Sports Science Exchange (2024) Vol. 37, No. 250, 1- 6

GATORADE SPORTS SCIENCE INSTITUTE

MAKING SENSE OF WEARABLES DATA

Marco Altini, PhD
HRV4Training

KEY POINTS

- Focus on wearable measurements under most circumstances, as they are directly captured by the wearable's sensors, while estimates are attempts to derive something that cannot be measured with the sensors available on the wearable device. Recognize that both measurements and estimates can have larger errors in certain contexts, as when there is movement.
- Focus on wearable physiological responses as opposed to made-up scores combining physiology and behavior. The emphasis should be on the body's physiological response rather than penalizing scores for changes in behavior or external factors. Behavior and external factors remain key as context.
- There is no objective quantification or reference system for many made-up scores. There's no objective way to quantify metrics like sleep quality, readiness, recovery or stress, and wearables may oversimplify physiological responses, lacking necessary context. Be skeptical about these.
- Start with a plan: Before interpreting wearable data, establish a plan, and use measurements (e.g., resting physiology) to capture responses to the plan, potentially making adjustments. Relying solely on made-up scores may otherwise lead to a reactive approach, responding to acute changes without a long-term focus.

INTRODUCTION

The wearables market is increasing day by day. Millions of devices are currently used by athletes as well as by the general population to track various metrics typically concerning sleep, physical activity and recovery. In this context, it is frequently observed that individuals either wholeheartedly embrace a wearable device or dismiss it entirely due to, for example, an actual or perceived inaccuracy in a provided metric. On one side, we might have sponsored athletes and somewhat over-enthusiast consumers, while on the other side, skeptical coaches and/or scientists.

Regrettably, a more nuanced approach is necessary when it comes to these devices and their utilization. In this Sports Science Exchange (SSE) article, the aim is to establish a framework that can enable the reader to derive more meaningful insights from the data, focusing on the parameters and situations that are underpinned by more solid scientific evidence. There is significant value in utilizing these devices for heightened awareness and actionable insights, but we must move beyond the simplistic viewpoints mentioned above.

To start with, it is crucial to distinguish the parameters provided by a wearable device into what exactly is being measured (and when) and what is only being estimated (and how). Once these distinctions are grasped, we can allocate our time and energy to aspects that are likely more dependable and beneficial.

PROPOSED FRAMEWORK

A wearable device typically provides various metrics, such as heart rate (Nelson et al., 2020), heart rate variability (Georgiou et al., 2018), temperature (Maijala et al., 2019), oxygen saturation (SPO2, Spaccarotella et al., 2022), calories burned (Fuller et al., 2020), stress level, readiness or recovery scores (Ibrahim et al., 2024), sleep stages (Lee et al., 2018), sleep quality scores (de Zambotti et al., 2023) and more.

A crucial aspect to comprehend is that these metrics are not uniformly derived, which leads to important implications for their accuracy. This distinction is vital because it would be erroneous to assume that an inaccuracy in a specific parameter (in a particular context, for instance, heart rate during exercise) renders the data inaccurate for all other parameters or even for the same parameter under different circumstances (e.g. heart rate measured at rest or during sleep).

We need to transcend this line of thinking, as wearables do not operate on an all-or-nothing basis. Just because one aspect may be lacking doesn't mean the entirety is compromised - it's akin to acknowledging that being subpar in one skill doesn't equate to incompetence in every skill.

The challenge lies in the fact that these devices are often marketed as excelling in every aspect, lacking transparency regarding signal quality, measurement error, estimate error, etc. Consequently, we find ourselves undertaking this evaluative task independently, which is no simple feat.

The framework that is proposed groups metrics according to how they are derived, the availability of reference devices to evaluate their accuracy, as well as the context in which they are acquired, which impacts both accuracy and interpretability. The following sections will first introduce differences between measurements and estimates (a first key distinction), and then provide insights on the importance of the context in which we measure. Then, estimates are discussed and further subdivided into estimates of known parameters (e.g. estimating calories, something we can verify with an indirect calorimeter) and estimates of unknown parameters (e.g. estimating “recovery”, something that does not have a reference device and is made-up by the wearable). Finally, the issue of determining when a change in a parameter is meaningful is discussed, and when it is just part of day-to-day variability - a step essential for practical actionability.

MEASUREMENTS VS ESTIMATES

When a parameter is measured, we determine its exact value, accounting for a margin of error. Measurements require a sensor designed for the task, according to what is an established measurement method for that specific parameter. For instance, a wearable might use an optical sensor to measure changes in blood volume in the microvascular bed of tissue during the cardiac cycle (Lemay et al., 2014), allowing it to determine pulse rate under certain conditions.

On the other hand, estimations involve making a guess. Estimations vary in complexity, ranging from simple methods (for example, determining maximal heart rate based on age, which is based on a simple regression model, Cruz-Martínez et al., 2014) to more intricate approaches (such as using a machine learning model to estimate sleep stages based on heart rate variability, movement, temperature and circadian features, Altini & Kinnunen, 2021). It's essential to recognize that estimations are fundamentally guesses. To truly measure maximal heart rate, a maximal test is necessary, and to accurately measure sleep stages, one would need to monitor brain waves, eye movement and muscle activity (Rundo & Downey III, 2019), factors not measured by most wearables that provide sleep stage data, as they are worn on the finger or the wrist.

Differentiating between measurements and estimates is crucial as it allows us to discern between data captured directly by the device's sensors and data derived from related parameters. This understanding helps us focus on what the device can reliably capture and what it is attempting to infer.

CONTEXT, ACCURACY AND INTERPRETABILITY OF MEASUREMENTS

In the previous sections, a distinction was made between measurements and estimates. As measurements are provided by a sensor designed for the purpose, they tend to be more accurate than estimates, which are guesses based on related parameters. However, even for measurements, we need to highlight two other key aspects that become important when using the data: context and interpretability.

The context of the measurement will impact both its accuracy and its interpretability. In simple terms, when we measure, might impact the accuracy and the interpretability of the data. Measurements are not flawless; they can have errors, influenced by when they are taken. For example, optical technology, like the sensors we typically find in wearables to measure pulse rate or its variability, are less accurate during movement, which affects signal quality and introduces higher error rates (Chow & Yang, 2020; Gillinov et al., 2017; Thomson et al., 2019; Xie et al., 2018). Evaluating measurements against a reference device, like an electrocardiogram or a chest strap, helps verify their accuracy in different contexts (Stone et al., 2021). We cannot extrapolate from measurements taken in a certain context and assume the device will behave similarly in a different one.

Context matters in the interpretability of measurements as well. Physiological parameters, such as heart rate variability (HRV), may only be meaningful when measured at specific times. For instance, HRV, used to capture the body's stress response, is most accurate when measured at rest, away from stressors, typically first thing in the morning or continuously during the night. Therefore, the time of measurement plays a crucial role in ensuring the collection of meaningful data. Measuring outside of these well-defined, specific contexts, would not lead to any meaningful use of the data. For example, in the context of HRV, the simplest things, like swallowing saliva (Yildiz & Doma, 2018) or drinking water (Grasser, 2020; Ragsdale et al., 2019) can create artifacts that last between a few minutes to hours. Hence, continuous HRV data are not representative of physiological stress the way it would be when acquired according to certain protocols (e.g. first thing in the morning, before eating, drinking, etc.).

This last point is key because we can always measure something, and wearables do market themselves on their ability to collect continuous data, but measurements do not necessarily lead to any useful or actionable information unless acquired in certain contexts and following certain protocols. We need to be aware of simplistic assumptions often made when it comes to wearables (e.g. higher is better for HRV) despite common associations between high HRV and poor health outcomes (Peyser et al., 2021).

ESTIMATES OF KNOWN AND UNKNOWN PARAMETERS

So far measurements have been covered, and while measurements are not perfect, they do provide sensors that are designed for the task. When looking at estimates, we often look at parameters that are guessed in more or less complex ways, based on other parameters that tend to be somewhat associated with the ones we are actually interested in. For example, if we are interested in sleep time but are not measuring brain waves, we can try to use movement and cardiac activity as proxies of sleep. It is easy to understand that for these reasons, estimates often carry larger errors (Düking et al., 2020). In the realm of estimates provided by wearables, a useful additional distinction can be made between estimates of known parameters and estimates of unknown parameters.

Estimates of known parameters refer to estimating a parameter that could be measured using different technology but is estimated by wearables due to practical constraints. Examples include calories and sleep stages, estimated from movement data and heart rate. While estimates should always be taken for what they are (i.e. guesses), estimates of known parameters give us the possibility to determine the accuracy of a given estimate by comparing it to a reference system, such as polysomnography for sleep stages or indirect calorimetry for calories. Reference systems do have their limitations, for example, they are often expensive and might lack ecological validity (e.g. measuring sleep in a sleep lab using polysomnography) and might be quite different from the actual sleep we would get in a different environment while not being monitored. If it is not feasible to use a reference system due to cost or other constraints, a useful approach could be to look for validations in research papers where participants are similar to the participants of interest for us (e.g. ourselves, athletes, or the population we are interested in studying, de Zambotti et al., 2023). Finally, given that wearable companies and the market move faster than academia and that it is not feasible to validate every single device and algorithm present on the market, nor it is possible to do it promptly, a different approach is recommended whenever possible. In particular, a simple way to determine if an estimate is valid and reliable is to compare multiple devices that provide such an estimate. For example, if wearables were able to estimate sleep stages or calories accurately, one should get very similar data when comparing multiple sensors, as well as ideally, very similar relative changes over time when monitoring an individual's longitudinally. If this is not the case (as found for the typical estimated metrics, such as calories, sleep stages, readiness, recovery, stress, oxygen saturation (SPO2), etc. - none of them are estimated consistently across devices), then this is a strong indication that we are currently unable to estimate such parameters with sufficient accuracy and reliability.

In concluding this section a few important points about estimates of known parameters must be made. It is common to generalize (e.g. a device is shown to be accurate in measuring or estimating a given parameter) and therefore we tend to believe that such a device will be equally accurate at measuring or estimating other parameters. This is unfortunately not the case, and we need to remember that just because a device is accurate at measuring one parameter (e.g. resting heart rate) it does not guarantee accuracy in estimating other parameters (e.g. SPO2, Spaccarotella et al., 2022). Secondly, error compounds. Estimates build on measurements that might include errors, or on other estimates, further compounding errors. For instance, estimating sleep stages using heart rate, itself measured with potential artifacts, introduces multiple layers of error.

BEHAVIOR OR RESPONSE?

Estimates of unknown parameters refer to parameters for which we do not have a reference system, and that are quite common in today's wearables (e.g. most made-up scores or metrics like readiness and recovery scores, sleep quality scores, strain, stress scores, etc.). Wearables provide these scores for a simple reason: they track many parameters and try to break down that information into something more digestible for the consumer, which means generating a single recovery or readiness score, or a sleep score. However, while aggregating information might give the false impression of providing a more complete picture, much of the context is lacking, and mixing physiology and behavior leads to a poor understanding of the individual response. For example, if a readiness or recovery score is lower, it might be due to the physiology being impaired (e.g. a lower HRV) or simply due to an assumption that the algorithm has made, based on the user's behavior (e.g. sleeping less time or being more active requires more recovery, according to the algorithm, or generic model assumptions, which might or might not be the case for the individual in question). Especially in sport settings, when wearables are used, we are not interested in blind guidance, but we are interested in analyzing the body's response to a given stimulus (training or else). We are not able to do so when using estimates of unknown parameters such as readiness, recovery, or sleep scores that mix behavior and physiology. Behavior remains key as context but hinders interpretation when mixed with physiology in a score. Given that estimates of unknown parameters are made up, inconsistent between devices, prone to change with software updates, have provided no correlation with the parameters they are trying to track (such as perceived stress and recovery, Lundstrom et al., 2023) and mix behavior and physiology in ways that prevent us to understand how the individual is responding to a given stimulus, they should probably have no place in the decision-making process. If our physiology is fine, it means that our body did not respond poorly to a disruption in sleep, for example, and therefore do not need to penalize our score because of sleep. All other parameters (sleep, activity, etc.) are inputs and become essential as contextual information. This is different from using behavioral parameters (also estimated with questionable accuracy) directly to determine our ability to perform on a given day which cannot be captured by a device that lacks information on essential parts of the equation, such as muscle soreness for example.

In conclusion, while wearables provide a wealth of data, critical consideration of the nature and reliability of estimates is essential. Focusing on measurements offers a more reliable basis for understanding the body's responses and making informed decisions. In the next section, a simple overview is provided to more effectively use data collected by wearables, without relying on estimates of unknown parameters (i.e. made-up scores) but relying on meaningful changes in physiological measurements.

SIGNAL VS NOISE AND MEANINGFUL CHANGES

Based on what has been discussed so far, we can start to define how to make use of wearables effectively. In particular, we need to start with a plan, depending on our goal. For athletes, the plan could simply be training periodization towards a certain event. In other contexts, we might be interested in a change in body composition, which similarly requires a targeted approach. Once we have a plan, the best use of wearables data, and where they can help us given the ease of use, comfort and accuracy of certain parameters, is to analyze relative changes in resting physiological data, to assess individual responses to our plan and other life stressors which might get in the way.

The focus should be on resting measurements given the higher accuracy and reliability, and not on estimates, and certainly not on estimates of unknown parameters (i.e. the made-up scores), using an accurate (validated) wearable. Once we have started collecting data, we need to be able to determine which changes in resting physiology are meaningful, and could be used to implement changes in training, and which variations in physiology are just part of day-to-day variability and should not be over-interpreted. To do so, sports science provides us with the concept of the smallest worthwhile change (Buchheit, 2014; Hopkins, 2000). Simply put, the smallest worthwhile change is what we can also call our normal range, or the range of values in which we expect the data to fall unless there are meaningful changes in the monitored parameter. If we take HRV as an example, a suppression below the normal range typically highlights significant stress on the body, and a poor physiological response (Altini & Plews, 2021). Only under these circumstances do we take action, by modulating training intensity for example (Carrasco-Poyatos et al., 2022; Javaloyes et al., 2019; Kiviniemi et al., 2007; Nuuttila et al., 2017). The normal range can be easily computed using historical data from an individual, as for example a range established as the mean plus or minus one standard deviation using data collected in the past 30-60 days (Figure 1).

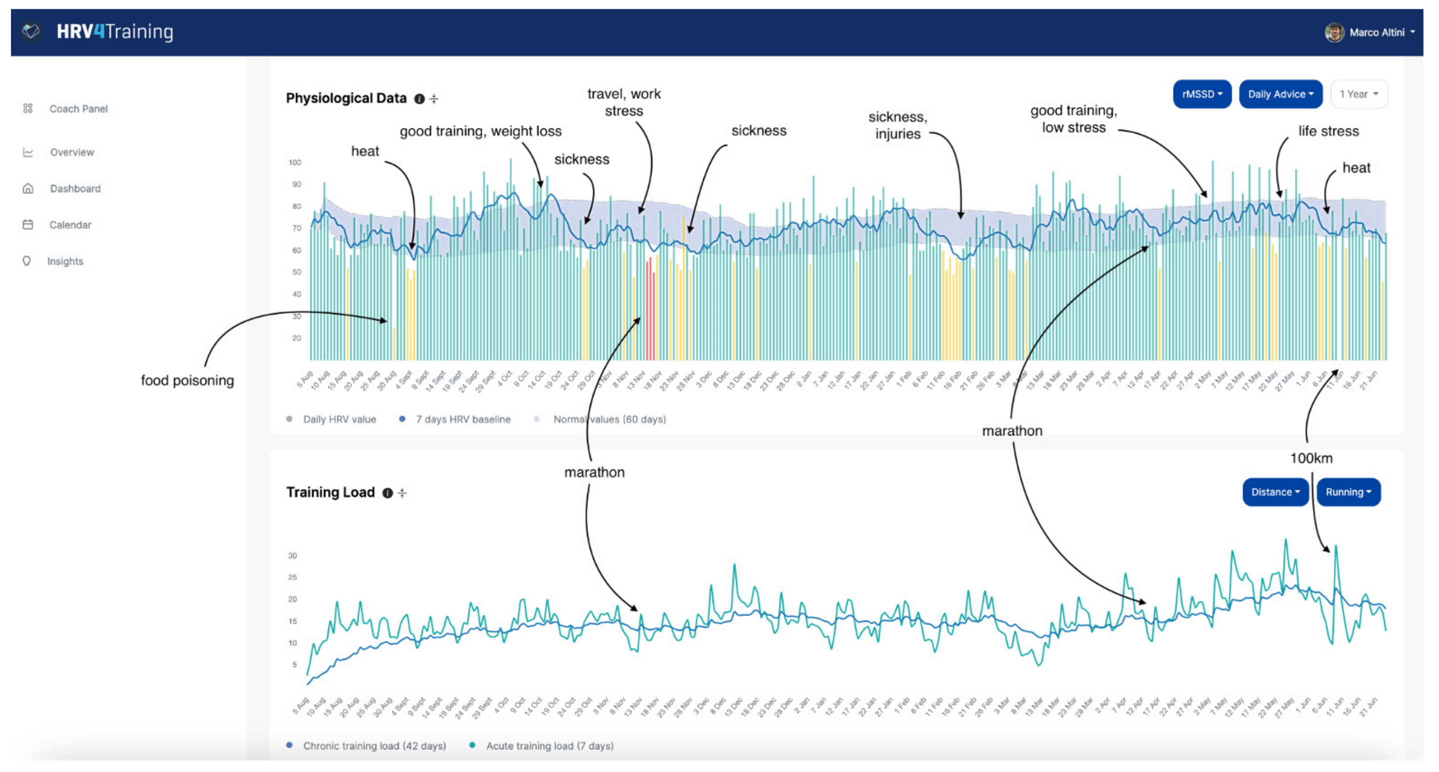

Figure 1: Example of resting physiological data (heart rate variability, HRV), normal range (shaded area) and context (training load and annotations).

PRACTICAL APPLICATION AND TAKEAWAYS

Measurements vs estimates is the first important classification when it comes to wearable-derived parameters:

- Focus on measurements under most circumstances, as they are directly captured by the wearable's sensors. Heart rate and HRV at rest, especially if provided as an average of the night or as a spot check measurement first thing in the morning, as well as skin temperature at rest, are often the only parameters measured accurately by wearables.
- Recognize that measurements can have larger errors in certain contexts, as when there is movement. Consider the right time for measurements to be meaningful in terms of interpretability. More data or continuous data does not necessarily mean useful data or that any additional actionable information is gathered.
- Focus on physiological responses: The key question when assessing these scores should be whether your behavior or physiology triggered a reduced score. The emphasis should be on the body's physiological response rather than penalizing scores for changes in behavior or external factors.
- No objective quantification: There's no objective way to quantify metrics like sleep quality or stress, and wearables may oversimplify physiological responses, lacking necessary context.

Considerations and recommendations on guidance and actionability:

- Goal Alignment: Consider whether you seek blind guidance or aim to understand your body's response to stressors. The first can make use of made-up scores, the second should rely on physiological changes with respect to an individual's normal range.
- Start with a Plan: Before interpreting wearable data, establish a plan, and use measurements (e.g., resting physiology) to capture responses to the plan, potentially making adjustments. Relying solely on made-up scores may otherwise lead to a reactive approach, responding to acute changes without a long-term focus.

SUMMARY

This SSE article advocates for a strategic approach to using wearables, emphasizing the importance of focusing on parameters that can lead to meaningful physiological insights, especially the ones measured at rest by sensors that were designed for the job, as opposed to the ones that are estimated from somewhat related variables. Despite the convenience of continuous data collection, wearables often overlook essential elements and contextual information, thereby challenging the notion that they offer a holistic view. A more intentional use is recommended, as opposed to the passive data collection often promoted. No amount of data can make up for poor protocols.

Due to issues with measurements during activity and the compounding of errors in many estimates, it is recommended to focus one's time and energy on the few parameters that can be actually measured. The suggestion is to minimize error by looking at resting data and maximize physiological meaning by examining night data or data collected first thing in the morning. The more we move away from resting measurements, the more noise, error and lack of context is introduced.

Users are encouraged to critically evaluate the capabilities of their wearables by asking questions such as whether the parameter of interest is actually measured or estimated and what the degree of error is. It is important to emphasize the importance of seeking validation papers for each parameter you are interested in and checking for consistency when using multiple wearables for the same parameter, especially when looking at estimates.

Aggregating information from wearables may create a false expectation of increased insight, whereas, in reality, it may only dilute the true understanding of the data. Taking a step back and relying on actual physiological data, observing deviations from the normal range and contextualizing this information with subjective reports and training data separately, will lead to greater insight. This approach aims to provide a balanced understanding of the device's capabilities without being overly influenced by hype or dismissing potential utility.

The views expressed are those of the author and do not necessarily reflect the position or policy of PepsiCo, Inc.

REFEFENCES

Altini, M., and H. Kinnunen (2021). The promise of sleep: A multi-sensor approach for accurate sleep stage detection using the oura ring. Sensors 21:4302.

Altini, M., and D. Plews (2021). What is behind changes in resting heart rate and heart rate variability? A large-scale analysis of longitudinal measurements acquired in free-living. Sensors 21:7932.

Buchheit, M. (2014). Monitoring training status with HR measures: do all roads lead to Rome? Front. Physiol. 5:73.

Carrasco-Poyatos, M., A. González-Quílez, M. Altini, and A. Granero-Gallegos (2022). Heart rate variability-guided training in professional runners: Effects on performance and vagal modulation. Physiol. Behav. 244:113654.

Chow, H.W., and C.C. Yang (2020). Accuracy of optical heart rate sensing technology in wearable fitness trackers for young and older adults: Validation and comparison study. JMIR MHealth and UHealth, 8:e14707.

Cruz-Martínez, L.E., J.T. Rojas-Valencia, J.F. Correa-Mesa, and C. Correa-Morales (2014). Maximum heart rate during exercise: Reliability of the 220-age and Tanaka formulas in healthy young people at a moderate elevation. Rev. de la Facult. de Med. 62:579-585.

de Zambotti, M., C. Goldstein, J. Cook, L. Menghini, M. Altini, P. Cheng, and R. Robillard (2023). State of the science and recommendations for using wearable technology in sleep and circadian research. Sleep zsad325 (online ahead of print).

Düking, P., L. Giessing, M.O. Frenkel, K. Koehler, H.C. Holmberg, and B. Sperlich (2020). Wrist-worn wearables for monitoring heart rate and energy expenditure while sitting or performing light-to-vigorous physical activity: validation study. JMIR MHealth and UHealth, 8:e16716.

Fuller, D., Colwell, E., Low, J., Orychock, K., Tobin, M. A., Simango, B., ... & Taylor, N. G. (2020). Reliability and validity of commercially available wearable devices for measuring steps, energy expenditure, and heart rate: systematic review. JMIR mHealth and uHealth, 8(9), e18694.

Georgiou, K., A.V. Larentzakis, N.N. Khamis, G.I. Alsuhaibani, Y.A. Alaska, and E.J. Giallafos (2018). Can wearable devices accurately measure heart rate variability? A systematic review. Folia Med. 60:7-20.

Gillinov, S., M. Etiwy, R. Wang, G. Blackburn, D. Phelan, A.M. Gillinov, P. Houghtaling, H. Javadikasgari, and M.Y. Desai (2017). Variable accuracy of wearable heart rate monitors during aerobic exercise. Med. Sci. Sports Exerc. 49:1697-1703.

Grasser, E.K. (2020). Dose-dependent heart rate responses to drinking water: A randomized crossover study in young, non-obese males. Clin. Autonom. Res. 30:567-570.

Hopkins, W.G. (2000). Measures of reliability in sports medicine and science. Sports Med. 30:1-15.

Ibrahim, A.H., C.T. Beaumont, and K. Strohacker (2024). Exploring regular exercisers' experiences with readiness/recovery scores produced by wearable devices: A descriptive qualitative study. Recovery scores produced by wearable devices: A descriptive qualitative study. Available at SSRN: https://ssrn.com/abstract=4694263 or http://dx.doi.org/10.2139/ssrn.4694263

Javaloyes, A., J.M. Sarabia, R.P. Lamberts, and M. Moya-Ramon (2019). Training prescription guided by heart-rate variability in cycling. Int. J. Sports Physiol. Perform. 14:23-32.

Kiviniemi, A.M., A.J. Hautala, H. Kinnunen, and M.P. Tulppo (2007). Endurance training guided individually by daily heart rate variability measurements. Eur. J. Appl. Physiol. 101:743-751.

Lee, J.M., W. Byun, A. Keill, D. Dinkel, and Y. Seo. (2018). Comparison of wearable trackers' ability to estimate sleep. Int. J. Environ. Res. Public Health 15:1265.

Lemay, M., M. Bertschi, J. Sola, R. Renevey, J. Parak, and I. Korhonen (2014). Application of optical heart rate monitoring. In: Wearable Sensors. Academic Press, pp. 105-129.

Lundstrom, E.A., M.J. De Souza, K.J. Koltun, N.C. Strock, H.N. Canil, and N.I. Williams (2023). Wearable technology metrics are associated with energy deficiency and psychological stress in elite swimmers. Int. J. Sports Sci. Coach. (online ahead of print). https://doi.org/10.1177/17479541231206424.

Maijala, A., H. Kinnunen, H. Koskimäki, T. Jämsä, and M. Kangas (2019). Nocturnal finger skin temperature in menstrual cycle tracking: Ambulatory pilot study using a wearable Oura ring. BMC Women's Health 19:1-10.

Nelson, B.W., C.A. Low, N. Jacobson, P. Areán, J. Torous, and N.B. Allen (2020). Guidelines for wrist-worn consumer wearable assessment of heart rate in biobehavioral research. NPJ Dig. Med. 3:90.

Nuuttila, O.P., A. Nikander, D. Polomoshnov, J.A. Laukkanen, and K. Häkkinen (2017). Effects of HRV-guided vs. predetermined block training on performance, HRV and serum hormones. Int. J. Sports Med. 38:909-920.

Peyser, D., B. Scolnick, T. Hildebrandt, and J.A. Taylor (2021). Heart rate variability as a biomarker for anorexia nervosa: A review. Eur. Eating Disord. Rev. 29:20-23.

Ragsdale, C.C., J.T. Ellis, J. Phelps, N. Foster, and A.A. Flatt (2019). Sports drink ingestion inflates heart rate variability: implications for pre-training measures. Department of Health Sciences and Kinesiology Faculty Presentations. Presentation 298. source: http://www.eventscribe.com/2019/posters/nsca/SplitViewer.asp?PID=NDkxMzgwMzlyNDE https://digitalcommons.georgiasouthern.edu/health-kinesiology-facpres/298.

Rundo, J.V., and R. Downey III (2019). Polysomnography. Handb. Clin. Neurol. 160:381-392.

Spaccarotella, C., A. Polimeni, C. Mancuso, G. Pelaia, G. Esposito, and C. Indolfi (2022). Assessment of non-invasive measurements of oxygen saturation and heart rate with an apple smartwatch: Comparison with a standard pulse oximeter. J. Clin. Med. 11:1467.

Stone, J.D., H.K. Ulman, K. Tran, A.G. Thompson, M.D. Halter, J.H. Ramadan, M. Stephenson, V.S. Finomore, Jr, S.M. Galster, A.R. Rezai, and J.A. Hagen (2021). Assessing the accuracy of popular commercial technologies that measure resting heart rate and heart rate variability. Front. Sports Act. Living 3:585870.

Thomson, E.A., K. Nuss, A. Comstock, S. Reinwald, S. Blake, R.E. Pimentel, B.L. Tracy, and K. Li (2019). Heart rate measures from the Apple Watch, Fitbit Charge HR 2, and electrocardiogram across different exercise intensities. J. Sports Sci. 37:1411-1419.

Xie, J., D. Wen, L. Liang, Y. Jia, L. Gao, and J. Lei (2018). Evaluating the validity of current mainstream wearable devices in fitness tracking under various physical activities: Comparative study. JMIR MHealth UHealth 6:e975.

Yildiz, M., and S. Doma (2018). Effect of spontaneous saliva swallowing on short-term heart rate variability (HRV) and reliability of HRV analysis. Clin. Physiol. Funct. Imag. 38:710-717.