9\%</math>), 60 (<math>12.4\%</math>), 15 (<math>11.6\%</math>), 10 (<math>11.0\%</math>), and 20 (<math>8.1\%</math>) second bin averages. This shows that studies collecting breath-by-breath data often lack detailed descriptions of data processing methods, particularly for outlier removal and interpolation. While averaging methods are more commonly reported, improved documentation across all processing methods will enhance reproducibility and facilitate future research comparing data processing choices.</p>

SCHOLARONE™  
Manuscripts

# Popularity And Prevalence Of Gas Exchange Data Processing Methods: A Scoping Review

## Abstract

Cardiopulmonary exercise testing involves collecting variable breath-by-breath data, sometimes requiring data processing of outlier removal, interpolation, and averaging before later analysis. These data processing choices, such as averaging duration, are known to affect calculated values such as  $\text{VO}_2\text{max}$ . However, assessing the effects of data processing without knowing popular methods worth comparing is difficult. In addition, such details aid study reproduction. We conducted a scoping review of articles with exercise testing that collected data breath-by-breath from three databases. Of the 8,351 articles, 376 ( $4.5 \pm 0.4\%$ ) and 581 ( $7.0 \pm 0.5\%$ ) described outlier removal and interpolation, respectively. An estimated  $66.8 \pm 2.8\%$  reported averaging methods ( $n = 1078$ ). Commonly documented outlier cutoffs were  $\pm 3$  or  $4$  SD ( $39.1\%$  and  $51.6\%$ , respectively). The dominating interpolation duration and procedure were one second ( $93.9\%$ ) and linear interpolation ( $92.5\%$ ). Averaging methods commonly described were  $30$  ( $30.9\%$ ),  $60$  ( $12.4\%$ ),  $15$  ( $11.6\%$ ),  $10$  ( $11.0\%$ ), and  $20$  ( $8.1\%$ ) second bin averages. This shows that studies collecting breath-by-breath data often lack detailed descriptions of data processing methods, particularly for outlier removal and interpolation. While averaging methods are more commonly reported, improved documentation across all processing methods will enhance reproducibility and facilitate future research comparing data processing choices.

## 1. Introduction

Clinicians and researchers commonly use cardiopulmonary exercise testing (CPET) to determine maximal aerobic capacity ( $\text{VO}_2\text{max}$ ), ventilatory thresholds, and  $\text{VO}_2$  kinetics. Such values help categorize fitness, predict disease risk, and guide exercise [1, p. 162]. Using CPET results to guide exercise, especially relative to thresholds, produces better improvements due to more consistent and predictable metabolic responses [2]. Therefore, incorrectly calculating or identifying these values limits CPET benefits.

Calculating the above values often requires data processing when CPET data is collected breath-by-breath (BBB) as it is highly variable [3]. CPET data processing usually involves outlier removal, optional interpolation to regular intervals, and averaging to more accurately reflect whole-body metabolism [3]. Previous research has shown that data averaging influences CPET values. Averaging over longer durations reduces  $\text{VO}_2\text{max}$  and  $\text{VO}_2$  plateau detection [3–17]. We are unaware of research on the effects of data processing and locating ventilatory thresholds.

Many studies remove outliers by finding points  $\pm 3$  or  $\pm 4$  standard deviations (SD) beyond the local mean (i.e., a prediction interval). These cutoffs are common because the relatively small sample size of BBB gas exchange data often contains more values beyond 3 or 4 SD than one would predict from an assumed Gaussian distribution [18]. More outliers appear than expected because of both conscious and unconscious alterations of breathing patterns, including swallowing and coughing [18]. We are

unaware of prior research that examines how different outlier removal strategies affect  $\text{VO}_2\text{max}$ , ventilatory thresholds, and  $\text{VO}_2$  kinetics.

Interpolation, often to one-second intervals, is common in  $\text{VO}_2$  kinetics research to “ensemble” average repeated transitions to minimize variability [18,19]. Although this does not affect parameter estimates, one-second interpolation has been criticized for artificially narrowing confidence intervals [20–23]. As before, we are unaware of research specifically investigating how interpolation affects  $\text{VO}_2\text{max}$  and ventilatory threshold identification.

Data processing choices, such as averaging and interpolation, impact CPET variables or their confidence intervals. Existing surveys [3] and studies [8] are small and focused on averaging methods only, finding time-based bin averages (e.g., 30-second averages) were popular. A larger sample can, therefore, better describe how often all data processing steps are described.

Before conducting this scoping review, we anecdotally observed that many articles using CPET data did not report all data processing steps, especially outlier removal and interpolation details. This may hamper reproduction or replication attempts, which have become a more prominent issue in science within the past decade [24,25]. Therefore, to assist with conducting future research on the effects of data processing on CPET values and to evaluate the methodological reproducibility of research using BBB gas exchange data generally, we conducted a broad scoping review to identify the frequency of reporting, and popularity of outlier removal, interpolation, and data averaging methods.

## 2. Methods

### 2.1 Design and Eligibility Criteria

This scoping review surveyed gas exchange data processing choices in original, peer-reviewed studies, summarizing the reporting frequency and methods for outlier removal, interpolation, and averaging. It is based on a dissertation chapter by the first author [26]. These methods [27] and results [28] are modeled on the PRISMA scoping review extension guidelines. Eligible articles were original, peer-reviewed articles, with BBB gas exchange data, human participants, in English, and with a DOI. We imposed no date restriction.

### 2.2 Information Sources and Search

We acquired data from the Ovid-MEDLINE, Scopus, and Web of Science databases with the guidance of a university librarian. The electronic search strategy for the Ovid-MEDLINE database can be found in the supplemental materials.

Our search output comprised article identifiers like DOIs. To find missing DOIs, we employed the PubMed Central ID Converter API [29] using Python. Full texts were accessed via publisher text and data mining APIs using Python, unpaywall.org [30] using the unpywall Python package, through custom-built web-scraping scripts, or manually. Our library subscription did not permit access to 1,549 articles.

## 2.3 Selection of Sources of Evidence

This study used a single screening process because it differs from most scoping reviews. It only requires an exercise test with BBB gas exchange data collection rather than a more complex assessment of the overall methodology and intervention.

### 2.3.1 Text Analysis and Screening

Despite database search filters, we screened additional non-English, non-human, and non-original articles such as reviews, meta-analyses, and protocol registrations, in addition to case studies. We manually analyzed a subset of articles to help build machine learning (ML) classifiers and construct RegExs described below. These ML classifiers and RegExs helped identify ineligible articles. This computerized screening required converting full-text PDF and EPUB documents into plain text files. Plain text files were normalized by transforming text to lowercase, removing hyphenations and extra whitespace, and correcting some plain text conversion-induced errors.

Following the normalization, we identified and removed articles that failed to correctly convert into text format, spotted non-English articles using the fasttext Python module [31] and employed a random forest classifier from the sklearn Python package [32] to detect ineligible articles based on our criteria. We manually reviewed potentially ineligible articles flagged by the ML classifier.

Next, we identified BBB articles using RegExs. Articles were considered BBB articles if their text contained variations of the phrase “breath-by-breath”, or if their text included the make or model of a known BBB analyzer. Breath-by-breath brands and analyzers we included were Oxycon and Carefusion brands, Medgraphics Ultima, CPX, CCM, and Cardio<sub>2</sub> models, Sensormedics Encore and 2900 models, Cosmed quark, k4, and k5 models, and the Minato RM-200, AE-280S, AE-300S, and AE-310S models. In total, we identified 8,417 articles.

Within this subset, we performed a similar RegEx search for studies that documented using Douglas Bags or mixing chambers and excluded those articles. The full details are described in the “data charting process” section.

### 2.3.2 Data Charting Process

RegExs identified the presence of short phrases likely indicating that the authors described these methodological details. If present, we extracted a “snippet” of text surrounding those phrases for later manual analysis by obtaining approximately 200 surrounding characters. We then recorded the methods from these snippets. In all cases, methods were only considered documented if the snippets provided at least some specific information. For example, articles stating outlying breaths were removed but without describing the outlier criteria were considered “not described.” Finally, we read the full-text article to accurately document the data when snippets were ambiguous.

The data charting subsections below provide text extraction examples. Extracted texts were normalized to lowercase, with end-of-line hyphenation and unnecessary white

space removed before capitalizing certain keywords for readability. Therefore, formatting varies and may include unconventional spacing and Unicode characters.

We analyzed all eligible BBB articles for outlier and interpolation methods because fewer articles described these methods (~5%) and the phrases were more distinct. In contrast, we analyzed a random subset of articles to document data averaging methods because far more articles described their averaging methods. Early estimates as we developed our RegExs were that ~60% or 5,050 articles had some averaging details. Furthermore, the phrases associated with averaging methods are more generic and often refer to other study aspects, such as heart rate averaging periods. Given the large number of articles, we needed a minimum sample size of 1,068 based on a 95% confidence interval and a maximum margin of error of  $\pm 3\%$ , assuming a proportion of 0.5. However, we raised this to 1,100 in anticipation of finding ineligible articles that eluded our previous text screening.

### 2.3.2.1 Outliers

Our outlier RegExs identified phrases like “swallowing”, “coughing”, “errant”, “aberrant”, and references to the “local mean,” “prediction interval,” or a specific standard deviation limit such as  $\pm 3$  or  $\pm 4$ . For example, our RegExs found “errant”; “local mean”; and “breath-by-breath vo2 data from each step transition were initially edited to exclude errant breaths by removing values lying more than 4 sd” from [33]. We gathered snippets surrounding those phrases and combined them when overlapping, thus producing

y[hb+mb] data (quaresima & ferrari, 2009). expressed as 2.5 data analysis and kinetic modelling the breath-by-breath vo2 data from each step transition were initially edited to exclude errant breaths by removing values lying more than 4 sd from the local mean determined using a five-breath rolling  
breese et al. and deoxy[hb+mb] responses were subavera

We recorded the outlier limit as  $\pm 4$  SD and the outlier function as a rolling 5-breath whole mean average.

### 2.3.2.2 Interpolation

Nearly all articles describing interpolation methods used variations of “interpolate.” The remaining phrases were infrequent and inconsistent enough that interpolation methods were only described for those articles when discovered by chance. To illustrate interpolation documentation, our RegExs extracted the snippet from [34].

the  $\dot{v}O_2$  data from gd and gl exercise bouts were modeled to characterize the oxygen uptake kinetics following the methods described by bell et al. (2001). breath-by-breath  $\dot{v}O_2$  data were linearly INTERPOLATED to provide second-by-second values. phase 1 data (i.e. the cardiodynamic component), from the first ~20 s of exercise, were omitted from the kinetics analysis because phase 1 is not directly repres

We documented the interpolation type as “linear” and the interpolation time as one second.



### 2.3.2.3 Averaging

We document averaging methods according to five criteria: type/units, subtype/calculation, amount, measure of center, and mean type (Figure 1). Type/units refer to the averaging units of time, breath, and digital filters. Subtype/calculation involves specific computations like bin and rolling averages or digital filter forms. The amount is the unit quantity. For example, 30 for a time average is 30 seconds but is 30 breaths for a breath average. Measure of center distinguishes between mean or median, and mean type delineates whole vs. trimmed mean. Trimmed (truncated) means exclude a number of the highest and lowest values in the quantity before averaging the remaining data.

Descriptions of averaging methods are also considerably more diverse and generic than outlier and interpolation descriptions. For example, “30-second averages” and “averaged every 30 seconds” invite complexity, leading to more snippets referring to averaging something besides BBB gas exchange data. Given that, we required that the text snippets include a reference to gas data such as the text “O<sub>2</sub>,” “breath,” “gas,” “ventilation,” etc.

In contrast to previous studies, we also documented every averaging method we found per paper instead of only describing the averaging method for VO<sub>2</sub>max. We also recorded multiple averaging methods when the authors described the sampling interval and the transformation applied to it. For example, the snippet from [35]

ath method using the vmax respiratory gas analyzer (sensormedics, yorba linda, ca). vo2max was defined as the mean of the three highest values of the averaged oxygen consumption measured consecutively OVER 20-S intervals. a total of 98% of the subjects achieved the respiratory exchange ratio of  $\geq 1.1$ . electrocardiography was recorded throughout the exercise test using cardiosoft software (ge medical systems,

states that oxygen consumption was measured every 20 seconds and that VO<sub>2</sub>max was calculated as the average of three 20-second intervals, or 60-seconds. For this article, we documented one averaging method as a 20-second time bin whole mean and another as a 60-second time bin whole mean.

In many cases, authors did not explicitly use the terms “average” or “mean” to describe their averaging methods, but we documented their methods when implied. For example, the snippet from [36] reading

red using a continuously monitored electrocardiograph. blood pressure was measured at the end of each workload increment using an automatic sphygmomanometer. peak v9o2 was defined as the v9o2 measured DURING THE LAST 30 S of peak exercise. oxygen pulse was calculated by dividing v9o2 by cardiac frequency. the anaerobic threshold was detected using the v-slope method [16]. the ventilatory equivalent for carbon dioxide w

states they calculated VO<sub>2</sub>peak using the last 30 seconds of exercise data. We documented such phrasing as a 30-second time-bin whole mean average.

### 2.3.3 Data Items

In all cases, articles that did not return any phrases were documented as “not described” for their respective data processing category. If snippets did not refer to the data processing category or if the snippet lacked sufficient information, those data processing variables were documented as “not described.” For example, interpolation variables were denoted as “not described” if interpolation was acknowledged but without details for the interpolation type or time.

#### 2.3.3.1 Outliers

We documented the outlier limit, for example,  $\pm 3$  standard deviations, and any outlier function used to compute the outlier limit, if described.

#### 2.3.3.2 Interpolation

We recorded the interpolation type (linear, cubic, Lagrange, specifically *uninterpolated*, and other) and time frame (e.g., every one second).

#### 2.3.3.3 Averaging

We noted the following averaging types: Time, breath, breath-time, time-breath, time-time, digital filter, ensemble, (explicitly) *unaveraged*, and other. Averaging subtypes included bin, rolling, bin-roll, rolling-bin, Butterworth low-pass, Fast Fourier Transform (FFT), and Savitsky-Golay. Next, we recorded the time in seconds or the number of breaths. We recorded the measure of center as mean or median. Finally, we noted if the mean was a whole or trimmed.

### 2.3.4 Synthesis of Results

Counts, percentages, and margin of error (95% confidence) were calculated for the reporting frequency of each data processing method using R [37] and RStudio [38].

## 3. Results

### 3.1 Selection of Sources of Evidence

Figure 2 shows the selection of sources of evidence flowchart. During our analysis, we identified 1,352 ineligible articles. We cross-referenced those against the breath-by-breath articles and removed another 354, leading to 8,351 articles.

### 3.2 Characteristics and Results of Individual Sources of Evidence

The PRISMA Extension for Scoping Reviews checklist normally requires a section to report the characteristics and results of individual sources of evidence, usually in a table format, including citations [27]. Given the vast nature of this scoping review, readers can instead view web links to our [outlier](#), [interpolation](#), and [averaging](#) data charting spreadsheets.

### 3.3 Synthesis of Results

We present our results according to the reporting prevalence followed by the specific characteristics when reported.

#### 3.3.1 Outliers

Of the 8,351 articles, 376 ( $4.5 \pm 0.4\%$ ) reported outlier removal methods. Of the articles reporting their outlier methods, the most prevalent methods were  $\pm 3$  (39.1%) and  $\pm 4$  (51.6%) standard deviations, respectively (Figure 3).

Only 102 ( $1.2 \pm 0.2\%$ ) articles reported details of the function they used to calculate their outlier limit. Of those, breath-based averages ( $n = 76$ , 74.5%) then time-based averages ( $n = 15$ , 14.7%) were the most common for calculating outlier boundaries. Specifically, 5-breath averages ( $n = 54$ , 52.9%) were the most prevalent functions to calculate outlier limits.

#### 3.3.2 Interpolation

We found that 581 ( $7.0 \pm 0.5\%$ ) out of 8,351 specified their interpolation methodology. When reported, the most common interpolation time was one second ( $n = 527$ , 93.9%). Although the majority of articles reporting interpolation procedures did not explicitly specify their interpolation method ( $n = 314$ , 54.0%), linear interpolation was the most popular stated method ( $n = 247$ , 92.5%) (see Table 1 and Figure 4).

Table 1: Most prevalent specified interpolation methods by type (a) and by time (b).

#### 3.3.3 Averaging

After removing 22 ineligible articles that we discovered during data documentation from the original 1,100 random articles, we analyzed 1,078 articles for our averaging analysis. We recorded that 852 ( $66.8 \pm 2.8\%$ ) reported some details of their data averaging methods. Time averages dominated in popularity (91.5%) (Table 2). Bin averages proved the most widespread averaging subtype (89.9%) (Table 2). Together, time-bin (86.8%) was the most frequent type-subtype averaging method combination.

Table 2: Averaging methods by type (a) and subtype (b).

When incorporating averaging amounts, 30-, 60-, 15-, and 10-second bin averages (Figure 5) were the most popular. The “other” methods category accounted for the second highest share of the total, but this represents many rarely used averaging methods.

## 4. Discussion

### 4.1 Summary of Evidence

This review shows that gas exchange data processing methods are infrequently reported for outlier removal and interpolation. We consider outlier removal



documentation important as it applies to many exercise test analyses. Removing outliers is important to  $\text{VO}_2$  kinetics and similar research with rapid intensity changes because they rely on high temporal resolution. Outlier removal is also relevant for maximal exercise testing as outliers near the end of a test may influence  $\text{VO}_{2\text{max}}$  or  $\text{VO}_{2\text{peak}}$ . Previous research indicates that a  $\text{VO}_{2\text{max}}$  below the 20th percentile for age and sex increases the risk of all-cause mortality [39], so accurate determinations of  $\text{VO}_{2\text{max}}$  are important for individuals with low cardiorespiratory fitness: an erroneous breath yielding an overestimated  $\text{VO}_{2\text{max}}$  may subdue the urgency to improve cardiovascular health for low-fitness individuals.

Outliers could also affect mathematical  $\text{VO}_2$  plateau determinations. Such methods test if neighboring  $\text{VO}_2$  values or a  $\text{VO}_2$  vs. time slope does not change or increase by more than a set rate (e.g., 50 mL/min) at the end of a maximal test. [10,40–43]. Though data averaging dampens their influence, outliers present near the conclusion of a maximal test could plausibly interfere with mathematical  $\text{VO}_2$  plateau determination.

We are currently unaware of research that has tested this, but outliers may interfere with submaximal thresholds found using algorithms, especially if they exist near likely breakpoints. Threshold algorithms often fit piecewise linear regressions and solve for the lowest sums of squares [44–46]. Points near the edges of the regression lines have more leverage when solving for the best-fit line and, therefore, are more likely to influence the slope or intercept. Such changes could alter the intersection point of the piecewise regression, and thus, the threshold values.

Finally, even fewer articles reported the outlier limit calculation function. As the function chosen impacts calculated outlier limit, it also affects where values are considered outliers. We are unaware of a recommended outlier removal function but encourage stating such details.

We find the low interpolation reporting more reasonable because this procedure is most relevant to less frequent  $\text{VO}_2$  kinetics studies. However, the V-slope method, one of the most common methods for determining the first ventilatory threshold, interpolates data in their original method [45]. Importantly, the V-slope algorithm is only part of the overall V-slope method, so it can be unclear if authors interpolated data when citing the V-slope method. Given this and the artificial confidence interval shrinkage, it may be prudent for future papers to specify interpolation or lack thereof.

Most studies use one-second linear interpolation, but different time frames and styles, such as cubic interpolation, may yield different results. Cubic spline interpolation produces a smooth curve but may slightly “overshoot” measured values [47]. Though likely small, we recommend authors specify the interpolation type.

Despite a much higher percentage of papers describing at least some of their averaging methods, a third of studies examined in this review neglected to document their process. Data averaging likely contributes more to the final calculated values of  $\text{VO}_{2\text{max}}$  and other variables than do outlier removal and interpolation. Indeed, the research on the effect of interpolation on  $\text{VO}_2$  kinetics parameters shows that interpolation does not significantly affect the values of parameter estimates [20–23]. Although we are unaware of studies comparing the effect of outlier removal or leaving

data as-is before proceeding with other calculations, the known impact of data averaging on  $\text{VO}_2\text{max}$  and the inherent dampening effect of averaging on outliers itself suggests that data averaging is the most important of the three steps when the goal is to reflect the underlying whole-body metabolic rate. Therefore, researchers should state their gas exchange data averaging methods to improve research reproducibility and study comparisons.

Stating averaging methods can also help correctly classify cardiorespiratory fitness against normative data. Research by [12] offers a strategy to compare two  $\text{VO}_2\text{max}$  values obtained with different averaging methods. Without such corrections, one could misclassify cardiorespiratory fitness based on  $\text{VO}_2\text{max}$  if  $\text{VO}_2\text{max}$  were calculated with a sufficiently different sampling interval than that used to generate the normative data. Importantly, the normative data offered by the American College of Sports Medicine [1, table 4.9, pp. 88-93] is based on a regression of  $\text{VO}_2$  vs. time-to-exhaustion using a modified Balke protocol and equations developed from [48] and [49] (Cooper Institute, personal communication, 9/2021), rather than directly measured. The system used to create the regression for males [49] and females [48] averaged the data every minute and every 30 seconds, respectively. Given that, stating the averaging methods used may allow for better comparisons to normative data.

The most frequent, fully specified data averaging method, the 30-second time average, fits the maximum recommended guideline by Robergs [3]. [3] also recommended the 15-breath rolling average or the low-pass digital filter, but we only documented these methods two ( $0.2 \pm 0.3\%$ ) and one ( $0.1 \pm 0.2\%$ ) times, respectively.

## 4.2 Limitations

This study presents the most extensive review of gas exchange data processing methods to date. However, due to its scope, not every article received a detailed examination, which means some data processing descriptions might have been missed due to the limitations of our RegExs, leading us to categorize these as “not described.” Articles that referred to previous works for their data processing techniques were also marked as “not described” for simplicity. We realize authors must balance adequate methodological documentation with journal word or character limits. Yet, methodological shortcut citations can mean missing details that prevent readers from fully reproducing the methods used [50]. Next, by chance, we found rare examples of articles using the median as the measure of center as we built our RegExs. However, we did not document any such cases in our random sample. A larger sample would likely find these and other rare data averaging methods. Finally, it is possible that a few ineligible articles eluded our screening. Taken together, our results are not entirely comprehensive and may slightly underestimate data processing methods’ true reporting frequency.

Another limitation of this scoping review is that our results do not indicate how different data processing methods were used. For example, we did not distinguish if a 60-second time-bin average was used to calculate  $\text{VO}_2\text{max}$  or a steady-state exercise period. Therefore, this review cannot estimate the prevalence of different processing methods

for specific analyses, such as  $\text{VO}_2\text{max}$ . Nevertheless, this is the first study we know of to document data processing methods *besides* those used to calculate  $\text{VO}_2\text{max}$ .

### 4.3 Conclusions

This scoping review found that data processing methods were seldom reported for outlier removal and interpolation, and that averaging reporting, though much higher, could further improve. The results reflect prevalent methods. While prevalence should not be conflated with quality, knowing the prevalent methods can allow others to test the influence data processing in this field by comparing relevant options. Finally, we hope these results motivate others to improve their methodological documentation and, thus, reproducibility in this field.

## References

- [1] Pescatello LS. ACSM's guidelines for exercise testing and prescription. 9th ed. Philadelphia: Wolters Kluwer/Lippincott Williams & Wilkins Health; 2014
- [2] Jamnick NA, Pettitt RW, Granata C, et al. An Examination and Critique of Current Methods to Determine Exercise Intensity. *Sports Med* 2020; 50: 1729–1756. doi:[10.1007/s40279-020-01322-8](https://doi.org/10.1007/s40279-020-01322-8)
- [3] Robergs RA, Dwyer D, Astorino T. Recommendations for Improved Data Processing from Expired Gas Analysis Indirect Calorimetry. *Sports Med* 2010; 40: 95–111. doi:[10.2165/11319670-000000000-00000](https://doi.org/10.2165/11319670-000000000-00000)
- [4] Robergs RA, Burnett AF. Methods Used To Process Data From Indirect Calorimetry And Their Application To  $\text{VO}_2$  Max. *J Exerc Physiol Online* 2003; 6
- [5] Sousa A, Figueiredo P, Oliveira N, et al. Comparison Between Swimming  $\text{VO}_{2\text{peak}}$  and  $\text{VO}_{2\text{max}}$  at Different Time Intervals. *Open Sports Sci J* 2010; 3: 22–24. doi:[10.2174/1875399X01003010022](https://doi.org/10.2174/1875399X01003010022)
- [6] Johnson JS, Carlson JJ, VanderLaan RL, et al. Effects of Sampling Interval on Peak Oxygen Consumption in Patients Evaluated for Heart Transplantation. *Chest* 1998; 113: 816–819. doi:[10.1378/chest.113.3.816](https://doi.org/10.1378/chest.113.3.816)
- [7] Sell KM, Ghigiarelli JJ, Prendergast JM, et al. Comparison of  $\text{V}\dot{\text{O}}_{2\text{peak}}$  and  $\text{V}\dot{\text{O}}_{2\text{max}}$  at Different Sampling Intervals in Collegiate Wrestlers. *J Strength Cond Res* 2021; 35: 2915–2917. doi:[10.1519/JSC.0000000000003887](https://doi.org/10.1519/JSC.0000000000003887)
- [8] Midgley AW, McNaughton LR, Carroll S. Effect of the  $\text{V}\dot{\text{O}}_2$  time-averaging interval on the reproducibility of  $\text{V}\dot{\text{O}}_{2\text{max}}$  in healthy athletic subjects. *Clin Physiol Funct Imaging* 2007; 27: 122–125. doi:[10.1111/j.1475-097X.2007.00725.x](https://doi.org/10.1111/j.1475-097X.2007.00725.x)
- [9] Astorino TA. Alterations in  $\text{VO}_2$  max and the  $\text{VO}_2$  plateau with manipulation of sampling interval. *Clin Physiol Funct Imaging* 2009; 29: 6067. doi:[10.1111/j.1475-097X.2008.00835.x](https://doi.org/10.1111/j.1475-097X.2008.00835.x)

- [10] Astorino TA, Robergs RA, Ghiasvand F, et al. Incidence of the oxygen plateau at VO<sub>2</sub>max during exercise testing to volitional fatigue. *J Exerc Physiol Online* 2000; 3: 112
- [11] Martin-Rincon M, González-Henríquez JJ, Losa-Reyna J, et al. Impact of data averaging strategies on V̇O<sub>2</sub>max assessment: Mathematical modeling and reliability. *Scand J Med Sci Sports* 2019; 29: 1473–1488. doi:[10.1111/sms.13495](https://doi.org/10.1111/sms.13495)
- [12] Martin-Rincon M, Calbet JAL. Progress Update and Challenges on VO<sub>2</sub>max Testing and Interpretation. *Front Physiol* 2020; 11: 1070. doi:[10.3389/fphys.2020.01070](https://doi.org/10.3389/fphys.2020.01070)
- [13] Scheadler CM, Garver MJ, Hanson NJ. The Gas Sampling Interval Effect on VO<sub>2</sub>peak Is Independent of Exercise Protocol. *Med Sci Sports Exerc* 2017; 49: 1911–1916. doi:[10.1249/MSS.0000000000001301](https://doi.org/10.1249/MSS.0000000000001301)
- [14] Jesus K de, Guidetti L, Jesus K de, et al. Which Are The Best VO<sub>2</sub> Sampling Intervals to Characterize Low to Severe Swimming Intensities? *Int J Sports Med* 2014; 35: 1030–1036. doi:[10.1055/s-0034-1368784](https://doi.org/10.1055/s-0034-1368784)
- [15] Hill DW, Stephens LP, Blumoff-Ross SA, et al. Effect of sampling strategy on measures of VO<sub>2</sub>peak obtained using commercial breath-by-breath systems. *Eur J Appl Physiol* 2003; 89: 564–569. doi:[10.1007/s00421-003-0843-1](https://doi.org/10.1007/s00421-003-0843-1)
- [16] Smart NA, Jeffriess L, Giallauria F, et al. Effect of duration of data averaging interval on reported peak VO<sub>2</sub> in patients with heart failure. *Int J Cardiol* 2015; 182: 530–533. doi:[10.1016/j.ijcard.2014.12.174](https://doi.org/10.1016/j.ijcard.2014.12.174)
- [17] Matthews JI, Bush BA, Morales FM. Microprocessor Exercise Physiology Systems vs a Nonautomated System. *Chest* 1987; 92: 696–703. doi:[10.1378/chest.92.4.696](https://doi.org/10.1378/chest.92.4.696)
- [18] Lamarra N, Whipp BJ, Ward SA, et al. Effect of interbreath fluctuations on characterizing exercise gas exchange kinetics. *J Appl Physiol* 1987; 62: 2003–2012. doi:[10.1152/jappl.1987.62.5.2003](https://doi.org/10.1152/jappl.1987.62.5.2003)
- [19] Keir DA, Murias JM, Paterson DH, et al. Breath-by-breath pulmonary O<sub>2</sub> uptake kinetics: Effect of data processing on confidence in estimating model parameters. *Exp Physiol* 2014; 99: 1511–1522. doi:[10.1113/expphysiol.2014.080812](https://doi.org/10.1113/expphysiol.2014.080812)
- [20] Benson AP, Bowen TS, Ferguson C, et al. Data collection, handling, and fitting strategies to optimize accuracy and precision of oxygen uptake kinetics estimation from breath-by-breath measurements. *J Appl Physiol* 2017; 123: 2272–2282. doi:[10.1152/jappphysiol.00988.2016](https://doi.org/10.1152/jappphysiol.00988.2016)
- [21] Francescato MP, Cettolo V, Bellio R. Confidence intervals for the parameters estimated from simulated O<sub>2</sub> uptake kinetics: Effects of different data treatments. *Exp Physiol* 2014; 99: 1871–1895. doi:[10.1113/expphysiol.2013.076208](https://doi.org/10.1113/expphysiol.2013.076208)

- [22] Francescato MP, Cettolo V. The 1-s interpolation of breath-by-breath O<sub>2</sub> uptake data to determine kinetic parameters: the misleading procedure. *Sport Sci Health* 2019; 16: 193. doi:[10.1007/s11332-019-00602-9](https://doi.org/10.1007/s11332-019-00602-9)
- [23] Francescato MP, Cettolo V, Bellio R. Interpreting the confidence intervals of model parameters of breath-by-breath pulmonary O<sub>2</sub> uptake. *Exp Physiol* 2015; 100: 475475. doi:[10.1113/EP085043](https://doi.org/10.1113/EP085043)
- [24] Goodman SN, Fanelli D, Ioannidis JPA. What does research reproducibility mean? *Sci Transl Med* 2016; 8. doi:[10.1126/scitranslmed.aaf5027](https://doi.org/10.1126/scitranslmed.aaf5027)
- [25] Open Science Collaboration. Estimating the reproducibility of psychological science. *Science* (1979) 2015; 349: aac4716. doi:[10.1126/science.aac4716](https://doi.org/10.1126/science.aac4716)
- [26] *blinded for peer review*
- [27] Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Ann Intern Med* 2018; 169: 467–473. doi:[10.7326/M18-0850](https://doi.org/10.7326/M18-0850)
- [28] Peters MDJ, Marnie C, Tricco AC, et al. Updated methodological guidance for the conduct of scoping reviews. *JBIM Evid Synth* 2020; 18: 2119–2126. doi:[10.11124/JBIES-20-00167](https://doi.org/10.11124/JBIES-20-00167)
- [29] NCBI. ID Converter API. PubMed Central (PMC) 2022; Im Internet: <https://www.ncbi.nlm.nih.gov/pmc/tools/id-converter-api/>; Stand: 06.04.2022
- [30] Unpaywall: An open database of 20 million free scholarly articles. Im Internet: <https://unpaywall.org/>; Stand: 24.07.2024
- [31] Bojanowski P, Grave E, Joulin A, et al. Enriching Word Vectors with Subword Information. *arXiv preprint arXiv:160704606* 2016;
- [32] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res* 2011; 12: 28252830
- [33] Breese BC, Saynor ZL, Barker AR, et al. Relationship between (non)linear phase II pulmonary oxygen uptake kinetics with skeletal muscle oxygenation and age in 1115 year olds. *Exp Physiol* 2019; 104: 1929–1941. doi:[10.1113/EP087979](https://doi.org/10.1113/EP087979)
- [34] Hartman ME, Ekkekakis P, Dicks ND, et al. Dynamics of pleasure-displeasure at the limit of exercise tolerance: conceptualizing the sense of exertional physical fatigue as an affective response. *J Exp Biol* 2018; jeb.186585. doi:[10.1242/jeb.186585](https://doi.org/10.1242/jeb.186585)
- [35] Hassinen M, Lakka TA, Savonen K, et al. Cardiorespiratory Fitness as a Feature of Metabolic Syndrome in Older Men and Women. *Diabetes Care* 2008; 31: 1242–1247. doi:[10.2337/dc07-2298](https://doi.org/10.2337/dc07-2298)



- [36] Deboeck G, Niset G, Lamotte M, et al. Exercise testing in pulmonary arterial hypertension and in chronic heart failure. *Eur Respir J* 2004; 23: 747–751. doi:[10.1183/09031936.04.00111904](https://doi.org/10.1183/09031936.04.00111904)
- [37] R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2021
- [38] Posit team. *RStudio: Integrated Development Environment for R*. Boston, MA: Posit Software, PBC; 2022
- [39] Blair SN, Kohl HW, Barlow CE, et al. Changes in physical fitness and all-cause mortality: a prospective study of healthy and unhealthy men. *JAMA* 1995; 273: 1093–1098. doi:[10.1001/jama.1995.03520380029031](https://doi.org/10.1001/jama.1995.03520380029031)
- [40] Robergs RA. An exercise physiologist's "contemporary" interpretations of the "ugly and creaking edifices" of the VO<sub>2</sub>max concept. *J Exerc Physiol Online* 2001; 4: 144
- [41] Myers J, Walsh D, Buchanan N, et al. Can Maximal Cardiopulmonary Capacity be Recognized by a Plateau in Oxygen Uptake? *Chest* 1989; 96: 1312–1316. doi:[10.1378/chest.96.6.1312](https://doi.org/10.1378/chest.96.6.1312)
- [42] Myers J, Walsh D, Sullivan M, et al. Effect of sampling on variability and plateau in oxygen uptake. *J Appl Physiol* 1990; 68: 404–410. doi:[10.1152/jappl.1990.68.1.404](https://doi.org/10.1152/jappl.1990.68.1.404)
- [43] Yoon B-K, Kravitz L, Robergs R. V O<sub>2</sub>max, protocol duration, and the V O<sub>2</sub> plateau. *Med Sci Sports Exerc* 2007; 39: 1186–1192
- [44] Jones RH, Molitoris BA. A statistical method for determining the breakpoint of two lines. *Anal Biochem* 1984; 141: 287–290. doi:[10.1016/0003-2697\(84\)90458-5](https://doi.org/10.1016/0003-2697(84)90458-5)
- [45] Beaver WL, Wasserman K, Whipp BJ. A new method for detecting anaerobic threshold by gas exchange. *J Appl Physiol* 1986; 60: 2020–2027. doi:[10.1152/jappl.1986.60.6.2020](https://doi.org/10.1152/jappl.1986.60.6.2020)
- [46] Orr GW, Green HJ, Hughson RL, et al. A computer linear regression model to determine ventilatory anaerobic threshold. *J Appl Physiol* 1982; 52: 1349–1352. doi:[10.1152/jappl.1982.52.5.1349](https://doi.org/10.1152/jappl.1982.52.5.1349)
- [47] Zhang Z, Martin CF. Convergence and Gibbs' phenomenon in cubic spline interpolation of discontinuous functions. *J Comput Appl Math* 1997; 87: 359–371. doi:[10.1016/S0377-0427\(97\)00199-4](https://doi.org/10.1016/S0377-0427(97)00199-4)
- [48] Pollock ML, Foster C, Schmidt D, et al. Comparative analysis of physiologic responses to three different maximal graded exercise test protocols in healthy women. *Am Heart J* 1982; 103: 363–373
- [49] Pollock ML, Bohannon RL, Cooper KH, et al. A comparative analysis of four protocols for maximal treadmill stress testing. *Am Heart J* 1976; 92: 39–46. doi:[10.1016/S0002-8703\(76\)80401-2](https://doi.org/10.1016/S0002-8703(76)80401-2)

[50] Standvoss K, Kazezian V, Lewke BR, et al. Shortcut citations in the methods section: Frequency, problems, and strategies for responsible reuse. PLoS Biol 2024; 22: e3002562. doi:[10.1371/journal.pbio.3002562](https://doi.org/10.1371/journal.pbio.3002562)

## Table and Figure Captions

### Table Captions

Table 1: Most prevalent specified interpolation methods by type (a) and by time (b).

Table 2: Averaging methods by type (a) and subtype (b).

### Figure Captions

Figure 1: Flowchart depicting the four major components of averaging method documentation.

Figure 2: Selection of sources of evidence flowchart. Dashed lines point to articles that were removed. Solid lines indicate the path of articles that remained in the analysis. Details are provided in the main text. “Unobtained” articles were those unavailable due to University library licensing limitations or faulty download links. “Resolvable” files could be analyzed by RegExs and ML.

Figure 3: Counts and percentages of outlier limits when specified.

Figure 4: Most prevalent specified interpolation methods by both type and time.

Figure 5: Prevalence of complete averaging procedures. The numbers in each column label are in seconds for time averages and the number of breaths for breath averages. The “other” column represents methods that accounted for less than 1% of the total stated methods.

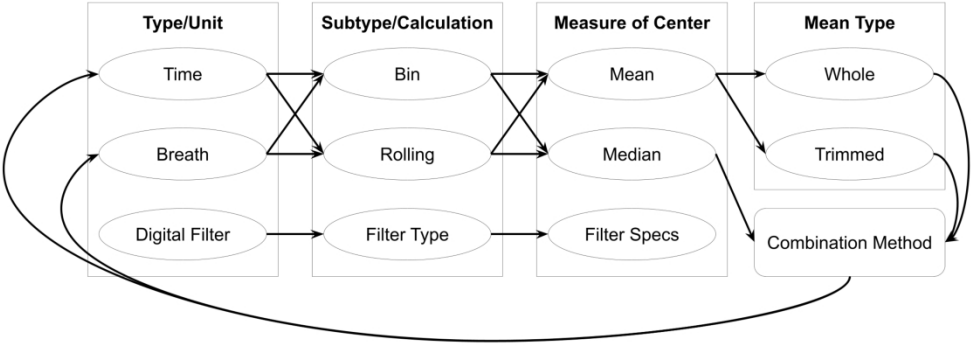


Figure 1: Flowchart depicting the four major components of averaging method documentation.

1438x1079mm (72 x 72 DPI)

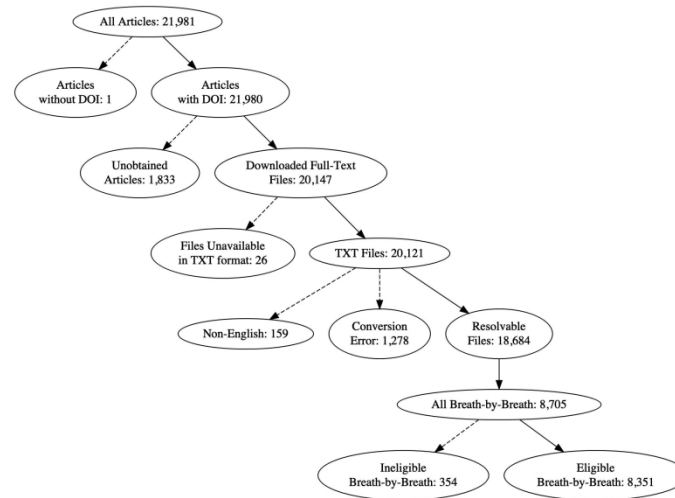


Figure 2: Selection of sources of evidence flowchart. Dashed lines point to articles that were removed. Solid lines indicate the path of articles that remained in the analysis. Details are provided in the main text. "Unobtained" articles were those unavailable due to University library licensing limitations or faulty download links. "Resolvable" files could be analyzed by RegExs and ML.

1058x787mm (72 x 72 DPI)

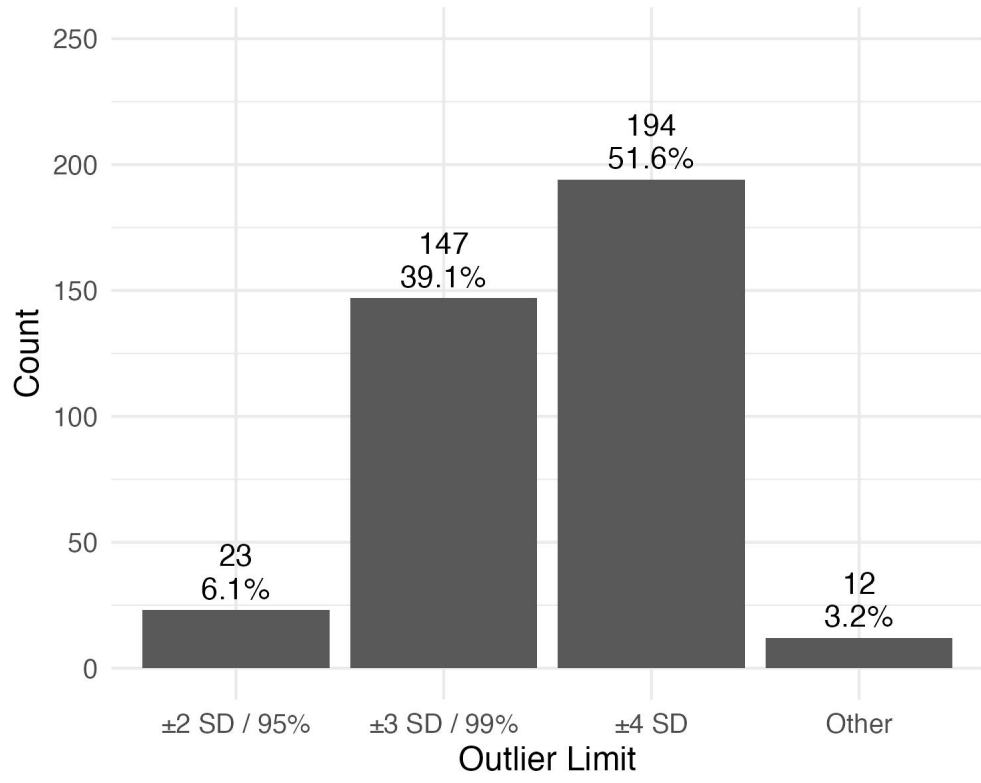


Figure 3: Counts and percentages of outlier limits when specified.

529x422mm (72 x 72 DPI)



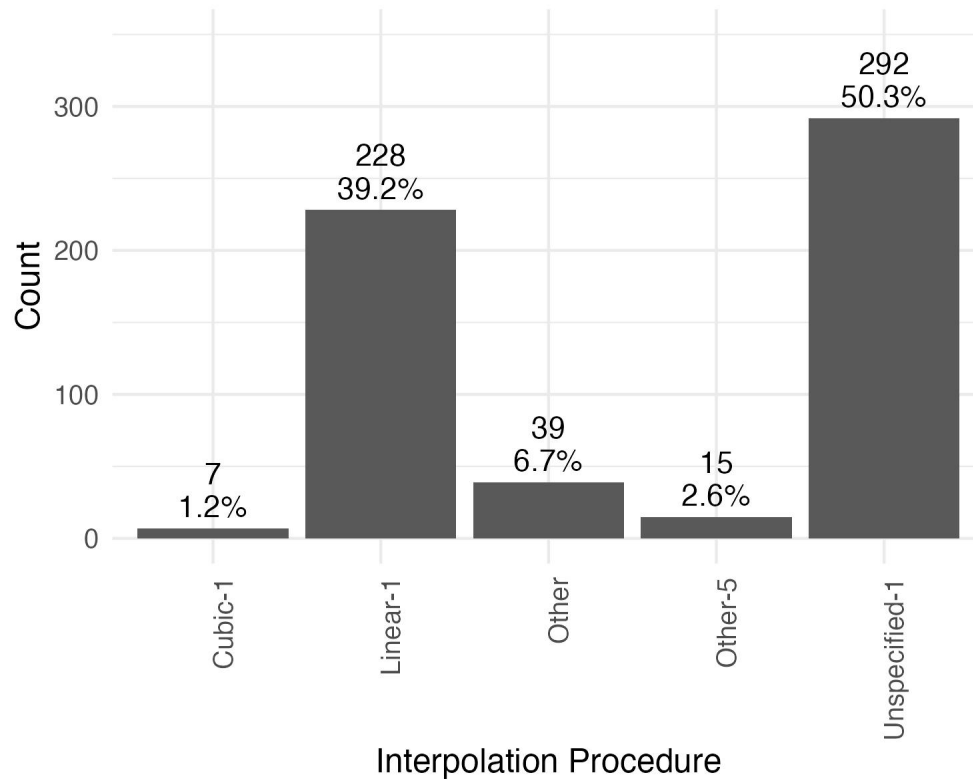


Figure 4: Most prevalent specified interpolation methods by both type and time.

529x422mm (72 x 72 DPI)

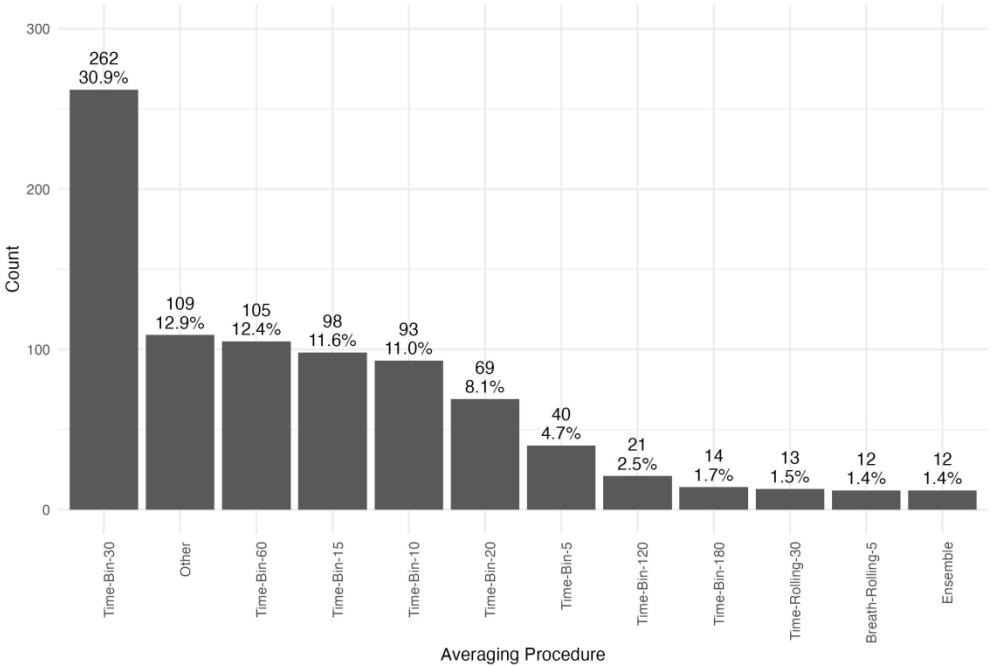


Figure 5: Prevalence of complete averaging procedures. The numbers in each column label are in seconds for time averages and the number of breaths for breath averages. The “other” column represents methods that accounted for less than 1% of the total stated methods.

917x621mm (72 x 72 DPI)

Interpolation Type	N	%
Linear	247	92.5
Cubic	10	3.7
Other	5	1.9
Uninterpolated	3	1.1
Lagrange	2	0.7

Table 1 a

Interpolation Time (s)	N	%
1	527	93.9
5	15	2.7
Other	19	3.4

Table 1 b

Table caption: Most prevalent specified interpolation methods by type (a) and by time (b)

Averaging Type	N	%
Time	776	91.5
Breath	38	4.5
Ensemble	13	1.5
Breath-Time	8	0.9
Other	4	0.5
Digital Filter	3	0.4
Time-Time	3	0.4
Unaveraged	2	0.2
Time-Breath	1	0.1

Table 2 a

Averaging Subtype	N	%
Bin	744	89.9
Rolling	58	7.0
Rolling-Bin	16	1.9
Bin-Roll	7	0.8
Fast Fourier Transform	2	0.2
Butterworth Low-Pass	1	0.1

Table 2 b

Figure caption: Averaging methods by type (a) and subtype (b).













Scopus search:

( TITLE-ABS-KEY ( "oxygen consumption" OR "oxygen uptake" OR vo2 ) AND TITLE-ABS-KEY ( cycli\* OR bicycl\* OR run OR runn\* OR treadmill\* OR swim\* OR ski OR skie\* OR skiing OR ergometer\* OR row\* ) AND TITLE-ABS-KEY ( exer\* ) ) AND ( LIMIT-TO ( DOCTYPE , "ar" ) ) AND ( LIMIT-TO ( LANGUAGE , "English" ) ) AND ( LIMIT-TO ( SRCTYPE , "j" ) ) AND ( LIMIT-TO ( EXACTKEYWORD , "Human" ) )

Web of Science Search

((ALL=("oxygen consumption" or "oxygen uptake" or VO2)) AND ALL=(cycli\* or bicycl\* or run or runn\* or treadmill\* or swim\* or ski or skie\* or skii or ergometer\* or row\*)) AND ALL=(exer\*) and English (Languages) and Articles (Document Types)

Ovid-MEDLINE

1. ("oxygen consumption" or "oxygen uptake" or VO2).mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]  
2. (cycli\* or bicycl\* or run or runn\* or treadmill\* or swim\* or ski or skie\* or skii or ergometer\* or row\*).mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]  
3. 1 and 2  
4. 3 and exer\*.mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]  
5. 4 and "Humans".sa\_suba.  
6. 5 and "Journal Article".sa\_pubt.  
7. 6 not ("Review" or "Systematic Review" or "Meta-Analysis").sa\_pubt.