Data Processing Strategies to Determine Maximum Oxygen Uptake: A Systematic Scoping Review and Experimental Comparison

Bachelor thesis

from

Simon Nolte

German Sport University Cologne

Cologne 2022

Thesis supervisor:

Dr. Oliver Jan Quittmann

Institute of Movement and Neurosciences

Affirmation in lieu of an oath

Herewith I affirm in lieu of an oath that I have authored this Master thesis independently and did not use any other sources and tools than indicated. All citations, either direct quotations or passages which were reproduced verbatim or nearby-verbatim from publications, are indicated and the respective references are named. The same is true for tables and figures. I did not submit this piece of work in the same or similar way or in extracts in another assignment.

_____

Personally signed

Zusammenfassung (German abstract)

Die maximale Sauerstoffaufnahme ($\dot{V}O_{2max}$) ist die wichtigste physiologische Kenngröße der Ausdauerleistungsfähigkeit. Zur Bestimmung der $\dot{V}O_{2max}$ werden während eines Rampentests die Atemgase kontinuierlich gemessen. Die üblicherweise im breath-by-breath-Verfahren erhobenen Daten erfordern eine Weiterverarbeitung um die $\dot{V}O_{2max}$ zu bestimmen. Unterschiedliche Verfahren der Datenverarbeitung können die gemessenen $\dot{V}O_{2max}$ beeinflussen. Ziel der vorliegenden Arbeit ist es, anhand eines scoping review zu bestimmen, welche Datenverarbeitungsstrategien in der wissenschaftlichen Literatur genutzt werden, um diese im Anschluss anhand experimenteller Daten zu vergleichen. Für den Review wurde eine Zufallsstichprobe von 500 Artikeln untersucht; der experimentelle Vergleich erfolgte anhand von 72 standardisierten Rampentests auf dem Laufband. Die im Review untersuchten Studien benutzen eine Vielzahl von verschiedenen Datenverabeitungsstrategien, welche häufig nur unzureichend beschrieben werden. Die unterschiedlichen Strategien können die $\dot{V}O_{2max}$ im Median zu mehr als 5% beeinflussen, mit deutlich höheren individuellen Abweichungen. Die Ergebnisse verdeutlichen den Bedarf einer Standardisierung und korrekter Darstellung der benutzten Datenverarbeitungsstrategien zur Bestimmung der $\dot{V}O_{2max}$. Diese Arbeit bietet hierfür Empfehlungen zur Verwendung und Vergleichbarkeit verschiedener Strategien, sowie Richtlinien zur Beschreibung dieser im Rahmen wissenschaftlicher Studien.

Table of contents

List of Figures

List of Tables

1 Introduction

1.1 Background

Performance in endurance sports is limited by the human physiology. Athletes need to supply energy to the contracting muscles for locomotion, a process that happens mainly via the oxidative phosphorylation. A higher maximal activity of the oxidative phosphorylation allows to supply more energy, and thus to move faster [1]. The highest activity of oxidative phosphorylation on a cellular level can be approximated by measuring the maximum oxygen uptake ($\dot{V}O_{2max}$) on a whole-body level.

$\dot{V}O_{2max}$ is defined as the highest rate a body can consume oxygen [2]. Determined by measuring gas exchange data, it is one of the most common assessed exercise parameters in sports science and medicine. $\dot{V}O_{2max}$ highly corresponds with endurance performance in heterogeneous groups [3–5] and can be regarded the most relevant physiological parameter for predicting endurance performance [2]. Many training programs aim to target the $\dot{V}O_{2max}$, and studies evaluate the quality of interventions by measuring changes in $\dot{V}O_{2max.}$

Researchers predominantly measure $\dot{V}O_{2max}$ during exercise tests to exhaustion. In a laboratory setting such tests commonly take place on treadmills or cycling ergometer. Fatigue prior to reaching the $\dot{V}O_{2max}$ results in underestimating its true value. Therefore most exercise protocols consist of a continuous or stepwise increase in load with an optimal duration of 8-12 minutes [6,7]. While these narrow limits have been challenged, considerably shorter or longer protocols should not be used [8]. Despite protocol type and protocol duration, factors such as motivation [9], exercise modality [10] and biological variability [11] bias the $\dot{V}O_{2max}$ determination.

Early research on $\dot{V}O_{2max}$ used bags to collect expired air for later analysis — the so called Douglas bag method [12]. Today's modern metabolic carts allow for simultaneous collection and analysis of gases, using either mixing chambers (which often sample data at fixed time intervals, e.g., 15 seconds) or breath-by-breath methods [13]. While mixing chambers are regarded as more reliable [14], measuring breath-by-breath generates data with a higher temporal resolution [15]. The measuring method and the metabolic cart model can influence the measured oxygen uptake — and as such the $\dot{V}O_{2max}$ [16,17]. But even the same oxygen uptake data generated during an exercise test may result in varying outcomes when processed differently.

Whether an observed peak in oxygen uptake corresponds to the true maximum has been extensively discussed in the past decades [18–22]. To identify a true maximum, researchers commonly evaluate a set of parameters measured during ramp tests — the so called '$\dot{V}O_{2max}$ criteria' [19]. If these criteria are (partly) not fulfilled, researchers are advised to speak of a peak oxygen uptake instead of $\dot{V}O_{2max}$ [19]. In this thesis I will not distinguish between peak and maximum oxygen uptake, as the criteria for $\dot{V}O_{2max}$ (e.g. the primary criterion of

a plateau in oxygen uptake or the secondary criterion of the maximum respiratory quotient) do heavily rely on the data processing strategy used [23]. Thus, I speak of $\dot{V}O_{2max}$ as the maximum oxygen uptake measured during an appropriate exercise test regardless of any $\dot{V}O_{2max}$ criteria.

1.2 Previous Research

Measured breath-by-breath data is noisy (see grey points in Figure 1). Both the biological variability of breathing patterns and the measurement error (as well as irregular breaths such as coughs and swallowing) lead to a highly fluctuating raw oxygen uptake. For interpretation the raw data requires some form of processing.

Different data processing strategies influence measured parameters of gas exchange. As soon as when the first automated systems for gas exchange measurement were available, Matthews et al. [24] reported that different processing strategies lead to different $\dot{V}O_{2max}$ values. Since then, many researchers investigated the influence of different data averaging intervals on $\dot{V}O_{2max}$, unsurprisingly showing that shorter calculation intervals lead to higher $\dot{V}O_{2max}$ values [23,25,26]. Midgley et al. [27] found differences between data processing strategies for $\dot{V}O_{2max}$, but no difference in the reliability of these, infering that no optimal strategy exists, but that consistency of strategies is key when comparing outcomes. Based on own data in sedentary and moderately trained individuals, Martin-Rincon et al. [28] provided linear-log equations to compare mean $\dot{V}O_{2max}$ values derived from processing strategies with differing averaging interval lengths. There is inconclusive evidence on whether the influence of different processing strategies interacts with different exercise protocols [29,30].

Differences in $\dot{V}O_{2max}$ due to data processing can cause serious implications in practice. They hinder the assessment of longitudinal data from athletes that participated in diagnostics with differing data processing or the comparability of data from studies within meta-analysis [28]. Data processing strategies directly affect the occurrence of a plateau in oxygen uptake, the primary criterion for $\dot{V}O_{2max}$ [23]. Crucially, in situations where individuals are classified by their $\dot{V}O_{2max}$ — for example when describing the training status of a study population [32,33] or evaluating patients for a heart transplantation [34] — differing processing strategies can lead to misclassifications [25].

Calls to standardize data processing strategies of gas exchange data are frequent [25,29,30,35]. Howley et al. [19] argued to use longer calculation intervals (60 seconds), as shorter intervals may introduce bias towards extreme data values and thus could systematically overestimate true $\dot{V}O_{2max}$. Based on synthetic and experimental data, Robergs & Burnett [36] opposed this conclusion, stating that shorter calculation intervals are less erroneous. They further argued to use breath-based moving averages instead of time averages. In terms of data processing, the current ATS/ACCP guidelines for cardiopulmonary exercise testing remain vague, stating that *“30 to 60-second intervals for averaging data are*

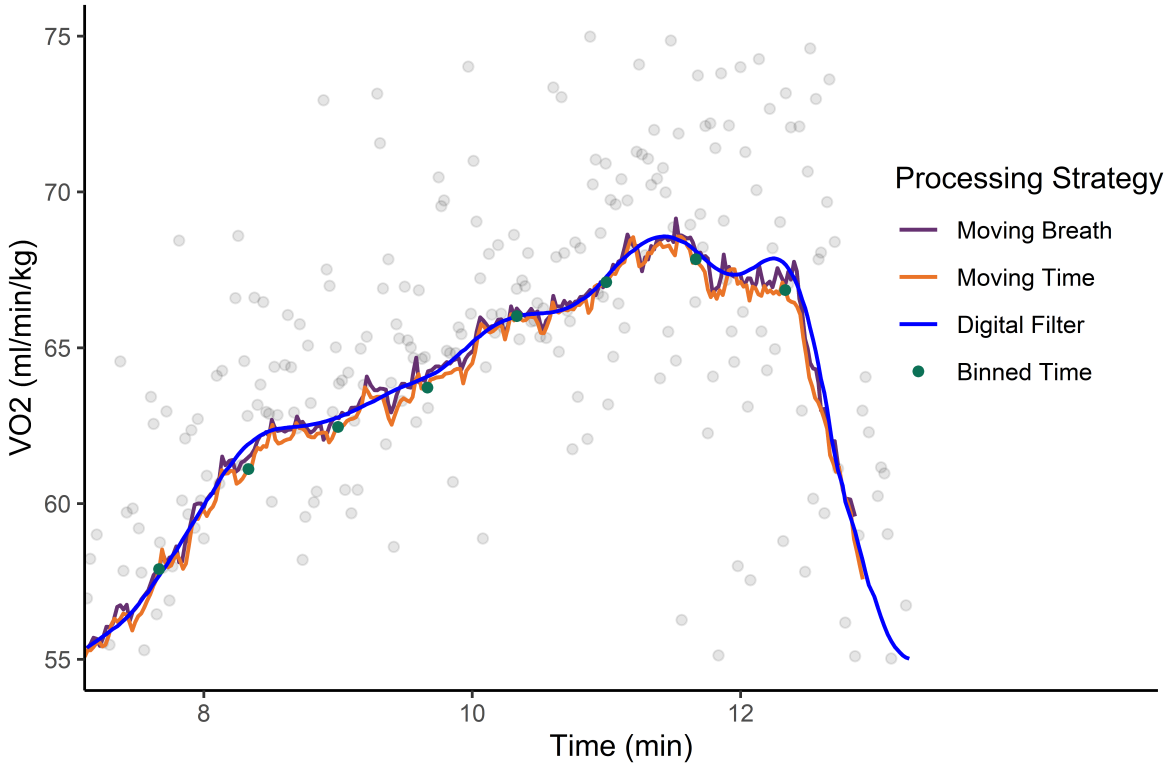

Figure 1: Example of raw breath-by-breath data processed by different strategies during the final minutes of a ramp test to exhaustion. Grey points display oxygen uptake from the single breaths. The moving averages were calculated over 30 breaths and seconds, respectively. The binned average was calculated over 30-second intervals with the mean values aligned to the center of each interval. For digital filtering a forward-backward (zero phase) low-pass Butterworth filter with the parameters 0.04 Hz (cut-off frequency) and 3 (order) for each filter was applied. For more details on the used data processing strategies, see the Section 2.2.2. Data from [31]

recommended, although 20-seconds intervals may be acceptable" [37]. In a classic opinion piece, Robergs et al. [38] argued for using digital filtering (i.e. a low-pass Butterworth filter) for processing gas exchange data, a strategy initially introduced for gas exchange measures by Weir et al. [39]. However, Robergs et al. [38] acknowledged the lack of accessible software implementing such procedures.

In absence of one commonly accepted method, data processing varies among the literature. In light of the influence on outcome variables, many articles highlighted the need to report processing strategies in research articles [23,25,30,35]. Midgley et al. [27] were the first to evaluate reported data processing strategies for breath-by-breath analyses in selected journals. They found that almost all studies reporting their methods choose binned time averaging, with only 1 in 117 using a moving time and a moving breath average, respectively. One third of the studies did not describe their processing method at all. While providing interesting first insights into reporting and processing practices, the search by Midgley et al. [27] was not systematic and its methods were not described in detail. Robergs et al. [38] wrote that to investigate the current state of data processing strategies, two possible approaches are *"(i) a summary of published research, and (ii) a survey circulated via the Internet to as many exercise physiologists as possible"*. They provided the latter one with a total of 75 respondents, who reported a great variety in data processing strategies. Most researchers reported the use of binned time averages over 30 or 60 seconds. Astonishingly, about half of the respondents admitted that their data processing strategy was rather chosen due to subjective influences as opposed to objective reasoning. Practices in reporting and processing may have changed more than a decade after the publication of such numbers.

1.3 Aim

In this thesis I will review the usage and reporting of different data processing strategies in the scientific literature and investigate their influence on the $\dot{V}O_{2max}$. To date, the practices of data processing to determine $\dot{V}O_{2max}$ in the actual scientific work remain largely unknown. Previous research on this topic has only performed unsystematic searches [27] or non-representative surveys [38]. Moreover, recent recommendation [38] as well as advances in measurement devices and analysis software may have changed processing practices in the past 15 years.

Selected data processing strategies have been extensively compared in the literature. Due to the absence of a systematic mapping of current practices, these works lacked the reasoning of which strategies to compare. Many studies compared different averaging intervals, but not averaging types (e.g. moving breath vs. binned time) [23,27]. [28] Martin-Rincon et al. provided formulas for comparing data processing strategies by investigating a data set of sedentary or recreational trained athletes, using two different metabolic carts. Therefore in this work motivation and measurement devices may have interacted with the influence of processing strategies.

Differences in $\dot{V}O_{2max}$ due to data processing strategies may have serious implication for individuals [25], yet almost all comparisons only report mean differences between strategies. No research has yet compared a variety of systematically derived strategies among a group of well-trained individuals using a standardized measurement set-up.

This thesis will first map current practices of data processing for $\dot{V}O_{2max}$ determination by the means of a systematic scoping review of current scientific literature. Secondly, I will compare different data processing strategies in relation to the most commonly applied on a set of 72 standardized treadmill tests in well-trained individuals. The results will help to compare $\dot{V}O_{2max}$ data derived from different processing methods among studies and in individuals. The review allows to assess the implementation of current data processing recommendations and to find malpractices in reporting. The results build a basis for providing new recommendations for data processing and its reporting for determining the $\dot{V}O_{2max.}$

Table 1: Search strings for the systematic scoping review.

Source	Search Strings
PubMed	‘(((((”maximum oxygen uptake”) OR (”maximal oxygen uptake”)) OR (”VO2max”)) OR (”maximum oxygen consumption”)) OR (”maximal oxygen consumption”)) AND ((”2017/01/01”[Date - Publication] : ”3000”[Date - Publication]))‘
Web of Science	‘(((((ALL=(”maximum oxygen uptake”)) OR ALL=(”maximal oxygen uptake”)) OR ALL=(”VO2max”)) OR ALL=(”maximum oxygen consumption”)) OR ALL=(”maximal oxygen consumption”)) AND PY=(2017-2022)‘

2 Methods

The work presented in this thesis was preregistered before the start of the project on the Open Science Framework [40], following the ‘Inclusive Systematic Review Registration Form’ [41]. Any deviations from the preregistration are indicated in the ‘Transparent Changes’ document (Section A.1). Major deviations will also be explicitly stated within the methods section. All data and code of this research project can be found on GitHub.

I conducted all analyses using R Version 4.2.0 [42] in the R Studio IDE Version 2022.2.2.485 [43]. The thesis was entirely written in Quarto Version 0.9.380 [44]. The attached packages and default settings are documented in Section A.2.

2.1 Systematic Scoping Review

The aim of the scoping review was to systematically map current practices of data processing for $\dot{V}O_{2max}$ determination in the scientific literature. Since determining $\dot{V}O_{2max}$ is a far too common procedure to perform an exhaustive search, I randomly sampled 500 articles that referred to $\dot{V}O_{2max}$ or similar keywords. Data on processing strategies were extracted from all sampled articles that directly measured $\dot{V}O_{2max}$ using an appropriate testing procedure in humans. The review was performed in accordance to the PRISMA extension for Scoping reviews [45]. Section A.3 contains the corresponding reporting checklist.

2.1.1 Search & Screening

The article search was conducted on 16th March 2022 using PubMed and Web of Science. The search included articles published from 2017 to 2022 referring to ‘maximum oxygen uptake’ or equivalent terms in title, abstract or keywords. Table 1 shows the exact search terms used.

Table 2: Exclusion criteria for the screening process.

	Criterion	Details
A*	not in English	Full text only available in non-English language
B*	no primary research	research was no original investigation or only a reanalysis of data
C*	research not in humans	research was conducted in animals
D*	$\dot{V}O_{2max}$ only estimated	$\dot{V}O_{2max}$ was only approximated by means of a predictive equation
E	no appropriate test protocol	Protocol for $\dot{V}O_{2max}$ testing did either not include exercise to voluntary exhaustion or was to long (>20 min) for a reliable estimate
F	no information regarding the exclusion criteria	no information regarding the exclusion criteria | crucial information on $\dot{V}O_{2max}$ testing that allowed the evaluation of the other exclusion criteria were missing

*During the abstract screening only the criteria marked with an asterix were evaluated

The search results from both data bases were merged and checked for the presence of a Digital Object Identifier (DOI). Entries without DOI were excluded to allow for automated removal of duplicates by DOI matching in the next step. After removing the duplicates I conducted an automated title scanning to exclude results that were likely no original research articles. All titles that contained one of the following words were excluded: review, correction, meta-analysis, comment, retraction, editorial, erratum, reply.

In accordance with the preregistration I drew a random sample from the search results. The goal of this process was to give an unbiased estimate of the current state of scientific $\dot{V}O_{2max}$ testing. Based on the procedure described in the preregistration, the sample included a total of 500 articles.

The abstracts from the articles included in the random sample were blinded for scanning. This meant removing any further information not relevant for the screening—such as authors or journal—leaving only the title, abstract and an ID of the article (see Section A.4 for an example). Two researchers independently scanned the abstracts to filter those that matched one of the exclusion criteria shown in Table 2. When the screeners disagreed in their assessments, they resolved the conflict by discussion.

After the abstract screening I retrieved the full-texts for all remaining articles. The full-texts were again independently scanned by two researchers to include only those articles that measured $\dot{V}O_{2max}$ using an appropriate testing procedure in humans (see Table 2 for the detailed full-text exclusion criteria). Conflicts were resolved by discussion between examiners. All data exclusion steps are documented in an Markdown script on GitHub.

2.1.2 Data Extraction

I retrieved data from all articles remaining after the abstract and full-text screening. Extraction included the following data:

- metabolic cart used
- measurement type (breath-by-breath, mixing chamber, …)
- type of outcome for $\dot{V}O_{2max}$ (primary, secondary, other)
- data preprocessing (e.g., filtering)
- data processing software
- interpolation procedure
- data processing/determination of $\dot{V}O_{2max}$:
  - type (time average, breath average, digital filtering)
  - alignment (rolling, binned, …)
  - interval (in seconds or breaths, parameters for filtering)
- reference for the used data processing strategy

The criteria 'type of outcome' and 'reference' were added to the extraction list after the abstracts had been scanned, thus they were not stated in the preregistration. All extracted data is available as a csv file on GitHub.

2.1.3 Data Synthesis

The extracted data is presented in a purely descriptive way. I calculated the relative and absolute frequency for the reporting of the extracted items. Similarly, I counted the use of different data strategies for processing data in all articles that reported measuring breath-by-breath. Total interval duration of averaging procedures were derived from the reported parameters.

2.2 Experimental Comparison

To determine the influence of the most common data processing strategies on the determination of $\dot{V}O_{2max}$, I compared them on a set of already collected gas exchange data from ramp tests in running.

2.2.1 Data Source

A total of N = 72 exercise tests were analysed for this study. Due to a miscalculation, the preregistration had incorrectly stated a number of 76 tests. The data were from previous research on the metabolic profile of endurance runners [31,46]. The tested individuals were

experienced distance runners (15 female, 54 male; three of the males participated in both studies). The $\dot{V}O_{2max}$ tests were conducted in March to September 2019 [46] and March to October 2021 [31] using an identical exercise protocol. Participants run on a treadmill (saturn 300/100, h/p/cosmos sports & medical 127 GmbH, Nussdorf-Traunstein, Germany) with 1% inclination for ten minutes at a velocity of 2.8 $m \cdot s^{-1}$ as a warm-up. After preparing the gas exchange measures, they started a ramp protocol with an initial speed of 2.8 $m \cdot s^{-1}$ for two minutes and subsequently increased velocity by 0.15 $m \cdot s^{-1}$ every 30 seconds. The researchers provided verbal encouragement and terminated the exercise when the participants reached subjective exhaustion.

Gas exchange data were recorded using a ZAN 600 USB device (nSpire Health, Inc., Longmont, CO, United States of America). The device was calibrated with a 3l-syringe pump (nSpire Health, Inc., Longmont, CO, 143 United States of America) and a reference gas (15% O_2, 6% CO_2) before each measurement. The measured breath-by-breath data is available on GitHub.

2.2.2 Data Processing

The spiro Package version 0.0.4 for R [47] processed the raw gas exchange data. It includes various algorithms to calculate $\dot{V}O_{2max}$ with user-defined parameters on given data. The full analysis script is available on GitHub.

Moving time-based averages were calculated by first linearly interpolating the breath-by-breath data to seconds. Subsequently a (center aligned) moving average was calculated over a defined timespan. These processing steps are implemented in the spiro package [47]. For calculating the moving average over 30 seconds for example, I used the functions `spiro(data) |> spiro_max(30)`.

Binned time averages were calculated using a custom function (available in the analysis script on GitHub). Breath-by-breath data were initially interpolated to full seconds and then binned into consecutive intervals of constant lengths. The average of each interval was aligned to its center. Incomplete intervals (i.e. the last seconds of measurement) were not considered in the analysis. Note that some authors use a different procedure for determining their bins, starting by the end point of the measurement. However, defining bins beginning by the start of the measurement is a common output option for many gas exchange data analysis software (e.g. Cosmed Omnia).

Breath based moving averages were calculated on the raw data. As this functionality is implemented in the spiro package, I used the functions `spiro(data) |> spiro_max("30b")` for an exemplary 30-breath long averaging interval.

Butterworth filters are a class of recursive digital filters. Robergs et al. [38] and Weir et al. [39] argue that these more advanced processing strategies are superior to the traditional moving or binned average approach for analysing gas exchange data. However Robergs et al. [38]

missed to account for the time lag introduced by Butterworth filters. Therefore I applied a zero-phase Butterworth filter by means of forward-backward filtering. While this effectively zeroes out the time lag, the resulting filter is non-causal, which means it cannot be used online (i.e. in real time). However, for the present application, an offline filter is sufficient. The forward-backward filtering also introduces transients at both ends of the signal [48]. I therefore padded the reverted signal at both the start and the end of the array (identical to the 'even padtype' in Python's `scipy.signal.filtfilt()` function [49]). I used 3 as the order parameter and 0.04 as the low-pass cut-off frequency for each filter [38]. Note that due to the forward-backward approach both the filter's overall order and overall cut-off frequency are different from these values. The described approach is implemented in the spiro package [47], and can be used by calling, for example, the functions `spiro(data) |> spiro_max("0.04fz3")`.

2.2.3 Comparison of methods

To compare different processing methods within and between individuals, I choose to express the $\dot{V}O_{2max}$ normalized to a reference procedure. The reference procedure was chosen as being the most commonly applied in current literature as determined by the systematic review. Individual $\dot{V}O_{2max}$ values were expressed in reference to this procedure, where a value of 1 means that the processing method yields exactly the same $\dot{V}O_{2max}$ value as the reference method. I calculated the data for all integer parameter values within the range of the values found in the literature during the review. On a group level, I calculated the median and 10%- and 90%-quantiles of each processing strategy.

In an additional exploratory analysis I investigated the respiratory rate (number of breaths per minute) during the ramp tests. This may help to understand how breath-based and time-based data processing methods relate to each other.

3 Results

3.1 Systematic Scoping Review

Initial search yielded 7529 results of which 4364 remained after automated filtering and removal of duplicates (see flow diagram in Figure 2). Out of the random sample (n = 500), 242 articles were included in the final analysis. All sampled studies and their inclusion/exclusion status and reason are available on GitHub.

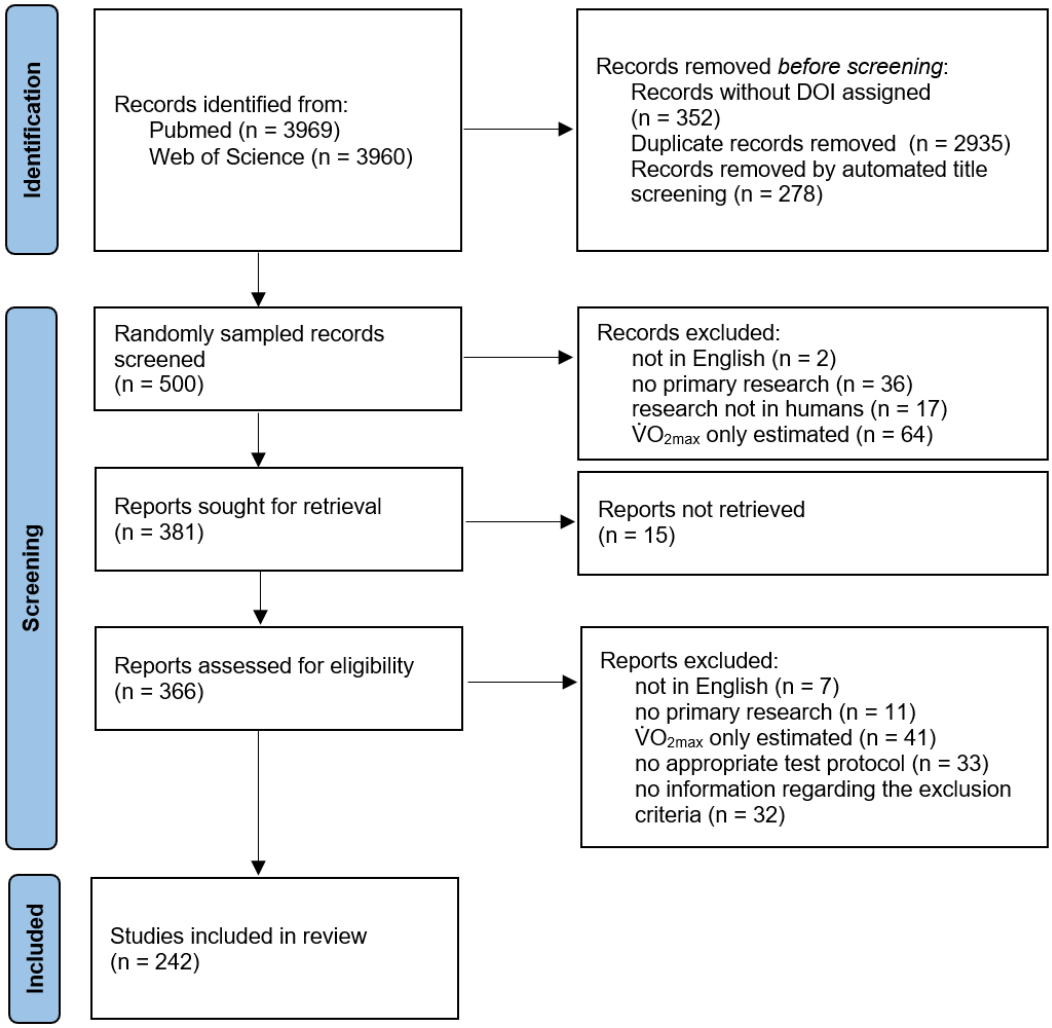

Figure 2: Flow diagram for the systematic scoping review in accordance with the PRISMA 2020 Statement [50]

Reporting practices of the methodology of gas exchange measures differed widely across the literature (see Table 3). Almost half (44.2%) of the articles did not report any information regarding their data processing strategy. One in twenty articles (5.8%) provided a rationale for their used strategy. Only a single article [51] reported information regarding all the investigated criteria.

Table 3: Percentage of studies that provided details on the different characteristics of oxygen uptake data processing.

Metabolic cart	Preprocessing	Software	Processing Strategy	Rationale
88.0%	5.6%*	15.0%*	55.8%	5.8%

*only examined within the subgroup of studies using breath-by-breath measurements

Out of the authors that provided information and collected breath-by-breath measurements, most (79.5%) utilized binned averages to determine $\dot{V}O_{2max}$. Moving time averages, or breath-based averages were uncommon (see Figure 3). No study used methods of digital filtering to determine $\dot{V}O_{2max}$.

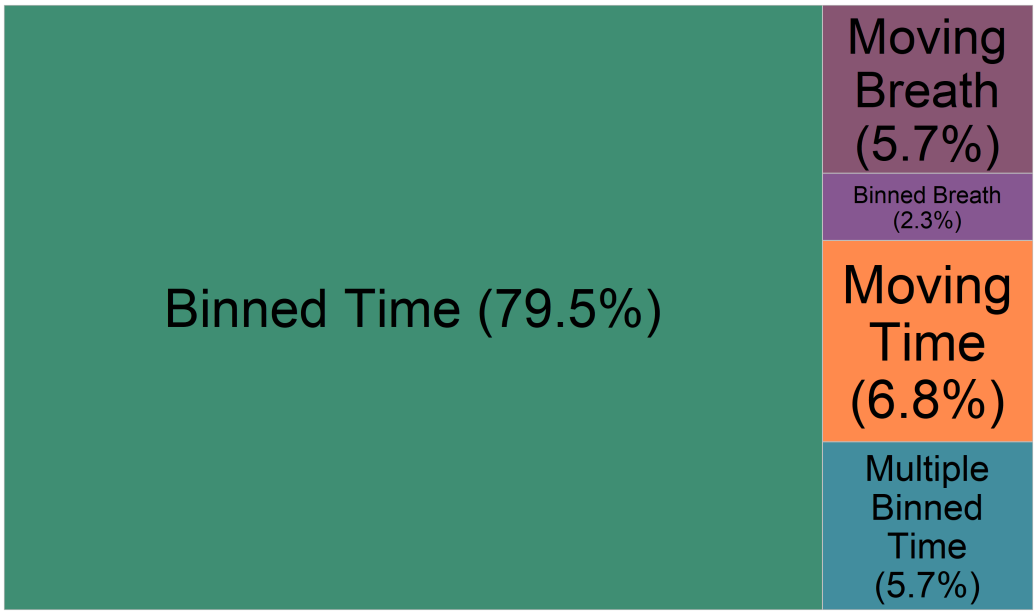

Figure 3: Data strategies for processing breath-by-breath data in the reviewed literature (n = 88).

For preprocessing, some authors reported the use of a (linear) interpolation for the breath-by-breath data to seconds (n = 7; 4.3%). A minority of researchers reported the use of data filtering strategies to remove outliers. This included the use of initial data smoothing by a short moving average (3 seconds, n = 1; 5 breaths, n = 3), the manual detection and removal of outliers (n = 2) or an automated removal of outliers (n = 5). For the automated outlier detection authors removed single data points differing from a not further defined local mean by a varying number of standard deviations (2, 3 or 4) or being outside of a 95% confidence interval. When reported, the software used for data processing varied among studies showing a total of more than 15 reported programs (for 30 studies that reported this parameter).

The calculation intervals for time-based averages of mixing chamber and breath-by-breath devices ranged from 5 to 60 seconds (see Figure 4). 30 second length intervals were most common to define $\dot{V}O_{2max}$, while authors also often employed shorter (10-20 s) and longer

(60 s) periods.

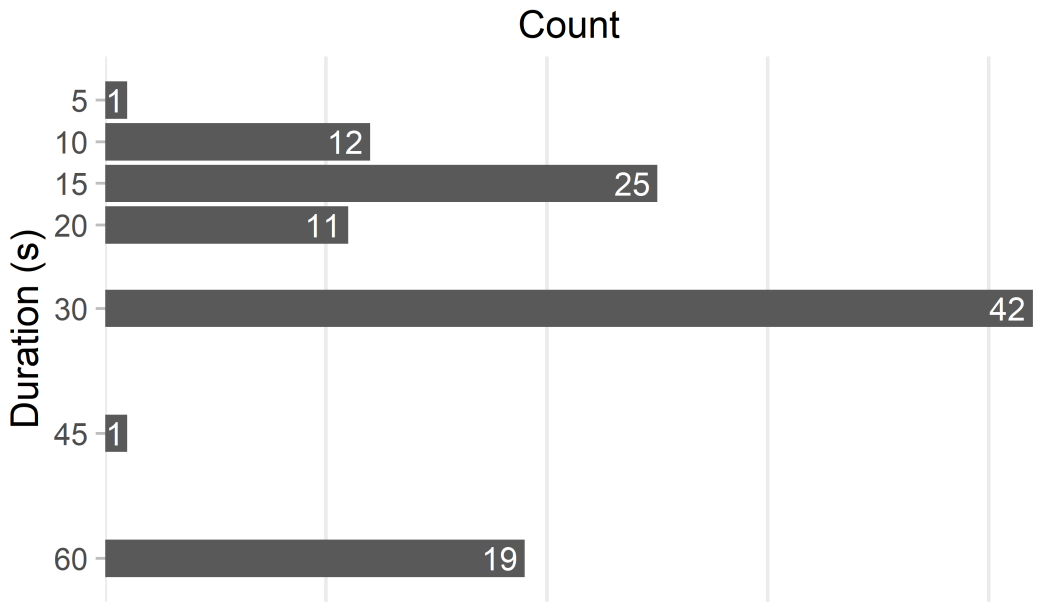

Figure 4: Total durations of the calculation interval of $\dot{V}O_{2max}$ in the reviewed studies.

3.2 Experimental Comparison

The average $\dot{V}O_{2max}$ as determined by a binned 30-second average was 62.2±6.3 $ml \cdot min^{-1} \cdot kg^{-1}$ (mean ± standard deviation). Applying different data processing strategies for $\dot{V}O_{2max}$ determination leads to different outcome values (see Figure 5).

Binned time averages systematically generate lower $\dot{V}O_{2max}$ values compared to moving averages. Decreasing the calculation interval to 5 seconds leads to a 3-4% median increase of $\dot{V}O_{2max}$ values. Notably on an individual level these increases may be lower (≈2%) or much higher (>10%) than the median value.

Moving time and moving breath averages yield nearly identical values for $\dot{V}O_{2max}$ over all calculation intervals. This corresponds to the trained athletes reaching respiratory rates around 60 min^{-1} in the final minute of the exercise test (see Figure 6). Using a zero-phase forward-backward filter (third order, 0.04 Hz cut-off frequency) lead to $\dot{V}O_{2max}$ values 0.4% [-0.2%; +1.4%] (median, 10th and 90th quantile) higher than the 30-second binned time average approach.

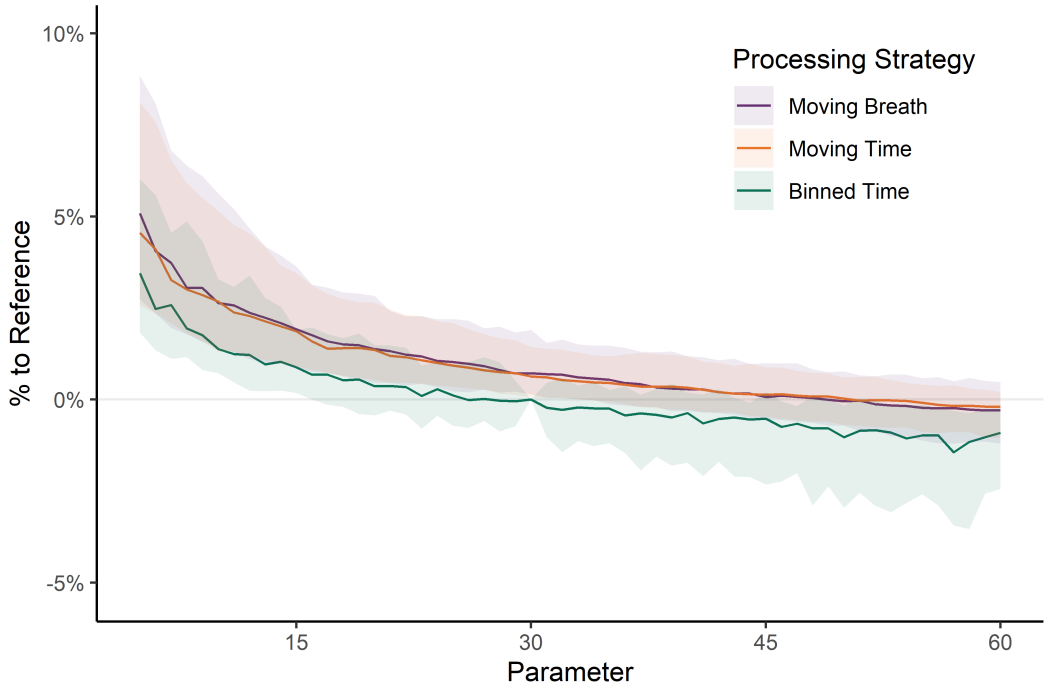

Figure 5: $\dot{V}O_{2max}$ varies by data processing strategy. Values are expressed relative to the $\dot{V}O_{2max}$ from a 30-second binned average — the most common strategy as determined by the review. Solid lines display the median, the shaded area marks the interval between 10th and 90th percentile. Using moving average leads to systematically higher $\dot{V}O_{2max}$ values compared to binned time averages. Changing the averaging interval (in seconds or breaths) can lead to median changes in $\dot{V}O_{2max}$ as large as 5%.

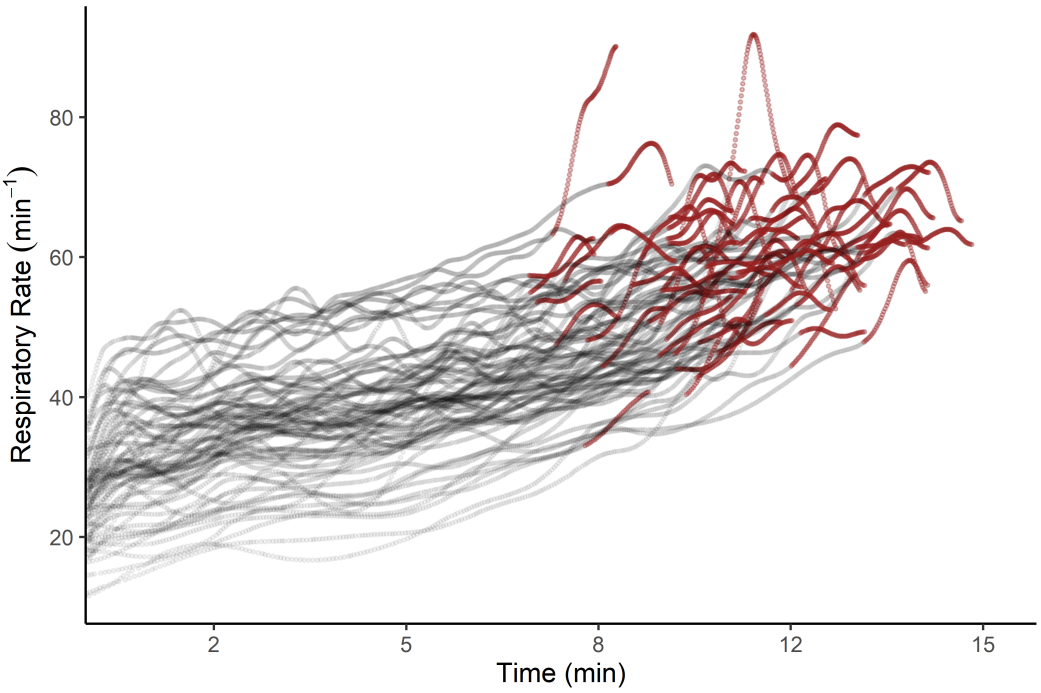

Figure 6: Respiratory rates peak around 60 min^{-1} in the ramp tests. The red segments correspond to the last minute before exhaustion of each individual (n = 72).

4 Discussion

The collected data shows, that current research uses a variety of strategies to determine $\dot{V}O_{2max}$, which directly influences the values obtained. Many articles only incompletely report on their methods and use outdated strategies. These practices hinder the validity and reproducibility of $\dot{V}O_{2max}$ measurement.

4.1 Current state of data processing

Despite calls to use moving averages [36,38], binned time averages remain the most commonly used data processing strategy to determine $\dot{V}O_{2max}$ in the reviewed literature (see Figure 3). These numbers are generally in line with the findings of the non-systematic search by Midgley et al. [27] and the survey by Robergs et al. [38]. It is somewhat surprising, that practices have not changed in recent years despite the publication of recommendations, that discourage researchers from using binned averages [38]. Using binned time averages leads to systematically lower $\dot{V}O_{2max}$ values as compared to moving averages (see Figure 5). The peak in oxygen uptake may fall in between two averaging intervals, resulting in an underestimation of $\dot{V}O_{2max}$ (see Figure 1 for an example). Binned time average revoke the most important argument for measuring breath-by-breath: the high temporal resolution of data. Despite these arguments speaking against the use of binned time averages, my review demonstrates that they remain extremely common in the scientific literature.

Breath-based averages seem to be more common (≈ 8%) than reported previously (< 1%) [27], but less common than assessed in self-reporting (≈ 17%) [38]. The increasing proportion of breath-based averages may be explained by publications in the recent years advocating for their use [36,38]. However the proportion of breath-based averages is still much smaller than that of binned time averages, which have not been recommended in such a way.

The length of the calculation interval for averaging is highly diverse within the literature (see Figure 4). This may reflect contradictory recommendations [19,36]. As long as there is no consensus on the arguments speaking for shorter or longer calculation intervals, there appears to be no optimal interval duration. As different interval durations can heavily influence $\dot{V}O_{2max}$ by more than 5% on a median level (see Figure 5), the exact reporting of the data processing strategy remains essential for interpretation.

The exact reasons for exercise scientist to ignore most recommendations by using binned time averages remain unknown. Some researchers may simply not be aware of the impact different data processing strategies have on the $\dot{V}O_{2max}$. But past publications on this issue have been widely cited [23,27,38] and should be known to most scientists in this field. Researchers may also use binned averages for traditional reasons. Douglas bags, as well as many mixing chamber devices, measure the oxygen uptake over fixed time intervals, producing data apparently similar to those by a binned time average of breath-by-breath data. But comparability with older data should only be an issue when using data acquired by different measurement methods within one analysis. Current studies using new breath-by-breath data do not reasonably need to rely on outdated methods of data processing.

A major source for choosing suboptimal processing strategies may be limitations by analysis software. My review shows that most researchers use the software of the metabolic cart's manufacturer to analyse the gas exchange data. These software may by default output binned time averages instead of raw breath-by-breath data. Moreover, further processing (e.g., interpolation, moving averages) may require the use of additional software. This may also explain why digital filtering has — despite recommended by Robergs et al. [38] — has not been used in a single study reviewed here: standard distributions of common data analysis software (e.g., Microsoft Excel) lack the capability to perform such operations. The common malpractices of data processing are likely to be both attributed to a lack of awareness and a lack of easy-to-use software solutions.

4.2 Impact of different data processing

The different data processing strategies found in the literature directly bias the $\dot{V}O_{2max}$ determination (see Figure 5), and as such can influence the classification of individuals, the evaluation of training success, and the assessment of $\dot{V}O_{2max}$ attainment. In accordance with previous findings [23,25,26] — and pure logic — longer calculation intervals lead to lower $\dot{V}O_{2max}$ values (see Figure 5). The analysed data shows median differences up to 5-7%

between processing strategies, which is in accordance with previous research [28]. Some studies reported even higher mean differences of up to 20% [35], but only when including raw breath-by-breath data in the comparison. The evaluation of unprocessed raw data for its maximum is highly erroneous (see the individual breath data points in Figure 1) and as such is not performed in research (see Figure 4); so there is no reason to include it in a comparison to other strategies. While previous research was often conducted in sedentary or recreational individuals, this thesis now presents evidence, that a similar effect of data processing strategies on $\dot{V}O_{2max}$ exists in well-trained athletes.

Binned time averages lead to systematically lower $\dot{V}O_{2max}$ values compared with moving averages, for the reasons explained above. While this general trend has been acknowledged previously [28], it has not been quantified. The data presented in this thesis suggest a ≈-1% lower median $\dot{V}O_{2max}$ when using binned averages compared with moving averages of the same calculation interval length. This difference is well within the measurement error of most if not all metabolic carts, but it is systematic and as such may bias the evaluation in scenarios where small changes in $\dot{V}O_{2max}$ are important (e.g. in elite sports).

Moving time and moving breath averages lead to almost identical $\dot{V}O_{2max}$ values on a median level (Figure 5). This seems natural in that the athletes in this study reached respiratory rates around 60 min^{-1} (see Figure 6), resulting in equivalent time- and breath-based interval lengths. For an athletic population, $\dot{V}O_{2max}$ values obtained by moving time and moving breath average can approximately used interchangeably. Given that less trained individual display lower respiratory rates during exercise tests to exhaustion [52], this finding will not translate to a sedentary population.

The exact impact of data processing strategies on the $\dot{V}O_{2max}$ is highly individual. Most research did present only comparisons of mean values, with results in accordance with those found here [28]. On an individual level, data processing strategies may impact the $\dot{V}O_{2max}$ in different severity. For example, for 10% of the investigated athletes a binned time average of 5 seconds leads to a $\dot{V}O_{2max}$ <3% higher than by a 30-second average, while for another 10% the $\dot{V}O_{2max}$ was >6% higher (see Figure 5). Current values reported and equation derived compare strategies on a group level [28], which improves comparability of group results for meta-analyses or group classifications. However on an individual level these can only be applied with a high margin of error. Changes in the impact of different data processing strategies on $\dot{V}O_{2max}$ range from 1-2% in some individuals to more than 10% in others. Hence when evaluating $\dot{V}O_{2max}$ data from different tests in a single individual obtained by using different processing methods, there is no way to accurately compare these values even when the processing strategies are reported. While the comparison of $\dot{V}O_{2max}$ from different processing strategies require their reporting for an sufficient analysis on a group level, on the individual level the raw data from each test is required.

4.3 Guidelines for data processing and reporting

This thesis focussed on the occurrence and impact of different data processing strategies, but did not investigate their validity in the first place. However, the results of the scoping review and experimental comparison allow to specify existing recommendation for data processing and set new guidelines for data reporting.

I highly disregard researchers from continuing to use binned time averages to determine $\dot{V}O_{2max}$. The main reason for such procedure is pure tradition, as it reduces the breath-by-breath data in an inappropriate way leading to a small, yet systematic, underestimation of $\dot{V}O_{2max}$. Moving time or moving breath averages are preferable to binned averages. In an athletic population they may be used equivalent, but in sedentary individuals breath-based averages will lead to lower $\dot{V}O_{2max}$ values than time-based averages with the same interval length parameter. While using a constant breath interval will lead to a similar degree of data smoothing regardless of the training status, using a constant time interval will lead to the same physiological time-frame for determining $\dot{V}O_{2max}$ (but with a different degree of data smoothing). From a data processing perspective, I prefer using breath-based moving averages, while keeping in mind, that this may underestimate true $\dot{V}O_{2max}$ in individuals with a low respiratory rate in the end of the exercise test.

Digital filtering seems to be a promising method, as it markedly reduces the variability moving averages have due to single data outliers (see Figure 1). I therefore agree with Robergs et al. [38], who recommend digital filters over any classical averaging procedure. The exact type of filter and values of filter parameters to be used have not yet been systematically investigated, and it is unclear whether it is possible at all to determine criteria for identifying an optimal filter. However, a Butterworth filter, as proposed by Robergs et al. [38], seems to produce reasonably smoothed data. Note, that a single Butterworth filter comprises a time lag, a fact highlighted by Weir et al. [39] that was not acknowledged by Robergs et al. [38]. To allow for a correct phasing of data, as well as for a correct $\dot{V}O_{2max}$ determination when the measurement is terminated shortly after reaching exhaustion, a zero-phase filter is needed. A zero-phase forward-backward Butterworth filter (as suggested by Weir et al. [39]) seems to produce reasonable smoothing when used with the parameters suggested by Robergs et al. [38] (low-pass cut-off frequency: 0.04 Hz; third order; see Figure 1). It should be noted, that despite using the same parameters the degree of filtering varies from that by Robergs et al. [38], as the filter is applied twice, which changes the overall magnitude of parameters. Other filtering parameters may be as reasonable as this, but it is unclear how to objectively investigate this issue.

I recommend to use a zero-phase forward-backward Butterworth filter on the raw breath-by-breath data. The $\dot{V}O_{2max}$ is then defined as the highest single filtered data point. A filter order of 3 and a low cut-off frequency of 0.04 Hz appear to be reasonable parameters for each filter. While such data processing has advantages over traditional data processing forms, it requires specialized software.

Table 4: Recommendations for reporting data processing strategies to determine the maximum oxygen uptake

Reporting Item	Description
Metabolic Cart	State the exact device model and manufacturer
Measurement Mode	State the measurement mode (e.g., mixing chamber, breath-by-breath,…)
Software	State the name and version of the software used for data analysis
Preprocessing	State if and how data underwent any initial modification (e.g., filtering of outliers, interpolation) before analysis
Processing Strategy	State the exact data processing strategy used to determine the $\dot{V}O_{2max}$ (e.g., binned time averages)
Processing Parameters	State the parameters used for the processing strategy (e.g., length of averaging interval)
Rationale	State the rationale for using the processing strategy (e.g., reference to recommendations)

To compare and evaluate $\dot{V}O_{2max}$ values from different studies, knowledge of the underlying data processing strategy is crucial. The review presented here shows that almost half of the studies that measured $\dot{V}O_{2max}$ did not describe their processing strategy. Other aspects of the data processing, such as outlier filtering or rationale for the chosen procedure were only in rare instances reported (see Table 3). Table 4 lists 7 item that should be reported to provide sufficient information on the data processing strategy used to determine $\dot{V}O_{2max}$. These items may be reported in form of a checklist, as an in-text enumeration or in sentence form. An example paragraph containing all the relevant information for the original data presented in this thesis [31,46] would be:

> *"We measured breath-by-breath data during the ramp tests with a ZAN 600 USB device (nSpire Health, Inc., Longmont, CO, United States of America). The raw data was analyzed without any previous filtering by using a low-pass forward-backward Butterworth filter (each filter: 3rd order, 0.04 Hz cut-off) implemented in the spiro Package for R [47]. This strategy produces similar results as that recommended by Robergs et al. [38], but does not produce a time lag."*

Note that the correct reporting of an exercise test to determine $\dot{V}O_{2max}$ requires more information than those on data processing. Further aspects to be reported include, but are not limited to, the study population, the exercise protocol, the device calibration, and criteria to stop the test. In cases where journals endorse word limits on articles, these reporting — including the reporting on data processing strategies — may be included in supplementary files. The correct and detailed reporting of data processing strategies, as well as other test characteristics, is crucial for interpreting presented $\dot{V}O_{2max}$ values.

The results of this thesis suggest that comprehensive reporting facilitates approximate com-

parisons of $\dot{V}O_{2max}$ data on a group level derived using different data processing strategies. But on an individual level and for a precise comparison, reporting is not enough, as differences between data processing strategies vary between individuals and are potentially influenced by training status. Sharing of the raw metabolic gas exchange data is the solution to this problem. It allows any researcher to recalculate the $\dot{V}O_{2max}$ using their preferred data processing strategy. Anonymization of raw gas exchange data files should not be a hinder sharing, since most of these files are structured in a simple way, which allows to easily remove any personal information (if this had not been done in the metabolic cart's system before). In terms of reproducibility of $\dot{V}O_{2max}$ determination, a way even superior to correct reporting and raw data sharing is to additionally share data analysis code. This allows any researcher to independently reproduce the $\dot{V}O_{2max}$ processing conducted within a study, but requires the data analysis to take place in a programming (or at least code generating) environment. Luckily such programs for the purpose of analyzing gas exchange data exist as free open-source software [47,53].

4.4 Limitations

Due to the sheer quantity of the publications investigating $\dot{V}O_{2max}$, it was not possible to perform an exhaustive review of the literature. The scoping review therefore relies on a random sample which not necessarily captures exact trends of the literature. However, efforts were made — such as random sampling and systematic article exclusions — to ensure the sample to be representative. Notable, almost half of the studies did not report their data processing strategy at all. The data processing strategies used in the literature could only be described when studies reported them.

Ambiguities in the reporting of the investigated studies may impact the analysis results. For example some studies using long binned averages (e.g., 60 seconds) may have in fact been using multiple binned averages of shorter durations (e.g., 4x15 second), without describing this correctly. Moreover the exact definitions for building binned averages varies within the literature: While most studies define the binning periods from the beginning of the exercise, some may define them from the endpoint. Additionally some studies reviewed did not define the maximum, but a preset binned average period as their $\dot{V}O_{2max}$ (for example the last bin, regardless of its value). In situations where the maximum in oxygen uptake is reached considerably before exhaustion (i.e., a long plateau in oxygen uptake exists) this may lead to different results than a traditional binned average processing. I did not separately consider such sub-categories of data processing strategies, as they may not be very common and are often hard to precisely investigate due to ambiguity in reporting.

This work treated each breath as the single data processing unit of cardiopulmonary exercise testing. However, metabolic carts sample gas fraction and gas flow data at a much higher frequency (e.g., 50 Hz). The data for each breath is subsequently calculated from these

raw signals by the means of an algorithm. Different algorithms to generate the breath-by-breath data can lead to different outcomes [54], and as such may also influence the $\dot{V}O_{2max}$. Hence documenting and reporting of the breath-by-breath algorithm is advisable. Yet many metabolic carts do not describe their default algorithm and limit access to the raw data signal.

The experimental comparison of different data processing strategies was conducted on a standardized data set of exercise tests. This standardization in terms of training status, exercise protocol, and measurement device helps to highlight the impact of different data processing even in a relatively homogeneous data set. However the results may only partly transfer to different settings, such as individuals with less training background. I did not conduct a formal analysis of the validity or reliability of different data processing strategies, so recommendations regarding their use rely on theoretical derivations and prior research. Since the reliability of different strategies appears to be similar [27,28] and currently no accepted methods to quantify the validity of $\dot{V}O_{2max}$ exist, the presented approach is sufficient to derive recommendations for data processing strategies to determine $\dot{V}O_{2max}$.

5 Conclusion

The determination of $\dot{V}O_{2max}$ from raw breath-by-breath data requires some form of data post-processing. As shown by the scoping review, researchers use a variety of strategies for this data processing. I demonstrated on experimental data, that different processing strategies systematically bias the $\dot{V}O_{2max}$ assessment, which can ultimately lead to misclassification of individuals or groups, or to incorrect evaluation of interventions. Despite contradictory recommendations, binned time average remain by far the most common data processing strategy. The calculation intervals for time averages range from 5 to 60 seconds in the reviewed literature, which heavily influences the $\dot{V}O_{2max}$ value obtained. Almost all research did only incompletely report on their data processing to determine $\dot{V}O_{2max}$, and as such can be regarded non-reproducible. To ensure valid and reproducible $\dot{V}O_{2max}$ assessments, researchers should at minimum report their processing strategy based on a set of guidelines I developed in this thesis. Despite calls for standardization of data processing, to date no evidence for an optimal strategy exists, but digital filtering (e.g., zero-phase Butterworth filter) appears to generate a more reasonable data smoothing compared to traditional (moving) averages. I encourage researchers to share their raw gas exchange data and use programming software for data analysis, to minimize the bias due to data processing when determining the $\dot{V}O_{2max}$.

References

1. Holloszy JO, Coyle EF. Adaptations of skeletal muscle to endurance exercise and their metabolic consequences. Journal of Applied Physiology. 1984;56:831–8. https://doi.org/10.1152/jappl.1984.56.4.831
2. Bassett DRJETH. Limiting factors for maximum oxygen uptake and determinants of endurance performance. Medicine & Science in Sports & Exercise. 2000;32:70. https://doi.org/10.1097/00005768-200001000-00012
3. Reaburn P, Dascombe B. Endurance performance in masters athletes. European Review of Aging and Physical Activity. 2008;5:31–42. https://doi.org/10.1007/s11556-008-0029-2
4. Costill DL, Thomason H, Roberts E. Fractional utilization of the aerobic capacity during distance running. Medicine & Science in Sports & Exercise. 1973;5:248252. https://doi.org/10.1249/00005768-197300540-00007
5. Tanaka K, Takeshima N, Kato T, Niihata S, Ueda K. Critical determinants of endurance performance in middle-aged and elderly endurance runners with heterogeneous training habits. European Journal of Applied Physiology and Occupational Physiology. 1990;59:443–9. https://doi.org/10.1007/BF02388626
6. Buchfuhrer MJ, Hansen JE, Robinson TE, Sue DY, Wasserman K, Whipp BJ. Optimizing the exercise protocol for cardiopulmonary assessment. Journal of Applied Physiology. 1983;55:1558–64. https://doi.org/10.1152/jappl.1983.55.5.1558
7. Yoon B-K, Kravitz L, Robergs R. V̇O2max, protocol duration, and the v̇O2 plateau. Medicine & Science in Sports & Exercise. 2007;39:11861192. https://doi.org/10.1249/mss.0b13e318054e304
8. Midgley AW, Bentley DJ, Luttikholt H, McNaughton LR, Millet GP. Challenging a Dogma of Exercise Physiology. Sports Medicine. 2008;38:441–7. https://doi.org/10.2165/00007256-200838060-00001
9. Midgley AW, Marchant DC, Levy AR. A call to action towards an evidence-based approach to using verbal encouragement during maximal exercise testing. Clinical Physiology and Functional Imaging. 2018;38:547–53. https://doi.org/10.1111/cpf.12454
10. Myers J, Buchanan N, Walsh D, Kraemer M, McAuley P, Hamilton-Wessler M, et al. Comparison of the ramp versus standard exercise protocols. Journal of the American College of Cardiology. 1991;17:1334–42. https://doi.org/10.1016/S0735-1097(10)80144-5
11. Katch VL, Sady SS, Freedson P. Biological variability in maximum aerobic power. Medicine & Science in Sports & Exercise. 1982;14:2125. https://doi.org/10.1249/00005768-198201000-00004

12. Shephard RJ. Open-circuit respirometry: a brief historical review of the use of Douglas bags and chemical analyzers. European Journal of Applied Physiology. 2017;117:381–7. https://doi.org/10.1007/s00421-017-3556-6

13. Macfarlane DJ. Automated Metabolic Gas Analysis Systems. Sports Medicine. 2001;31:841–61. https://doi.org/10.2165/00007256-200131120-00002

14. Beijst C, Schep G, Breda E van, Wijn PFF, Pul C van. Accuracy and precision of CPET equipment: A comparison of breath-by-breath and mixing chamber systems. Journal of Medical Engineering & Technology. 2013;37:35–42. https://doi.org/10.3109/03091902.2012.733057

15. Roecker K, Prettin S, Sorichter S. Gas Exchange Measurements with High Temporal Resolution: The Breath-by-Breath Approach. International Journal of Sports Medicine. 2005;26:S11–8. https://doi.org/10.1055/s-2004-830506

16. Winkert K, Kirsten J, Kamnig R, Steinacker JM, Treff G. Differences in V̇O2max Measurements Between Breath-by-Breath and Mixing-Chamber Mode in the COSMED K5. International Journal of Sports Physiology and Performance. 2021;16:1335–40. https://doi.org/10.1123/ijspp.2020-0634

17. Miles DS, Cox MH, Verde TJ. Four commonly utilized metabolic systems fail to produce similar results during submaximal and maximal exercise. Sports Medicine, Training and Rehabilitation. 1994;5:189–98. https://doi.org/10.1080/15438629409512016

18. Taylor HL, Buskirk E, Henschel A. Maximal oxygen intake as an objective measure of cardio-respiratory performance. Journal of Applied Physiology. 1955;8:73–80. https://doi.org/10.1152/jappl.1955.8.1.73

19. Howley ET, Bassett DR, Welch HG. Criteria for maximal oxygen uptake: Review and commentary. Medicine & Science in Sports & Exercise. 1995;27:12921301. https://journals.lww.com/acsm-msse/Abstract/1995/09000/Criteria_for_maximal_oxygen_uptake__review_and.9.aspx

20. Poole DC, Wilkerson DP, Jones AM. Validity of criteria for establishing maximal O2 uptake during ramp exercise tests. European Journal of Applied Physiology. 2008;102:403–10. https://doi.org/10.1007/s00421-007-0596-3

21. Poole DC, Jones AM. Measurement of the maximum oxygen uptake v̇o2max: V̇o2peak is no longer acceptable. Journal of Applied Physiology. 2017;122:997–1002. https://doi.org/10.1152/japplphysiol.01063.2016

22. Green S, Askew C. V̇o2peak is an acceptable estimate of cardiorespiratory fitness but not v̇o2max. Journal of Applied Physiology. 2018;125:229–32. https://doi.org/10.1152/japplphysiol.00850.2017

23. Astorino TA. Alterations in VO2max and the VO2 plateau with manipulation of sampling interval. Clinical Physiology and Functional Imaging. 2009;29:60–7. https://doi.org/10.1111/j.1475-097X.2008.00835.x

24. Matthews JI, Bush BA, Morales FM. Microprocessor Exercise Physiology Systems vs a Nonautomated System: A Comparison of Data Output. Chest. 1987;92:696–703. https://doi.org/10.1378/chest.92.4.696

25. Johnson JS, Carlson JJ, VanderLaan RL, Langholz DE. Effects of sampling interval on peak oxygen consumption in patients evaluated for heart transplantation. CHEST. 1998;113:816–9. https://doi.org/10.1378/chest.113.3.816

26. Sell KM, Ghigiarelli JJ, Prendergast JM, Ciani GJ, Martin J, Gonzalez AM. Comparison of V̇o2peak and V̇o2max at Different Sampling Intervals in Collegiate Wrestlers. Journal of Strength and Conditioning Research. 2021;35:2915–7. https://doi.org/10.1519/JSC.0000000000003887

27. Midgley AW, McNaughton LR, Carroll S. Effect of the o2 time-averaging interval on the reproducibility of o2max in healthy athletic subjects. Clinical Physiology and Functional Imaging. 2007;27:122–5. https://doi.org/10.1111/j.1475-097X.2007.00725.x

28. Martin-Rincon M, González-Henríquez JJ, Losa-Reyna J, Perez-Suarez I, Ponce-González JG, La Calle-Herrero J de, et al. Impact of data averaging strategies on V̇O2max assessment: Mathematical modeling and reliability. Scandinavian Journal of Medicine & Science in Sports. 2019;29:1473–88. https://doi.org/10.1111/sms.13495

29. Hill DW, Stephens LP, Blumoff-Ross SA, Poole DC, Smith JC. Effect of sampling strategy on measures of V̇O2peak obtained using commercial breath-by-breath systems. European Journal of Applied Physiology. 2003;89:564–9. https://doi.org/10.1007/s00421-003-0843-1

30. Scheadler CM, Garver MJ, Hanson NJ. The gas sampling interval effect on v̇O2peak is independent of exercise protocol. Medicine & Science in Sports & Exercise. 2017;49:19111916. https://doi.org/10.1249/MSS.0000000000001301

31. Quittmann OJ, Schwarz YM, Nolte S, Fuchs M, Gehlert G, Slowig Y, et al. Relationship between physiological parameters and time trial performance over 1, 2 and 3 km in trained runners.

32. Pauw KD, Roelands B, Cheung SS, Geus B de, Rietjens G, Meeusen R. Guidelines to Classify Subject Groups in Sport-Science Research. International Journal of Sports Physiology and Performance. 2013;8:111–22. https://doi.org/10.1123/ijspp.8.2.111

33. Decroix L, Pauw KD, Foster C, Meeusen R. Guidelines to Classify Female Subject Groups in Sport-Science Research. International Journal of Sports Physiology and Performance. 2016;11:204–13. https://doi.org/10.1123/ijspp.2015-0153

34. Mancini DM, Eisen H, Kussmaul W, Mull R, Edmunds LH, Wilson JR. Value of peak exercise oxygen consumption for optimal timing of cardiac transplantation in ambulatory patients with heart failure. Circulation. 1991;83:778–86. https://doi.org/10.1161/01.cir.83.3.778

35. Myers J, Walsh D, Sullivan M, Froelicher V. Effect of sampling on variability and plateau in oxygen uptake. Journal of Applied Physiology. 1990;68:404–10. https://doi.org/10.1152/jappl.1990.68.1.404

36. Robergs RA, Burnett A. Methods used to process data from indirect calorimetry and their application to VO2max. Journal of Exercise Physiology Online. 2003;6:44–57. http://connection.ebscohost.com/c/articles/21794702/methods-used-process-data-from-indirect-calorimetry-their-application-vo2max

37. ATS/ACCP. ATS/ACCP statement on cardiopulmonary exercise testing. American Journal of Respiratory and Critical Care Medicine. 2003;167:211–77. https://doi.org/10.1164/rccm.167.2.211

38. Robergs RA, Dwyer D, Astorino T. Recommendations for Improved Data Processing from Expired Gas Analysis Indirect Calorimetry. Sports Medicine. 2010;40:95–111. https://doi.org/10.2165/11319670-000000000-00000

39. Weir J, Koerner S, Mack B, Masek J, Vanderhoff D, Heiderscheit B. VO2 plateau detection in cycle ergometry. Journal of Exercise Physiology Online. 2004;7:55–62.

40. Foster ED, Deardorff A. Open Science Framework (OSF). Journal of the Medical Library Association. 2017;105:203–6. https://doi.org/10.5195/jmla.2017.88

41. Akker O van den, Peters G-J, Bakker C, Carlsson R, Coles NA, Corker KS, et al. Inclusive systematic review registration form. https://doi.org/10.31222/osf.io/3nbea

42. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2022. https://www.r-project.org/

43. RStudio Team. RStudio: Integrated development environment for r. Boston, MA: RStudio, PBC; 2022. http://www.rstudio.com/

44. Allaire JJ, Teague C, Scheidegger C, Xie Y, Dervieux C. Quarto. 2022. https://doi.org/10.5281/zenodo.5960048

45. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. Annals of Internal Medicine. 2018;169:467–73. https://doi.org/10.7326/M18-0850

46. Quittmann OJ, Foitschik T, Vafa R, Freitag F, Spearmann N, Nolte S, et al. Augmenting the metabolic profile in endurance running by maximal lactate accumulation rate.

47. Nolte S. Spiro: Manage data from cardiopulmonary exercise testing. 2022. https://doi.org/10.5281/zenodo.5816170

48. Gustafsson F. Determining the initial states in forward-backward filtering. IEEE Transactions on Signal Processing. 1996;44:988–92. https://doi.org/10.1109/78.492552

49. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods. 2020;17:261–72. https://doi.org/10.1038/s41592-019-0686-2

50. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71. https://doi.org/10.1136/bmj.n71

51. Maturana FM, Schellhorn P, Erz G, Burgstahler C, Widmann M, Munz B, et al. Individual cardiovascular responsiveness to work-matched exercise within the moderate- and severe-intensity domains. European Journal of Applied Physiology. 2021;121:2039–59. https://doi.org/10.1007/s00421-021-04676-7

52. 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;100:136–42. https://doi.org/10.1378/chest.100.1.136

53. Maturana FM. Whippr: Tools for manipulating gas exchange data. 2022. https://github.com/fmmattioni/whippr

54. Koschate J, Cettolo V, Hoffmann U, Francescato MP. Breath-by-breath oxygen uptake during running: Effects of different calculation algorithms. Experimental Physiology. 2019;104:1829–40. https://doi.org/10.1113/EP087916

A Appendix

A.1 Transparent Changes

This document includes all deviation and modifications of code and methods in the final project compared to its preregistration.

A.1.1 Major Changes

A.1.1.1 Number of exercise tests for comparison

Due to a miscalculation, the preregistration provided an incorrect number of exercise test (n = 76). The correct number of exercise tests is n = 72, with 44 from [46] (one test only partly included in the original work due to missing other data) and 28 from [31] (three test only partly included in the original work due to missing other data).

A.1.1.2 Additional variables extracted from included research

`outcome`: Which type of outcome VO2max is. Either `primary`, `secondary` or `other`

`source`: Which source is provided for the data processing method used.

A.1.2 Minor Changes

A.1.2.1 Code changes for automated article filtering and screening preparation

- advanced detection of missing DOIs: `is.na(merge_data$doi) | (merge_data$doi == "")` instead of `is.na(merge_data$doi)`.
- Improved function to retrieve missing PubMed abstracts: Handles case when input (PMID) is missing (`if (is.na(pmid)) return(NA)`).
- save/load of the sampling results as an .Rda file to reduce computation time when working on parts of the workflow.

A.1.2.2 Unblinding of single abstracts

- Manual retrieval of abstracts for articles, as these were neither present in the search result data, nor could be automatically scraped. Abstracts were saved and imported as .txt files. Temporary unblinding only applied to the primary researcher. This concerns the abstract with the Sampling ID (sid): 50, 238, 288, 416, 488, 490, 500

- Manual retrieval of abstracts for articles, as the automatically collected abstract contained html-tags that could not be removed for later abstract plots. Abstracts were saved and imported as .txt files. Temporary unblinding only applied to the primary researcher. This concerns the abstract with the Sampling ID (sid): 344, 356
- Unblinding during screening to assess the implications of title given in squared brackets. This concerns the abstract with the Sampling ID (sid): 262, 303
- Consulting of online abstract due to incomplete abstract plot. This concerns the abstract with the Sampling ID (sid): 275

A.1.2.3 Minor Modification of exclusion criteria

Changes are in italics:

`'r'`: Is the article no original research *(i.e. no primary analysis of experimental data)* ? (if yes, indicate ‘r’; if no, continue)

`'t'`: Was no full-text ~~available~~ *accessible* for the corresponding article? (if yes, indicate ‘t’, if no continue)

A.1.2.4 Minor screening error for two articles

For two articles (sampling ids: 194, 282) I only realized during data extraction that they matched the exclusion criteria (‘c‘: no continuous measurement of oxygen uptake). In agreement of both screeners, the screening data was retrospectively changed.

A.2 Technical Details

A.2.1 Session Info

```
sessionInfo()
```

```
R version 4.2.0 (2022-04-22 ucrt)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 10 x64 (build 22000)

Matrix products: default

locale:
[1] LC_COLLATE=German_Germany.utf8  LC_CTYPE=German_Germany.utf8
[3] LC_MONETARY=German_Germany.utf8 LC_NUMERIC=C
[5] LC_TIME=German_Germany.utf8

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

loaded via a namespace (and not attached):
 [1] compiler_4.2.0    pillar_1.7.0      tools_4.2.0       digest_0.6.29
 [5] tibble_3.1.7      jsonlite_1.8.0    evaluate_0.15     lifecycle_1.0.1
 [9] viridisLite_0.4.0 pkgconfig_2.0.3   rlang_1.0.2       cli_3.3.0
[13] rstudioapi_0.13   yaml_2.3.5        xfun_0.31         fastmap_1.1.0
[17] kableExtra_1.3.4  httr_1.4.3        stringr_1.4.0     dplyr_1.0.9
[21] xml2_1.3.3        knitr_1.39        generics_0.1.2    vctrs_0.4.1
[25] systemfonts_1.0.4 tidyselect_1.1.2  rprojroot_2.0.3   webshot_0.5.3
[29] svglite_2.1.0     glue_1.6.2        here_1.0.1        R6_2.5.1
[33] fansi_1.0.3       rmarkdown_2.14    purrr_0.3.4       magrittr_2.0.3
[37] scales_1.2.0      htmltools_0.5.2   ellipsis_0.3.2    rvest_1.0.2
[41] colorspace_2.0-3  utf8_1.2.2        stringi_1.7.6     munsell_0.5.0
[45] crayon_1.5.1
```

A.2.2 Packages

```
p_used <- unique(renv::dependencies(path = "../")$Package)
```

```
Finding R package dependencies ... Done!
```

```r
p_inst <- as.data.frame(installed.packages())
out <- p_inst[p_inst$Package %in% p_used, c("Package", "Version")]
rownames(out) <- NULL
out
```

```
      Package  Version
1  colorspace    2.0-3
2       dplyr    1.0.9
3     ggplot2    3.3.6
4      ggtext    0.1.1
5        here    1.0.1
6  kableExtra    1.3.4
7       knitr     1.39
8   MetBrewer    0.2.0
9       purrr    0.3.4
10     readxl    1.4.0
11    rentrez    1.2.3
12       renv   0.15.4
13  rmarkdown     2.14
14     scales    1.2.0
15      shiny    1.7.1
16      spiro    0.0.4
17    stringr    1.4.0
18      tidyr    1.2.0
19 treemapify    2.5.5
20        XML 3.99-0.9
21       grid    4.2.0
```

A.3 Prisma Reporting Checklist

Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) Checklist

SECTION	ITEM	PRISMA-ScR CHECKLIST ITEM	REPORTED IN SECTION
TITLE			
Title	1	Identify the report as a scoping review.	Title
ABSTRACT			
Structured summary	2	Provide a structured summary that includes (as applicable): background, objectives, eligibility criteria, sources of evidence, charting methods, results, and conclusions that relate to the review questions and objectives.	Zusammenfassung (German abstract)
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known. Explain why the review questions/objectives lend themselves to a scoping review approach.	1.3 Aim
Objectives	4	Provide an explicit statement of the questions and objectives being addressed with reference to their key elements (e.g., population or participants, concepts, and context) or other relevant key elements used to conceptualize the review questions and/or objectives.	1.3 Aim
METHODS			
Protocol and registration	5	Indicate whether a review protocol exists; state if and where it can be accessed (e.g., a Web address); and if available, provide registration information, including the registration number.	2 Methods
Eligibility criteria	6	Specify characteristics of the sources of evidence used as eligibility criteria (e.g., years considered, language, and publication status), and provide a rationale.	2.1.1 Search & Screening
Information sources*	7	Describe all information sources in the search (e.g., databases with dates of coverage and contact with authors to identify additional sources), as well as the date the most recent search was executed.	2.1.1 Search & Screening
Search	8	Present the full electronic search strategy for at least 1 database, including any limits used, such that it could be repeated.	Table 1
Selection of sources of evidence†	9	State the process for selecting sources of evidence (i.e., screening and eligibility) included in the scoping review.	2.1.1 Search & Screening; Table 2
Data charting process‡	10	Describe the methods of charting data from the included sources of evidence (e.g., calibrated forms or forms that have been tested by the team before their use, and whether data charting was done independently or in duplicate) and any processes for obtaining and confirming data from investigators.	2.1.2 Data Extraction
Data items	11	List and define all variables for which data were sought and any assumptions and simplifications made.	2.1.2 Data Extraction
Critical appraisal	12	If done, provide a rationale for conducting a	N/A

SECTION	ITEM	PRISMA-ScR CHECKLIST ITEM	REPORTED IN SECTION
of individual sources of evidence§		critical appraisal of included sources of evidence; describe the methods used and how this information was used in any data synthesis (if appropriate).	
Synthesis of results	13	Describe the methods of handling and summarizing the data that were charted.	2.1.3 Data Synthesis
RESULTS			
Selection of sources of evidence	14	Give numbers of sources of evidence screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally using a flow diagram.	Figure 2
Characteristics of sources of evidence	15	For each source of evidence, present characteristics for which data were charted and provide the citations.	3.1 Systematic Scoping Review; online file
Critical appraisal within sources of evidence	16	If done, present data on critical appraisal of included sources of evidence (see item 12).	N/A
Results of individual sources of evidence	17	For each included source of evidence, present the relevant data that were charted that relate to the review questions and objectives.	3.1 Systematic Scoping Review; online file
Synthesis of results	18	Summarize and/or present the charting results as they relate to the review questions and objectives.	3.1 Systematic Scoping Review; Table 3; Figure 3-4
DISCUSSION			
Summary of evidence	19	Summarize the main results (including an overview of concepts, themes, and types of evidence available), link to the review questions and objectives, and consider the relevance to key groups.	4.1 Current State of Data Processing
Limitations	20	Discuss the limitations of the scoping review process.	4.3 Limitations
Conclusions	21	Provide a general interpretation of the results with respect to the review questions and objectives, as well as potential implications and/or next steps.	4.1 Current State of Data Processing
FUNDING			
Funding	22	Describe sources of funding for the included sources of evidence, as well as sources of funding for the scoping review. Describe the role of the funders of the scoping review.	*No funding was received for this scoping review*

JBI = Joanna Briggs Institute; PRISMA-ScR = Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews.
* Where *sources of evidence* (see second footnote) are compiled from, such as bibliographic databases, social media platforms, and Web sites.
† A more inclusive/heterogeneous term used to account for the different types of evidence or data sources (e.g., quantitative and/or qualitative research, expert opinion, and policy documents) that may be eligible in a scoping review as opposed to only studies. This is not to be confused with *information sources* (see first footnote).
‡ The frameworks by Arksey and O'Malley (6) and Levac and colleagues (7) and the JBI guidance (4, 5) refer to the process of data extraction in a scoping review as data charting.
§ The process of systematically examining research evidence to assess its validity, results, and relevance before using it to inform a decision. This term is used for items 12 and 19 instead of "risk of bias" (which is more applicable to systematic reviews of interventions) to include and acknowledge the various sources of evidence that may be used in a scoping review (e.g., quantitative and/or qualitative research, expert opinion, and policy document).

From: Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMAScR): Checklist and Explanation. Ann Intern Med. 2018;169:467–473. doi: 10.7326/M18-0850.

A.4 Blinded Abstract Example

Development in Adolescent Middle-Distance Athletes: A Study of Training Loadings, Physical Qualities, and Competition Performance

Sampling ID: 030

Jones, TW, Shillabeer, BC, Ryu, JH, and Cardinale, M. Development in adolescent middle-distance athletes: a study of training loadings, physical qualities, and competition performance. J Strength Cond Res 35(12S): S103-S110, 2021-The purpose of this study was to examine changes in running performance and physical qualities related to middle-distance performance over a training season. The study also examined relationships between training loading and changes in physical qualities as assessed by laboratory and field measures. Relationships between laboratory and field measures were also analyzed. This was a 9-month observational study of 10 highly trained adolescent middle-distance athletes. Training intensity distribution was similar over the observational period, whereas accumulated and mean distance and training time and accumulated load varied monthly. Statistically significant ($p < 0.05$) and large effect sizes (Cohen's d) (0.80) were observed for improvements in: body mass (5.6%), 600-m (4.6%), 1,200-m (8.7%), and 1,800-m (6.1%) time trial performance, critical speed (7.1%), V?o(2)max (5.5%), running economy (10.1%), vertical stiffness (2.6%), reactive index (3.8%), and countermovement jump power output relative to body mass (7.9%). Improvements in 1,800 m TT performance were correlated with increases in V?o(2)max ($r = 0.810$, $p = 0.015$) and critical speed ($r = 0.918$, $p = 0.001$). Increases in V?o(2)max and critical speed were also correlated ($r = 0.895$, $p = 0.003$). Data presented here indicate that improvements in critical speed may be reflective of changes in aerobic capacity in adolescent middle-distance athletes.

ID: 130; PMID = 31809463