DEVELOPMENT AND USAGE OF ANALYTIC TOOLS AND RECOMMENDATIONS FOR VALIDATION AND RELIABILITY STUDIES USING CONSUMER-GRADE WEARABLE TECHNOLOGY

By

Bryson Carrier

Bachelor of Science – Biology
Utah Valley University
2017

Master of Science – Kinesiology
University of Nevada, Las Vegas
2021

A dissertation submitted in partial fulfillment
of the requirements for the

Doctor of Philosophy – Interdisciplinary Health Sciences

The Graduate College

University of Nevada, Las Vegas
May 2024

UNLV | GRADUATE COLLEGE

Dissertation Approval

The Graduate College
The University of Nevada, Las Vegas

April 1, 2024

This dissertation prepared by

Bryson Carrier

entitled

Development and Usage of Analytic Tools and Recommendations for Validation and Reliability Studies Using Consumer-Grade Wearable Technology

is approved in partial fulfillment of the requirements for the degree of

Doctor of Philosophy – Interdisciplinary Health Sciences
The Graduate College

James Navalta, Ph.D.
Examination Committee Chair

John Mercer, Ph.D.
Examination Committee Member

Jennifer Bunn, Ph.D.
Examination Committee Member

Chad Cross, Ph.D.
Graduate College Faculty Representative

Alyssa Crittenden, Ph.D.
Vice Provost for Graduate Education & Dean of the Graduate College

Abstract

Introduction: Wearable technology is increasingly utilized across various fields, yet the validity and reliability of the physiological data these devices provide are often unverified due to a lack of rigorous testing standards.

Purpose: This dissertation contains three primary works, and therefore multiple purposes. The purpose of the first project (Chapter 2) is to introduce a new risk of bias assessment tool, specifically for assessing methodological quality and the risk of bias in validity and reliability studies using wearable technology, with a focus on consumer-grade wearable technology. The purpose of the second project (Chapter 3) was to perform a systematic review and meta-analysis that served a dual purpose: to review the current validity and reliability literature concerning consumer-grade wearable technology measurements/estimates of physiological variables (e.g. heart rate, energy expenditure, etc.) during exercise. Additionally, we sought to perform risk of bias assessments utilizing the novel **WEA**rable technology **R**isk of **B**ias and **O**bjectivity **T**ool (WEAR-BOT) and perform meta-analytic calculations on the reported data. The purpose of the third project (Chapter 4) was to evaluate the accuracy (validity) of maximal oxygen consumption (VO2max) estimates and blood oxygen saturation (BOS) measured via pulse oximetry using the Garmin fēnix 6 with a general population participant pool.

Methods: Chapter 1: The development of WEAR-BOT through a multi-institutional collaboration, employing iterative discussions, Delphi-style surveys, and pilot testing. Chapter 2: A systematic review and meta-analysis using WEAR-BOT to assess the risk of bias and analyze the validity and reliability data of physiological measurements from consumer-grade wearables during exercise. Chapter 3: A validation study employing WEAR-BOT guidelines to test the accuracy of a wearable device in measuring aerobic capacity (VO2max) and pulse oximetry in the general population.

Results: The development of WEAR-BOT established a detailed and structured approach to evaluate wearable technology studies. The systematic review highlighted a prevalent high risk of bias within the field, indicating the need for standardization. The validation study demonstrated the practical application of WEAR-BOT, confirming its effectiveness in guiding rigorous research methodologies and producing reliable data.

Conclusion: By introducing and applying the WEAR-BOT, this dissertation significantly contributes to the standardization and enhancement of research methods in the domain of wearable technology. The tool not only aids researchers in designing and evaluating studies but also ensures that the data generated from wearable devices are both reliable and valid, fostering greater trust and broader application in health-related and athletic settings.

Acknowledgements

I am deeply grateful to my dissertation committee for their invaluable guidance throughout this journey. My sincere thanks to Dr. James Navalta, who served not only as the chair of my advisory committee but also as my mentor during my Master's Degree and Doctoral Degree. His insights and encouragement have been pivotal to my academic and personal growth. Dr. John Mercer, Dr. Jen Bunn, and Dr. Chad Cross, as members of my advisory committee, have provided critical feedback and support that greatly enhanced the quality of my research.

I would also like to acknowledge my former advisors, Dr. Andrew Creer, Dr. Tyler Standifird, Dr. Lauren Brooks, and Dr. T. Heath Ogden, whose early guidance established a research foundation for my academic pursuits, culminating in this dissertation.

My co-authors on the various components of this dissertation deserve special mention for their collaboration and expertise. For the development of the WEAR-BOT checklist, I am thankful to Dr. Jen Bunn, Dr. Chris Eschbach, Dr. Joel Reece, Dr. Gregory Welk, Dr. Brett Dolezal, and again Dr. James Navalta. In the systematic review and meta-analysis, I was fortunate to work alongside Dr. Jen Bunn, Dr. Chris Eschbach, Dr. Joel Reece, Dr. Charli Aguilar, and Dr. James Navalta. For the validation study on VO2max and pulse oximetry, my gratitude extends to Brenna Barrios, Sofia Marten Chaves, and once more to Dr. Navalta for their dedication and hard work.

I must also express my gratitude to the many lab members, classmates, and friends throughout the years who have provided both academic support and much-needed escape from the rigors of higher education. Thank you to all my professors, who have taught me so much and helped me to this point.

Above all, my deepest appreciation goes to my wife, Stacy Carrier, who has stood by me with unwavering support throughout ten long years of higher education. Her emotional support and companionship have

been the cornerstone of my path to pursue and fulfill my academic goals. Stacy, you are truly the best friend and partner I could ever ask for, thank you for all you do.

Table of Contents

List of Tables

List of Figures

Chapter 1 - Introduction

Wearable technology is becoming increasingly pervasive in society (Benson et al., 2018; Vogels, 2020). Its application can be seen in the general population, collegiate and professional athletics, military, construction, healthcare, academic research, as well as others. Wearable technology has the potential to revolutionize physiological research, due to its constant monitoring and production of granular physiological and physical data, examining many aspects of human life (Carrier et al., 2020; Wright et al., 2017). However, the need for independent validation of this technology is needed, as there exists no governing entity to ensure accuracy. The demand for independent validity and reliability studies has been met by academic researchers, performing validity and reliability testing on consumer-grade wearable technology to establish whether these devices are valid and reliable, and under what circumstances and scenarios they perform well. This can be seen by the dramatic increase in validation and reliability studies that have been published over the last 10 years. This increase in research has also introduced the need for standardized and appropriate practices when evaluating validity and reliability, as well as risk of bias assessment tools.

This dissertation was inspired by my personal research journey in wearable technology. Our lab group was performing a systematic review previously and experienced the lack of appropriate risk of bias assessment tools specific to wearable technology (Carrier et al., 2020). The review not only highlighted the need for such tools but also underscored the limitations of existing tools, including COSMIN, in fully addressing the nuances of consumer-grade wearable devices, especially related to the field of exercise physiology, which has been my concentration in graduate school (Mokkink et al., 2006; Mokkink et al., 2010; Prinsen et al., 2018).

This realization birthed the development of the **WEA**rable technology **R**isk of **B**ias and **O**bjectivity **T**ool (WEAR-BOT) checklist, a tool meticulously designed to fill a gap in current risk of bias tools. The WEAR-

BOT represents a significant leap forward, offering an easy-to-use tool for evaluating bias in wearable technology studies by reviewers performing systematic reviews, as well as suggesting best practices for researchers looking to design their own validity or reliability studies. The WEAR-BOT was developed by a multi-institutional team of experts in wearable technology research and who's research specifically focused on validity and reliability of wearable technology. Through numerous hours of discussion, literature review, and pilot testing, it was developed and encapsulates a set of best practices and recommendations that can be used by reviewers performing a systematic review, or individual researchers looking to perform their own studies. It is meticulously crafted to ensure that studies on wearable devices are not only conducted with the utmost rigor but also presented in a manner that facilitates critical evaluation and replication. This standardization is crucial for advancing the field, enabling researchers to build upon each other's work with confidence in the reliability and validity and to clearly establish the validity and reliability of these devices, so they can further influence many other fields of research.

Diving deeper into the practical application of this tool, this dissertation features a systematic review and meta-analysis (D2) that employs the WEAR-BOT checklist to critically evaluate validation and reliability studies focusing on physiological variables measured by consumer-grade wearable devices during exercise. This exploration serves a dual purpose: validating the WEAR-BOT's efficacy in a real-world research scenario and refining its parameters to ensure broad applicability and robustness. This phase of the project is pivotal, acting as a proving ground for the WEAR-BOT checklist and setting a precedent for its adoption in future wearable technology research.

The final project of this dissertation (D3) is a validation study of wearable technology on aerobic capacity (VO2max) and pulse oximetry. This study illustrates the type of research the WEAR-BOT can be used to evaluate, or to guide researchers in planning to reduce bias. By conducting this validation study, the dissertation not only contributes to the body of knowledge on wearable technology's capabilities but

also demonstrates the practical application of the WEAR-BOT tool in guiding and assessing research quality.

Overall, this doctoral dissertation propels the field of validation and reliability studies with consumer-grade wearable technology design forward. The first paper details the WEAR-BOT checklist and guidelines for use. The second study provides an example of a systematic review and utilizes the WEAR-BOT checklist to determine risk of bias. The final investigation displays an example of how an individual study can be designed to align with WEAR-BOT best practices so that valid and reliable physiological data can be produced from commercially available wearable devices.

Chapter 1 References

Benson, L. C., Clermont, C. A., Bošnjak, E., & Ferber, R. (2018). The use of wearable devices for walking and running gait analysis outside of the lab: A systematic review. Gait & Posture, 63, 124-138.

Carrier, B., Barrios, B., Jolley, B. D., & Navalta, J. W. (2020). Validity and Reliability of Physiological Data in Applied Settings Measured by Wearable Technology: A Rapid Systematic Review. Technologies, 8(4), 70.

Mokkink, L. B., Terwee, C. B., Knol, D. L., Stratford, P. W., Alonso, J., Patrick, D. L., Bouter, L. M., & De Vet, H. C. (2006). Protocol of the COSMIN study: COnsensus-based Standards for the selection of health Measurement INstruments. BMC Medical Research Methodology, 6, 1-7.

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., Bouter, L. M., & De Vet, H. C. (2010). The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Quality of Life Research, 19, 539-549.

Prinsen, C. A., Mokkink, L. B., Bouter, L. M., Alonso, J., Patrick, D. L., De Vet, H. C., & Terwee, C. B. (2018). COSMIN guideline for systematic reviews of patient-reported outcome measures. Quality of Life Research, 27, 1147-1157.

Vogels, E. A. (2020, Jan. 9,). About one-in-five Americans use a smart watch or fitness tracker. www.pewresearch.org. Retrieved 3/8/2024, from https://www.pewresearch.org/short-reads/2020/01/09/about-one-in-five-americans-use-a-smart-watch-or-fitness-tracker/

Wright, S., Brown, T., Collier, S., & Sandberg, K. (2017). How consumer physical activity monitors could transform human physiology research. American Journal of Physiology-Regulatory Integrative and Comparative Physi, 312(3), R358-R367. 10.1152/ajpregu.00349.2016

Chapter 2 - The WEAR-BOT Checklist: A Risk of Bias Tool for Evaluating Validity and Reliability Research in Wearable Technology.

Abstract

This paper proposes an innovative tool designed to standardize the evaluation of validity and reliability studies in the rapidly evolving field of wearable technology. We introduce the WEArable Technology Risk of Bias and Objectivity Tool (the WEAR-BOT), a tool that addresses the need for a comprehensive and systematic way to assess bias in studies examining consumer-grade and research-grade wearable devices. The development of the WEAR-BOT involved extensive collaboration among experts, encompassing iterative, open-ended discussions, several rounds of anonymous Delphi-style questionnaires, and pilot testing. The tool comprises detailed checklists for both validity and reliability studies, with subdivisions focusing on study design, methodology, statistical analysis methods, and other critical aspects. The tool balances need for rigor with ease of use. It incorporates a variety of questions to rigorously evaluate the risk of bias in these studies, and aims to enhance and standardize methodological approaches in the field. The tool is practical, easily available, and easy to use, as it is built in Google Sheets and contains macros that are intuitive and easy to use that allow the user to work more efficiently. The WEAR-BOT represents a significant advancement in the standardization of research methods and statistical analysis in the domain of wearable technology.

Introduction

The popularity of wearable technology has led to corresponding increases in research on the reliability and validity of these devices. For instance, using the search terms “wearable technology or fitness tracker or activity monitor + validity or reliability” in Google Scholar (Alphabet Inc., Mountain View, CA, USA) produces 486, 1080, 3,640, 11,100, and 14,900 results for the years 2005, 2010, 2015, 2020, and 2022, respectively. There has been a steady increase in this type of research, with a nearly 1000% increase from 2010 to 2020. This nascent technology is being used by a range of different people for numerous use cases, frequently without a clear indication of the validity or reliability of the devices. Individual users, organizations, and even researchers often assume that the objective data on these devices are valid and reliable, frequently without any evidence to support such conclusions. These devices may be used to make training decisions or health assessments, though they may have provided poor results from inaccurate devices. Independent research determines the validity and reliability of these devices so users may be aware of their accuracy under different use cases is warranted. This is especially important as consumer-grade devices and many “research-grade” devices are not regulated by any governing entity. The responsibility of testing the accuracy of these devices’ rests upon independent researchers (other than the testing specific organizations may do internally, which are not generally made available to the public). As researchers have sought to validate these devices, the studies have used varied methodologies, statistical analyses, and reporting practices across studies and researchers (BUNN et al., 2018; Carrier et al., 2020; Evenson et al., 2015; Patel et al., 2021; Welk et al., 2019). Methodological and statistical best practices have been suggested by some researchers (Carrier et al., 2020; Keadle et al., 2019; van Lier et al., 2020; Welk et al., 2012; Welk et al., 2019), but a consensus has not been adequately reached. Thus, a standardizing of the methods, statistical analyses, and reporting practices is needed. We have addressed this gap with the development and refinement of an easy-to-use checklist.

Risk of bias assessment tools have several use cases. For instance, they are frequently used (i) when performing systematic reviews to evaluate the literature being reviewed, or (ii) by journal reviewers to assess the bias risk in a particular study, or (iii) by researchers aiming to design a study that will reduce the risk of bias in their own research. There are many tools that have been developed for almost all study designs. Common tools are the Cochrane Risk of Bias (ROB) 2.0 (Sterne et al., 2019), the Cochrane Risk Of Bias In Non-randomised Studies-of Interventions (ROBINS-I) (Sterne et al., 2016), JBI's critical Appraisal Tools (Aromataris et al., 2015; Barker et al., 2023; Campbell et al., 2020; Munn et al., 2020), along with many others. For assessing the risk of bias and methodological quality for studies on measurement properties, the COnsensus-based Standards for the selection of health status Measurement Instruments (COSMIN) checklist was published in 2006 with additional publications regarding clarifications and guidelines being published subsequently (Gagnier et al., 2021; Mokkink et al., 2006; Mokkink et al., 2010; Prinsen et al., 2018). The COSMIN is designed to evaluate measurement properties associated with patient-reported outcome measures (PROM), and thus spends time on aspects that may not be relevant to validity and reliability literature broadly, or with consumer-grade wearable technology. It has been utilized by researchers performing a risk of bias assessment in conjunction with a systematic review and meta-analysis; however, it lacks the specificity needed to systematically evaluate the rigor and methodological approaches in validity and reliability studies using consumer-grade wearable technology. Since the publication of the COSMIN checklist, the field of wearable technology has grown rapidly, and acceptable practices have been further established for validity and reliability studies. An updated risk of bias checklist should is needed that reflects the many changes that have occurred in wearable technology and make recommendations based on updated best practices for methodology, analysis, and reporting practices. Therefore, the purpose of this paper is to introduce a new risk of bias assessment tool, specifically for assessing methodological quality and the risk of bias in validity and reliability studies using wearable technology, with a focus on consumer-grade

wearable technology. The **WEA**rable Technology **R**isk of **B**ias and **O**bjectivity **T**ool (WEAR-BOT) is introduced here, with accompanying instructions for use below.

Tool Development Methodology

This tool was developed through iterative discussions and subsequent use with academics and professionals involved in the field of wearable technology, with a focus on testing the validity and reliability of consumer-grade devices. Eight researchers were invited to participate in the development of this tool, and seven agreed to participate. Discussions with individual researchers and the primary investigators were first conducted to introduce the general idea and discuss the need for development of a novel risk of bias tool. Regular group meetings then commenced, spanning several months, where published literature was reviewed and discussed and suggestions for tool development were made. This included reviewing published risk of bias tools, validity and reliability studies, and other recommendations made in published literature, while gradually developing the checklist. Discussion on recommendations, questions, wording, and the overall scope were constantly evaluated over the months of tool development. As researchers were located throughout the United States, all meetings were completed virtually. All meetings were an open-ended discussion, where each researcher contributed as they wished. Decisions on scope, content, wording, and all other aspects of the tool were discussed in an open manner until all researchers were satisfied with the initial results.

Once the questions and tool were mostly established, a pilot study or proof-of-concept systematic review was performed. The results of this systematic review will be published elsewhere. After the completion of the review, any suggestions for changes to the tool from the researchers involved in the systematic review were considered by the entire group, and instituted where consensus was reached. After several iterations of the tool were developed, criticized, and pilot tested, several rounds of an

anonymous, Delphi-style questionnaire were conducted to determine if consensus had been reached on several aspects of the checklist. These included aspects such as: question wording, category scope, and other aspects. Minor changes were made as a result of the Delphi questionnaires and are reflected in the current version of the WEAR-BOT checklist. Consensus was reached on all aspects of the tool that were questioned in the Delphi questionnaire process.

The final tool can be seen below, but the use of the tool must be done using Google Sheets (Alphabet Inc., Mountain View, CA, USA), as macros specific to Google Sheets do not transfer over to other spreadsheet products. The link for tool use is: https://docs.google.com/spreadsheets/d/1npIGT9SJl0_E6RfLRi7ZFKeQnjR2O05bvuGSkhhtoWM/edit?usp=sharing.

Results

The novel WEAR-BOT tool consists of two checklists, one for validity studies, and one for reliability studies. Each checklist is split into two broad categories, “Study Design and Methodology”, and “Statistical Analysis Methods”. There are subcategories, that each have questions intended to evaluate the risk of bias found in the study being evaluated. The researcher using the tool must answer one of the following, “Yes”, “Probably Yes”, “Probably No”, “No”, or “Not Applicable”. The subcategories for the validity checklist are 1. Test Variables, 2. Criterion Device, 3. Test Devices, 4. Test Protocols and Parameters, 5. Participants, 6. Data Processing, 7. Statistical Tests – Continuous Variables, and 8. Statistical Tests – Categorical Variables. Subcategories 1-5 are under “Study Design and Methodology”, whereas 6-8 are under “Statistical Analysis Methods” (see Figure 1). The validity checklist also has an “Areas of Consideration”, that the answers are not factored into the overall risk of bias calculations but may be considered by researchers looking to design their own studies (see Figure 2). The reliability

checklist has only two subcategories, which are both under “Statistical Analysis Methods”. Therefore, the reliability checklist contains questions for “Study Design and Methodology” (no subcategories), and “Data Processing” and “Statistical Tests” under “Statistical Analysis Methods” (see Figure 3). Some instructions for use are published in the tables/tool itself, for guidance on common issues researchers may run into. For more complete instructions, see below under the “General Guidelines and Instructions for Use” section of this paper.

WEArable Technology **R**isk of **B**ias and **O**bjectivity **T**ool (WEAR-BOT) - Validation Studies								
Authors:								
Year:								
Study Design and Methodology			Yes	Probably Yes	Probably No	No	Not Applicable	Result
1	Test Variables							
	a	Are the units of measurement (or estimated values) between test device and criterion the same?	☐	☐	☐	☐	☐	Low Risk of Bias
2	Criterion Device							
	a	Is there clear evidence that the criterion device/method used is valid?	☐	☐	☐	☐	☐	
	b	Was the criterion device properly calibrated, synced, and updated prior to testing?	☐	☐	☐	☐	☐	
	d	Was the software used for analysis reported and appropriate?	☐	☐	☐	☐	☐	Low Risk of Bias
3	Test Devices							
	a	Were the test devices properly synced, updated, and/or calibrated (if applicable) prior to testing?	☐	☐	☐	☐	☐	
	b	If necessary, was the data and/or settings reset between each test?	☐	☐	☐	☐	☐	
	c	If necessary, were participant demographics input for each test?	☐	☐	☐	☐	☐	
	d	Was device placement on the participant standardized and appropriate?	☐	☐	☐	☐	☐	
	e	Were the devices used in a way the manufacturers would approve of?	☐	☐	☐	☐	☐	
	f	Were all devices, software, and accessories used reported?	☐	☐	☐	☐	☐	Low Risk of Bias
4	Test Protocols and Parameters							
	a	Were measurements between the criterion and test device taken concurrently?	☐	☐	☐	☐	☐	
		If tested sequentially, was an appropriate amount of time given between tests?	☐	☐	☐	☐	☐	
	b	i. Were participants provided a sufficient amount of time so the measurement would not be affected by the previous test bout, or too long that physiological or physical traits may have changed? Select N/A if tested concurrently.						
	c	Was the data collection time interval reported?	☐	☐	☐	☐	☐	
	d	Was the test setting reported (e.g. laboratory, free-living, field)?	☐	☐	☐	☐	☐	
	e	Were steps taken to control for any potential confounding variables in the test environment?	☐	☐	☐	☐	☐	Low Risk of Bias
5	Participants							
	a	Did researchers perform an *a priori* power analysis or state other justifications to determine sample size?	☐	☐	☐	☐	☐	
	b	Was the inclusion/exclusion criteria and sample population reported and described (e.g. age, BMI, fitness level, disease status)?	☐	☐	☐	☐	☐	
	c	Were any potential confounding variables regarding participants identified that would influence the measurements (see limitations section)?	☐	☐	☐	☐	☐	Low Risk of Bias
Statistical Analysis Methods								
6	Data Processing							
	a	Were the data processing methods described appropriately and in a reproducible manner?	☐	☐	☐	☐	☐	
		Was the amount of missing and/or cleaned data reported?	☐	☐	☐	☐	☐	
	b	i. Select N/A if data was cross-sectional (no repeated measures) or cleaning not necessary.						
	c	Was it reported how missing data from the criterion and/or test devices were handled?	☐	☐	☐	☐	☐	
		Was justification provided for any data removed by the researchers?	☐	☐	☐	☐	☐	
	d	i. Select N/A if removed data was not reported (if answer to question 6c was "No" or "Probably No").						
		If necessary, was the method of aligning data reported and reasonable (e.g. aligned on timestamp, elapsed time)?	☐	☐	☐	☐	☐	
	e	i. Select N/A if not necessary (e.g. if data was cross-sectional).						
		If necessary, were steps taken to ensure there was no lag in the data (or signal processing), compared to the criterion, to ensure appropriate alignment (e.g. cross correlations)?	☐	☐	☐	☐	☐	
	f	i. Select N/A if not necessary (e.g. if data was cross-sectional).						
	g	Was the time interval (epoch) for data aggregation reported and appropriate (e.g. 1 sec, 5 sec, 1 min)?	☐	☐	☐	☐	☐	
	h	Was all software, programming scripts, or other resources used for statistical analysis disclosed?	☐	☐	☐	☐	☐	Low Risk of Bias
7	Statistical Tests - Continuous Variables							
		Were the test variables continuous?	☐	☐	☐	☐	☐	
	a	i. If 7a answer was no, answer the categorical variables section (section 8) and put N/A for all subsequent section 7 questions.						
	b	Were multiple tests used to determine validity?	☐	☐	☐	☐	☐	
	c	Was a test of error performed (e.g. MAPE, MAE, RMSE)?	☐	☐	☐	☐	☐	
	d	Was linearity between the test device and criterion established via correlation and/or regression?	☐	☐	☐	☐	☐	
		If correlation was used, was appropriate correlation tests used (Pearson's, Lin's Concordance, Spearman's, etc.)?	☐	☐	☐	☐	☐	
	e	i. Select N/A if correlation was not performed.						
		If necessary, was repeated measures correlation determined if there were non-independent samples?	☐	☐	☐	☐	☐	
	f	i. Select N/A if repeated measures correlation analysis was not necessary or correlation not performed.						
		If regression was used, was appropriate regressions utilized (e.g. Simple Linear, Deming, Passing-Bablok)?	☐	☐	☐	☐	☐	
	g	i. Select N/A if regression was not performed.						
		If regression was used, were appropriate model fit statistics reported?	☐	☐	☐	☐	☐	
	h	i. Select N/A if regression was not performed.						
	i	Was equivalence testing performed (e.g. TOST test, confidence interval for difference in means)?	☐	☐	☐	☐	☐	
	j	Was bias/agreement plotted via a Bland-Altman plot?	☐	☐	☐	☐	☐	
		If performed, was bias and limits of agreement estimates reported for the Bland-Altman analysis?	☐	☐	☐	☐	☐	
	k	i. Select N/A if Bland-Altman analysis was not performed.						
		If effect size was calculated, was it based on linear association (e.g. R^2) rather than less appropriate calculations using descriptive statistics (e.g. Cohen's D)?	☐	☐	☐	☐	☐	
	l	i. Select N/A if effect size was not calculated or effect size based on descriptive statistics was not used in the interpretation of the validity of the device (if it was calculated but not used to determine if validity was achieved).						
	m	Were validity thresholds reported?	☐	☐	☐	☐	☐	Low Risk of Bias
8	Statistical Tests - Categorical Variables							
		Were the test variables categorical?	☐	☐	☐	☐	☐	
	a	i. If 8a answer was "No", answer the continuous variables section and put N/A for all subsequent section 8 questions.						
	b	Were diagnostic tests performed to assess predictive validity (e.g. sensitivity, specificity, accuracy, AUC)?	☐	☐	☐	☐	☐	
	c	Was the classification table (confusion matrix) reported?	☐	☐	☐	☐	☐	
	d	Was association between the test device and criterion established with appropriate categorical correlation statistics (e.g. Cohen's Kappa, Cramer's V, tetrachoric [for nominal variables], rank-based correlations, polychoric [for ordinal variables])?	☐	☐	☐	☐	☐	
	e	Were validity thresholds reported?	☐	☐	☐	☐	☐	Low Risk of Bias

Figure 2.1. WEAR-BOT Checklist for Validity Studies.

WEArable Technology **R**isk of **B**ias and **O**bjectivity **T**ool (WEAR-BOT) - Validation Studies							
Authors:							
Year:							
Study Design and Methodology			Yes	Probably	Probably	No	Not
Areas of Consideration (Answers not considered in risk of bias calculation)							
	Inference Testing						
	a	Were any tests of mean differences performed that are unable to determine the validity of the test devices (e.g. ANOVA, t-test)?	☐	☐	☐	☐	☐
		If yes to question Ia, were the results of the inference test(s) used in the interpretation of the validity of the device? Inference testing is incapable of determining validity, and should only be used to assess adherence to the null model.	☐	☐	☐	☐	☐
I	b	i. If no to question Ia, answer N/A.					
	Environmental Factors						
II	a	Were environmental factors reported (e.g. temperature, humidity, altitude)?	☐	☐	☐	☐	☐
	Participant Biological Variability						
III	a	Were any steps taken to assess or control for participant biological variability, such as potential bilateral asymmetries in participants (differences between left and right sides) or other intrinsic biological variability?	☐	☐	☐	☐	☐

Figure 2.2. WEAR-BOT Areas of Consideration Checklist for Validity Studies.

WEArable Technology **R**isk of **B**ias and **O**bjectivity **T**ool (WEAR-BOT) - Reliability Studies											
Authors:											
Year:											
Study Design and Methodology				Yes	Probably Yes	Probably No	No	Not Applicable	Results
1	a	Was reliability tested concurrently (using two devices at the same time) as opposed to sequentially (using two trials with one device)?		☐	☐	☐	☐	☐	
		i.	If tested concurrently, answer questions b-d.						
		ii.	If tested sequentially, answer questions e-h.						
	b	Was device placement on the participant standardized and appropriate for each device?		☐	☐	☐	☐	☐	
	c	Were the devices the exact same model?		☐	☐	☐	☐	☐	
	d	Did the devices have the same software/firmware updates?		☐	☐	☐	☐	☐	
	e	Were steps taken to ensure consistent intensity and other testing parameters (time, modality, environmental conditions, etc.) between trials?		☐	☐	☐	☐	☐	
	f	Was the software/firmware the same for each trial?		☐	☐	☐	☐	☐	
	g	Was device placement on the participant the same for each trial?		☐	☐	☐	☐	☐	
	h	Was an appropriate amount of time given between tests?		☐	☐	☐	☐	☐	
		i.	For instance, was an appropriate duration provided so the previous trial does not effect the next trial, or too long that physiological or physical traits may have changed.						Low Risk of Bias
Statistical Analysis Methods				Yes	Probably Yes	Probably No	No	Not Applicable	Results
2	Data Processing								
	a	Were the data processing methods described appropriately and in a reproducible manner?		☐	☐	☐	☐	☐	
		Was the amount of missing and/or cleaned data reported?		☐	☐	☐	☐	☐	
	b.	i.	Select N/A if data was cross-sectional (no repeated measures) or cleaning not necessary.						
		Was justification provided for any data removed?		☐	☐	☐	☐	☐	
	c.	i.	Select N/A if no data was removed or if justification was not provided.						
	d	If necessary, was it reported how missing data from the test devices and/or trials were handled?		☐	☐	☐	☐	☐	
	e	Was any software used for analysis disclosed?		☐	☐	☐	☐	☐	Low Risk of Bias
3	Statistical Tests								
	a	Were multiple measures of reliability reported?		☐	☐	☐	☐	☐	
	b.	Was a test of absolute reliability reported (e.g. coefficient of variation, standard error of measurement)?		☐	☐	☐	☐	☐	
	c.	Was a test of relative reliability reported (e.g. ICC)?		☐	☐	☐	☐	☐	
	d	Were the reliability thresholds stated?		☐	☐	☐	☐	☐	Low Risk of Bias

Figure 2.3. WEAR-BOT Checklist for Reliability Studies.

General Guidelines and Instructions for Use

This tool is intended for researchers to evaluate previously published research and the risk of bias they may contain, as it relates to validity and reliability studies using consumer-grade wearable technology. This is frequently performed as part of a systematic review. However, researchers looking to perform their own validity or reliability study may also use the checklist to ensure they are designing studies that have a minimal risk of bias. The checklist is designed to reduce bias and improve the validity of the studies, including internal and external validity, by ensuring appropriate study design and methodology, as well as statistical analysis. It can be utilized by researchers evaluating wearable technology concurrently or sequentially, thus improving the evaluation of the test devices concurrent, predictive, and/or criterion validity.

This section will provide general instructions for use, describing the overall intent of each section and some general guidelines. Detailed instructions for use can be found below this section, which will address specific questions researchers may have when utilizing the tool.

Validity Checklist

Study Design and Methodology

Subsection 1: Test Variables

This section focuses on the alignment of measurement units between the test device and the criterion, ensuring that the validity of the device is assessed against intended measures. It emphasizes the importance of appropriate study design in validation research, discouraging the testing of variables not targeted by the device manufacturers.

Subsection 2: Criterion Device

This section evaluates the selection and utilization of criterion devices, requiring evidence of validity, proper calibration, and detailed reporting on the software used for analysis. This section highlights the need for clear justification when non-standard criterion methods are employed, ensuring the criterion's relevance and validity.

Subsection 3: Test Devices

This section addresses the standardized use and reporting of test devices, including their calibration, reset procedures, input of participant demographics, and placement on participants. It aims to establish control over potential confounding variables while balancing internal and external validity, ensuring that devices are used as intended by their manufacturers.

Subsection 4: Test Protocols and Parameters

This section emphasizes the importance of controlling all appropriate factors, especially when the criterion measure and test device are not tested concurrently, but are tested sequentially. It is important to control for all possible testing parameters, where appropriate. It recommends the reporting of data collection intervals, test settings, and measures taken to control potential confounders, ensuring thorough evaluation of the test environment.

Subsection 5: Participants

This section focuses on the justification of sample size through power analysis or other means, the reporting of inclusion/exclusion criteria, and the description of sample demographics. Additionally, it calls for the identification of potential confounding variables related to participants that could influence measurements.

Statistical Analysis Methods

Subsection 6: Data Processing

This section is concerned with the transparency and reproducibility of data processing methods, reporting on missing or cleaned data, and the alignment of data from test devices with the criterion. It stresses the importance of clear methodology in data handling and processing to ensure accurate comparisons in validity studies.

Subsection 7: Statistical Tests - Continuous Variables

This section is the longest section and evaluates the appropriateness of statistical tests for continuous variables, recommending specific tests for 3 different aspects of validity, 1. error, 2. linearity, and 3. equivalence testing, while also recommending a Bland-Altman plot be generated to visually represent measurement bias. It guides the choice of tests based on data characteristics and emphasizes the importance of reporting validity thresholds and proper effect size calculations.

Subsection 8: Statistical Tests – Categorical Variables

This section ensures the use of appropriate diagnostic tests for categorical variables, such as sensitivity, specificity, and accuracy, and the reporting of classification tables. It suggests association tests suitable for nominal or ordinal variables and stresses the importance of establishing and reporting validity thresholds.

Areas of Consideration

Finally, this tool mentions additional factors that, while not directly contributing to the risk of bias calculation, may prove valuable for researchers to consider when designing their studies. These considerations are meant to further mitigate bias and enhance the validity of research involving wearable technology.

Reliability Checklist

Study Design and Methodology

This section prompts researchers to detail their approach to reliability testing, distinguishing between concurrent and sequential methodologies. It emphasizes the importance of standardizing device placement, ensuring device and software uniformity, and maintaining consistent testing parameters across trials. These elements are crucial for minimizing variability and bias, thus enhancing the reliability of study findings.

Statistical Analysis Methods

Subsection 2: Data Processing

Within this section, the tool addresses the handling and processing of data, including the description of data cleaning methods, reporting of missing or cleaned data, and the justification for any data exclusion. This ensures that the data analysis process is transparent and reproducible. Additionally, it queries the use of specific software for analysis, promoting methodological integrity.

Subsection 3: Statistical Tests:

This section encourages the reporting of multiple measures of reliability, differentiating between absolute reliability measures, such as the coefficient of variation and standard error of measurement, and relative reliability, typically assessed using the Intraclass Correlation Coefficient (ICC). This approach to reliability testing provides a thorough understanding of a device's performance, producing a more robust understanding of the device’s reliability.

Detailed Instructions for Use

Validity Checklist

Study Design and Methodology

Subsection 1: Test Variables

Question 1a: Are the units of measurement (or estimated values) between test device and criterion the same?

This question is important to determine whether the validity of the device is being tested, or simply correlation between other variables. It would be inappropriate to test whether the device can measure variables the manufacturers did not intend for it to measure in the context of a validation study. While this could be performed in an exploratory manner, the analysis would be different than in validation studies.

Subsection 2: Criterion Device

Question 2a: Is there clear evidence that the criterion device/method used is valid?

This question is important because there must be clear evidence that the device chosen for the criterion is accurate and/or reliable enough to provide the correct values. There have been several studies that use a "criterion device" that is not widely agreed upon to be accurate and/or reliable enough to be considered a criterion, which introduces bias to the study. Although at times it is clear whether a certain measure is accepted as the gold standard, and should be used as the criterion, this is not always the case. In such instances where researchers are using data collection methods that are not widely

accepted as the gold standard, it is especially important that researchers evaluate the “criterion” before use and report it in the paper. For example, body composition testing using a DEXA scanner would be widely accepted as a criterion device, but researchers who utilize bioelectric impedance analysis (BIA) may need to establish the device as an appropriate criterion for their study based on previously published data. As has been noted by previous literature, thresholds for validity and reliability are not widely established (Carrier et al., 2021), nor are thresholds for the criterion devices, so researchers may need to use their best judgement when citing a device as a “criterion”, until threshold for criterion devices can be established.

Question 2b: Was the criterion device properly calibrated, synced, and updated prior to testing?

This is an important aspect for researchers to report to ensure that there was not a systematic or random bias in the data due to improper calibration, which may cause the validity and/or reliability measures of the test device to be inaccurate due to methodological errors on the researchers’ part. In addition, appropriate syncing with devices or having updated devices for some participants rather than others may introduce differing results, and thus introduce bias into the study.

Question 2c: Was the software used for analysis reported and appropriate?

This question requires the evaluator to determine the suitability of the software used for analysis within the study. This is, in part, a judgement call performed by the evaluator to determine whether it was appropriate. This task, of assessing whether a measure or aspect of a study was appropriate, is used several times in the WEAR-BOT. Given the tool's design to accommodate a broad range of applications, we rely on those using the tool to be experts in their fields, and to use their best judgment as to whether something was appropriate, based on their experience and current best-practices.

Subsection 3: Test Devices

These questions are mainly concerned with ensuring that the use of the test devices was standardized and reported, or justification provided if standardization was deliberately not prioritized.

Question 3a: Were the test devices properly synced, updated, and/or calibrated (if applicable) prior to testing?

As with calibration of the criterion device, this step is necessary to complete and report for proper bias evaluation. Authors should describe how data alignment or syncing between the test device and the criterion was performed. Additionally, the version of the operating system or firmware used in the test devices should be reported, especially if the system was updated within the study timeframe.

Question 3b: If necessary, was the data and/or settings reset or adjusted between each test?

In many wearable devices, previous data may influence the generation of new estimates. For example, a device may use accelerometer data in conjunction with GPS data to determine the stride length of the user. If this data is not reset between tests, and especially between participants, it represents a potential confounding variable. Therefore, it is recommended to reset the settings between tests, when necessary. However, if the researchers are sure that the device does not use previous data to influence the physiologic or physical estimates, then this step is not needed and should be noted in the manuscript.

Question 3c: If necessary, were participant demographics input for each test?

This is important for similar reasons to question “b”. If the test device participant demographics for their estimates, such as bodyweight being used in the calculation of energy expenditure, then not resetting these settings between tests may alter the algorithms used by the manufacturers and create a confounding variable and ultimately misrepresentation of the validity of the devices. Again, if the

researchers are sure that demographics are not used in any calculations by the device, this step is not necessary, but should be justified in the manuscript.

Question 3d: Was device placement on the participant standardized and appropriate?

The placement of test devices should be used as the manufacturers intended, and researchers should report the anatomical attachment point in the paper. With that being said, it should be noted that devices are meant to be used by the general population, and most manufacturers allow for a level of variety in how they are used. As long as the placement and use are properly reported and in-line with manufacturer recommendations, this question should be answered "Yes". This point is assessed further in question "3e".

Question 3e: Were the devices used in a way the manufacturers would approve of?

This question assesses in general terms if the device was used appropriately (rather than specific device placement as evaluated in Question 3d). This may require those assessing the study to use their best judgement and read device manuals if there is a question regarding methodology. Similar to device placement, test devices should be used in the manner for which the device was designed, and authors should note this in the manuscript.

Question 3f: Were all devices, software, and accessories used reported?

Some wearable devices have additional accessories that can be used in conjunction with the base device to improve accuracy or broaden the number of variables it can track/estimate. These should be reported, if not, comparison across studies cannot be done appropriately. Additionally, whatever software or applications were used to collect the data should be reported, whether this be the native application for the device, or a third party application.

Subsection 4: Test Protocols and Parameters

Question 4a: Were measurements between the criterion and test device taken concurrently?

This aspect of testing is important to establish because there are additional factors that need to be controlled for when testing sequentially, including device placement, elapsed time, exercise intensity (if being used during exercise), environmental factors, among many others. While testing sequentially can be used for testing, most consumer-grade devices are relatively inexpensive, and thus testing with two devices concurrently, in addition to the criterion (three devices total), is a realistic possibility, and reduces the risk of introducing confounding variables associated with time and multiple data collections.

Question 4b: If tested sequentially, was an appropriate amount of time given between tests?

As stated above, testing sequentially requires researchers to attempt to control many more variables than if testing concurrently to ensure appropriate internal validity. An important factor is time. This is specified in the tool, with instructions stating, “Were participants provided a sufficient amount of time so the measurement would not be affected by the previous test bout, or too long that physiological or physical traits may have changed?”. The amount of time between tests will vary from study to study, and possibly across testing bouts and modalities within the same study. If testing something that does not require a lot of rest time, such as activities of daily living, or walking, minutes may be enough time between trials. However, if testing an exercise modality that requires larger effort, such as running or circuit training, minutes may not be long enough, and researchers may be looking at having hours to days between trials. In addition, waiting too long, such as weeks to months, could be too long between trials, and a person’s physiology may change based on their training status (if testing exercise modalities). This question answer should be N/A if the devices were tested concurrently, as stated in the tool.

Question 4c: Was the data collection time interval reported?

The length of time the data was collected for should be reported, whether this was five minutes or five hours. This will be especially important for compiling results from multiple studies, as MAPE from a 5-minute bout of exercise should not be weighted the same as MAPE from a 5-hour bout of exercise when compiling an overall MAPE for devices.

Question 4d: Was the test setting reported (e.g. laboratory, free-living, field)?

This question is asked to determine whether the test setting was reported and what environments the device has been tested in, and under what circumstances the device may be considered valid. Knowing the environment the device was tested in is important for contextualizing the results and understanding the conditions under which the device was evaluated. Whether the measurements were taken in a controlled laboratory environment, during free-living conditions, or in a field setting may influence the use cases of the device, limiting it to certain scenarios. By clearly stating the test setting, researchers enable a deeper understanding of the context in which the device performs as expected or reveals potential limitations, guiding both users and developers in making informed decisions about its practical applications.

Question 4e: Were steps taken to control for any potential confounding variables in the test environment?

Authors should report whether potential confounding factors, specifically in their environment, were present during testing. As testing environments may vary greatly, it will be difficult to develop a questionnaire or checklist to address every potential confounding variable. Therefore, we rely on the expertise of those using the tool when evaluating the literature to use appropriate judgement based on previous literature and their experience to identify any potential issues in the methodology of the study being evaluated.

Subsection 5: Participants

Question 5a: Did researchers perform an *a priori* power analysis or state other justifications to determine sample size?

Performing an *a priori* power analysis is important to ensure appropriate power in the study without wasting time and resources by testing too many participants. Researchers should utilize a correlation-based effect size when calculating the necessary sample size (such as Pearson's Product Moment Correlation Coefficient), as using effect sizes based on descriptive statistics (such as Cohen's D) would likely result in far too many participants being tested to achieve appropriate power in the study. If a power analysis was not performed, authors should justify the sample size tested in another manner, otherwise the evaluator can select "No" for this question.

Question 5b: Was the inclusion/exclusion criteria and sample population reported and described (e.g. age, BMI, fitness level, disease status)?

An appropriate reporting of participant demographic characteristics should be provided by the researchers. As the goal of validation and reliability studies are to determine if and when wearable devices can produce accurate results, a critical component is the demographics of the population that the device was tested on. Comparing device performance during stride length for the general population vs. individuals with a musculoskeletal disorder would be improper, thus researchers should report the aspects of their population in a thorough manner.

Question 5c: Were any potential confounding variables regarding participants identified that would influence the measurements (see limitations section)?

Authors should report whether any potential confounding factors, specifically regarding participants, were present during testing. This question relies on the researchers utilizing the tool to use their

expertise and their best judgement to identify any potential confounding variables that may be present in the unique participant pool for each specific study they assess. The rationale should be sound, cautiously measured, and be stated in the systematic review or report they produce if they identify any potential confounding variables to ensure transparency.

Statistical Analysis Methods

Subsection 6: Data Processing

This subsection is the first under the broader category of "Statistical Analysis Methods," emphasizing the critical role of proper data processing in research involving wearable technology. It aims to ensure that data processing methods are thoroughly reported and adhere to standards that allow for reproducibility. These questions collectively aim to ensure the methodological rigor and transparency of data processing in research involving wearable technologies, providing a foundation for the subsequent statistical analysis.

Question 6a: Were the data processing methods described appropriately and in a reproducible manner?

This question assesses whether the methodology section provides a detailed account of the data processing steps, and could include reporting of specific software tools, versions, and settings used. The aim is to determine if another researcher could replicate the study based on the information provided. Data cleaning is "the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset" (). Data processing is specific to the device and measurement. If the measurement is taken several times throughout the data collection period (such as heart rate or accelerometry data), there may be missing data, or data that needs to be cleaned. However, if the data is cross-sectional, and only a single value is provided by the device (such as estimated VO_{2max}), then data

cleaning is not necessary. Whether the methods were described “appropriately” will require the evaluator to make a judgement call. The researchers should examine whether the methods were detailed enough to accurately reproduce the data processing methods, and whether they believe anything was omitted from the processing methodology. It is important that the researchers using the WEAR-BOT be familiar with at least basic types of data processing that may be required when using wearable technology in research, as specific devices have specific data processing needs.

Question 6b: Was the amount of missing and/or cleaned data reported?

This question is pivotal to ensuring data processing, and specifically cleaning of the data, is done transparently and with integrity. In a validity study, removing data should be done cautiously, and with good justification to prevent biased results. Researchers should report the volume and reasons for data exclusion, such as outliers, errors in the criterion data, or other reasons. In-text instructions to select N/A applies if the data was cross-sectional or if cleaning was deemed unnecessary due to the nature of the data collection method.

Question 6c: Was it reported how missing data from the criterion and/or test devices were handled?

This question is centered on the methodology for addressing missing data, ensuring that such handling does not bias the results. It may be as simple as stating the amount of missing data from each device. The amount of missing data will be relative to the granularity of the epoch. For example, devices that aggregate sensor data every second compared to every minute will likely have more missing data. If the device outputs raw data (unusual in consumer-grade electronics), the scale of processing and missing data will be much greater. This should be noted in the report and taken into account by the researchers evaluating the study. If imputation was performed, justification must be reported, as in most cases, imputation for validity studies of consumer-grade wearable devices would be inappropriate.

Question 6d: Was justification provided for any data removed by the researchers?

This question seeks to ensure that any decision to exclude data from the analysis is transparent and justified in the report. It underscores the necessity for researchers to provide a clear and well-founded rationale for any data they decide to remove or exclude from their study. Such justifications are crucial for understanding the boundaries and conditions the device being tested can be found valid or not valid. The criteria for data removal can vary widely, from errors in the criterion device (which should be removed so the test device is not faulted due to the criterion's mistake), to errors in data collection or entry. The explicit reporting of these criteria not only enhances the study's reproducibility but also allows for a critical assessment of its findings. In-text instructions to select N/A if no data was removed or if the removal was not explicitly reported are present.

Question 6e: If necessary, was the method of aligning data reported and reasonable (e.g. aligned on timestamp, elapsed time)?

The proper alignment of data points is crucial for comparison across devices or time points. This question evaluates whether the study described how data from different sources were synchronized. Options for alignment could be either by timestamps, event markers, or elapsed time. Other methods may be acceptable if the researcher described the process, why it was chosen, and why it was deemed necessary. Ensuring that measurements are properly aligned must be done in order for the validity of the device to be properly tested. Alignment of data, however, may not be necessary for all devices. In the instance that the data is cross-sectional (e.g. total hours of sleep), alignment is not necessary as there are not repeated measures. Therefore, evaluators have in-text instructions to select N/A if alignment was not a necessary step in the data processing phase.

Question 6f: If necessary, were steps taken to ensure there was no lag in the data (or signal processing), compared to the criterion, to ensure appropriate alignment?

This question also deals with data alignment but is a recommended step to ensure any potential lags between devices were identified and corrected. Techniques such as cross-correlation analysis might be employed to verify temporal alignment. Cross correlation is a statistical method used to measure the similarity or correlation between two datasets as a function of the lag of one relative to the other. Essentially, it helps to identify the degree to which two series are correlated at different time shifts, enabling the detection of patterns or relationships that may not be immediately evident in unshifted or simultaneously collected data. If there are significant correlations in the shifted data, an offset should be used to properly align the data to account for the lag present. Again, evaluators are provided in-text instruction to select N/A if cross correlations are not necessary (if the data is cross-sectional).

Question 6g: Was the time interval (epoch) for data aggregation reported and appropriate?

This question examines whether the study clearly reported the time intervals or epochs over which data was aggregated and assesses the suitability of these intervals for the study's objectives. The granularity of the data, or the smallest time unit of aggregation, is crucial as it influences the level of detail captured and the amount of data processing needed. For example, data aggregated at shorter intervals (e.g., every 5 seconds) can capture more detailed variability but may require more extensive cleaning than data aggregated at longer intervals (e.g., every 5 minutes), which would smooth over finer variations. The choice of aggregation interval impacts not only the processing and analysis of data within the study but also the comparability and synthesis of findings across studies and devices, especially in future meta-analyses. Therefore, it's important that the selected intervals are justified in the context of the study's goals and the characteristics of the data collected. Each device and measurement may justify different epoch’s and evaluators should use their expertise in the field and common practices in previous literature to determine if the level of aggregation is appropriate for each study they evaluate.

Question 6h: Was all software, programming scripts, or other resources used for statistical analysis disclosed?

Transparency in the tools and software used for analysis is critical for reproducibility. This includes not only the names and versions of statistical packages but also any custom scripts or code developed for the study.

Subsection 7: Statistical Tests - Continuous Variables

This subsection discusses the statistical methodologies applied to continuous variables in validation studies, recommending several statistical tests and reporting methods as best practices. This detailed approach to evaluating the application of statistical tests to continuous variables ensures that the methodologies employed are rigorously scrutinized for appropriateness, comprehensiveness, and transparency, when evaluating the credibility and risk of bias of the validation study.

Question 7a: Were the test variables continuous?

This preliminary question establishes the nature of the data used by researchers, confirming that the subsequent questions are relevant to the study's statistical analysis. If the test variables are not continuous, the evaluator should go on to subsection 8, which has questions pertaining to categorical variables, and mark subsequent questions in this section as N/A.

Question 7b: Were multiple tests used to determine validity?

This question is important to understand whether a comprehensive approach was taken in analyzing the data, employing multiple statistical tests to evaluate the validity of the findings. This approach is crucial for a thorough evaluation of wearable technology. While we recommend specific tests later in the tool, this question does not require those specific tests to be performed, and evaluators may select “Yes” or

“Probably Yes” if the original researchers utilized multiple tests that they claim to be used for validity analysis.

Question 7c: Was a test of error performed (e.g., MAPE, MAE, RMSE)?

This question asks about error testing, which is among the most common aspects of validity that is currently tested in published works. While we recommend that a test of error is performed, we do not specify which error measurement should be used. While mean absolute percentage error (MAPE) is the most common and makes it easy to compare results across different variables, as percentages are widely understood metrics, some researchers may prefer root mean square error (RMSE) or mean absolute error (MAE) which both will maintain the original units of whatever estimate or calculation the devices use, or other error measurement. Ultimately, the selection of an appropriate error metric should be chosen based on the nature of the data, study protocols, and specific objectives of the study.

Question 7d: Was linearity between the test device and criterion established via correlation and/or regression?

This question addresses the fundamental aspect of establishing a linear relationship between the measurements obtained from the test device and those from the criterion device in validation studies. Linearity is crucial because it indicates that the test device can accurately reflect changes in the variable of interest across the range of measurements in a manner consistent with the criterion device. It is necessary to demonstrate that for any increase or decrease in the measured variable, the test device's response is directly proportional to that of the criterion device, without systematic overestimation or underestimation at specific ranges. Establishing linearity involves statistical methods such as correlation analysis, which assesses the strength and direction of the relationship between two variables, and regression analysis, which models the relationship between a dependent variable (test device readings) and an independent variable (criterion device readings). A strong linear relationship (e.g., high

correlation coefficient, regression line closely fitting the data points) provides evidence that the test device is capable of accurately tracking the criterion across its measurement spectrum. The question of which correlation and regression tests should be used will be specified below in subsequent questions.

Question 7e: If correlation was used, was the appropriate correlation test employed (Pearson's, Lin's Concordance, Spearman's, etc.)?

This question examines the selection of correlation tests in the study. We do not recommend a specific correlation test, as different data may require specific tests. However, some considerations as to when you would use each test can be found here. Pearson's correlation coefficient can be used for continuous variables with a normal distribution, offering straightforward linear relationship insights. It's widely used, facilitating comparisons across studies and allowing for sample size calculations with common statistical software. Spearman's rank correlation is suited for non-normally distributed data, as it assesses the relationships through ranking, thus providing a viable option for non-linear associations. Lin's Concordance Correlation Coefficient, on the other hand, is generally recommended for validation studies due to its comprehensive assessment of both precision and accuracy between two variables. This makes Lin's particularly valuable when evaluating the agreement between a test device and a criterion standard, capturing the essence of variability in measurement. The choice between these tests hinges on the data's distribution and the study's specific needs. Justification should be provided by the original researchers as to why they used certain tests, and the evaluators best judgement should be used to determine if it was appropriate for the study being evaluated.

Question 7f: If necessary, was repeated measures correlation determined if there were non-independent samples?

This question is important and often overlooked in validation studies, which is why it is further specified from the previous question. The aim of this question is to address the analysis of data from studies with

measurements that are not independent, typically seen with repeated measures on the same subjects. Traditional correlation assessments may not accurately depict the relationship between variables due to the interrelated nature of these data points. The use of repeated measures correlation tests, such as the repeated measures correlation or an intraclass correlation coefficient (ICC) designed to handle multiple measurements is necessary to properly evaluate linearity in validation studies with repeated measures.

Question 7g: If regression was used, was the appropriate regression analysis utilized (e.g., Simple Linear, Deming, Passing-Bablok)?

This question evaluates the tests for linearity, specifically different regression models. Choosing the right type of regression analysis is important to accurately assess the relationship between the measurements obtained from the test device and those from the criterion device. While we do not recommend specific tests in the WEAR-BOT checklist, the reader can find brief explanations of when to use different models for validity testing. Simple linear regression is the most straightforward approach, modeling the relationship between a single independent variable and a dependent variable by fitting a straight line through the data points. This is the most widely known form of regression, and due to its simplicity, the most digestible for the reader. However, it assumes that the independent variable (criterion device measurements) is measured without error, which may not always be the case in validation studies where both devices could have measurement errors. Deming regression, also known as errors-in-variables regression, extends beyond simple linear regression by accounting for measurement errors in both the test and criterion devices. This method adjusts the regression line based on the ratio of the variances of the measurement errors, offering a more accurate estimation of the relationship when both variables have associated uncertainties. Deming regression is generally the preferred model for validity studies where the data is normally distributed. Passing-Bablok regression is a non-parametric approach that, like Deming regression, does not assume one of the variables to be error-free. It is robust against outliers and does not require the distribution of measurement errors to be normal, making it suitable for a wide

range of data types. Therefore, Passing-Bablok will be the better regression model if the data is not normally distributed. Taking into account these considerations will enable the researchers to utilize the correct regression model and allow evaluators to properly assess the risk of bias in the statistical methods of the studies in question.

Question 7h: If regression was used, were appropriate model fit statistics reported?

This question evaluates if appropriate results were reported from the regression model used. In validation studies where regression analysis is employed to examine the relationship between a test device and a criterion standard, reporting model performance is recommended for a comprehensive understanding of the model's fit and predictive accuracy. The coefficient of determination (R^2), residual sum of squares, y-intercept, and slope of the regression line is appropriate for simple linear regression, while Deming and Passing-Bablok regression should report the y-intercept and slope, as they do not produce a true R^2 value. Together, these metrics provide a detailed account of the linear relationship, allowing for an evaluation of how well the test device's measurements align with those of the criterion across the range of values tested. Reporting these metrics will allow the reader to better understand the amount of linearity between the devices, and will allow for comparisons between studies in future meta-analyses.

Question 7i: Was equivalence testing performed (e.g., TOST test, confidence interval for difference in means)?

This question investigates whether the study included equivalence testing to statistically determine if the test device's measurements are acceptably close to those of the criterion device. Equivalence testing, such as the Two One-Sided T-Tests (TOST test) and analysis using confidence intervals for the difference in means, is another method to determine validity (or equivalence) in validation studies. These can be used to establish that the test device's measurements are practically equivalent to a criterion standard

within a pre-defined margin. Unlike traditional hypothesis tests aiming to find significant differences, equivalence testing flips the null hypothesis and verifies that any deviations between devices are within acceptable limits. The TOST test procedure, for instance, checks if differences fall within specified equivalence bounds, offering a stringent criterion for method validation. This ensures the test device performs closely to the standard, supporting its use for the intended applications with confidence. This is particularly relevant for validation studies aiming to establish that two measurement methods agree within a tolerable margin of error.

Question 7j: Was bias/agreement plotted via a Bland-Altman plot?

This question simply asks if the researchers utilized Bland-Altman plots for assessing agreement/bias between the test device and the criterion. This graphical method is a widely utilized method for identifying any systematic bias and the limits of agreement in validation studies. By providing a visual representation of how the differences between the two measurement methods vary across the range of measurements, Bland-Altman plots facilitate the identification of any systematic bias or trends, such as a tendency for differences to increase as the magnitude of the measurement increases.

Question 7k: If performed, were bias and limits of agreement estimates reported for the Bland-Altman analysis?

This question is to evaluate whether, in addition to plotting, the study reports quantitative estimates of bias (average difference) and limits of agreement as calculated from the Bland-Altman analysis, providing a clear indication of the test device's accuracy and consistency. Reporting this is valuable to the readers and may be used in the future for comparisons between devices or modalities.

Question 7l: If effect size was calculated, was it based on linear association (e.g., R^2) rather than less appropriate calculations using descriptive statistics (e.g., Cohen's D)?

This question ensures the appropriate use of effect sizes for studies performing validity testing, such as those based on linear associations, rather than those more suited to comparing group means. Utilizing effect sized based on descriptive statistics (group means) will produce small effect sizes for most validation studies, as the goal of the test device is to be as close to the criterion in its measurements as possible. Therefore, it would be inappropriate to use effect sizes based on descriptive statistics and association-based effect sizes should be utilized. However, if the effect size based on descriptive statistics was not utilized in the interpretation of the validity of the device, that would not introduce bias into the study. Therefore, we include in-text instructions that state, "Select N/A if effect size was not calculated or effect size based on descriptive statistics was not used in the interpretation of the validity of the device (if it was calculated but not used to determine if validity was achieved)."

Question 7m: Were validity thresholds reported?

This question gets to the very heart of validity studies, to answer the question of whether a device was valid or not. As thresholds for validity have not been widely established (as of the publication of this paper), it is left up to the individual researchers to determine whether the device meets their standards. There have been several authors who propose varying thresholds for validity, some more conservative, and others more liberal. Whatever thresholds the researcher chooses should be established prior to data collection and reported in the published work.

Subsection 8: Statistical Tests - Categorical Variables

This subsection addresses the application and analysis of categorical variables within the context of consumer-grade wearable device validation studies. It emphasizes the importance of selecting appropriate statistical methodologies for analyzing categorical data, ensuring the validity and reliability of the devices under study.

Question 8a: Were the test variables categorical?

This question serves as a preliminary filter, confirming whether the data analyzed in this section is indeed categorical. If the data is not categorical, in-text instructions direct the reader to fill out the previous section on continuous variables and select “Not Applicable” (N/A) for all subsequent section 8 questions.

Question 8b: Were diagnostic tests performed to assess predictive validity (e.g. sensitivity, specificity, accuracy, AUC)?

This question evaluates the use of diagnostic accuracy tests to determine how well the device can correctly classify or predict outcomes compared to a criterion standard. This includes assessing whether measures such as sensitivity (true positive rate), specificity (true negative rate), overall accuracy, and area under the receiver operator curve (AUC) were calculated and reported, providing insight into the device's performance in categorical terms. This could be used for human activity recognition, where devices are attempting to predict what activity is being performed (e.g. walking, washing dishes), or classifying exercise intensity into light, moderate, and vigorous intensity exercise based on metabolic equivalents (METs), or other categories. Overall, these tests are fundamental in evaluating the predictive validity of a test device in classifying or predicting categories against a criterion standard.

Question 8c: Was the classification table (confusion matrix) reported?

This question simply seeks to confirm that the study provided a confusion matrix, detailing the number of true positives, true negatives, false positives, and false negatives. This matrix is an important reporting metric for understanding the device's classification accuracy and for calculating the diagnostic tests mentioned in question 8b, as well as being an avenue to improve transparency in the results.

Question 8d: Was association between the test device and criterion established with appropriate categorical correlation statistics (e.g., Cohen's Kappa, Cramer's V, tetrachoric [for nominal variables], rank-based correlations, polychoric [for ordinal variables])?

This question attempts to ensure that appropriate association statistics were run for the categorical variables, as correlation tests for continuous variables (such as Pearson's) are inappropriate to use for categorical variables. In validation studies, establishing the association between the test device and the criterion standard requires selecting the appropriate correlation statistics tailored to the data's nature. Cohen's Kappa is a robust measure used to assess the agreement between two raters or methods categorizing data into nominal categories, correcting for chance agreement. It is particularly useful when the categories are mutually exclusive and exhaustive. Cramer's V expands this concept to cases with more than two categories, providing a measure of association between nominal variables. For data that fall into ordered categories, rank-based correlations like Spearman's rho can be applied to assess the relationship between two variables. When the data is dichotomous or ordinal but assumed to follow an underlying continuous distribution, tetrachoric (for dichotomous variables) and polychoric (for ordinal variables) correlations are preferred as they estimate the Pearson correlation coefficient that would have been obtained if the underlying continuous variables were observed. These statistics can be used in validation studies for assessing the strength of the association between categorical outcomes measured by the test device and the criterion, ensuring that the chosen method aligns with the data's characteristics and the study's objectives.

Question 8e: Were validity thresholds reported?

As stated previously in question 7n, this question gets to the very heart of validity studies, to answer the question of whether a device was valid or not. As thresholds for validity have not been widely established (as of the publication of this paper), it is left up to the individual researchers to determine

whether the device meets their standards. There have been several authors who propose varying thresholds for validity, some more conservative, and others more liberal. Whatever thresholds the researcher chooses should be established prior to data collection and reported in the published work. Developing thresholds for categorical variables may pose more challenges than for continuous variables, and there have been few thresholds suggested based on diagnostic tests. Fortunately for evaluators, there is no judgement call to be made whether the researchers thresholds were appropriate, but rather whether they were reported. As widely accepted thresholds are developed, this tool may need to change to reflect the updated practices.

Areas of Consideration

This section addresses additional factors that, while not directly affecting the risk of bias score, may provide context for interpreting the validity and reliability of wearable device studies. These considerations encompass inference testing, environmental factors, and participant biological variability, offering insights into the risk of bias of the studies. These areas were of concern to many of the authors, but consensus was not able to be reached by the entire group to include it into the risk of bias tool calculations, therefore we offer it here, as areas of consideration.

I. Inference Testing

Question Ia: Were any tests of mean differences performed that are unable to determine the validity of the test devices (e.g., ANOVA, t-test)?

This question probes the use of inferential statistical tests designed to compare group means, such as ANOVA or t-tests, which are not directly applicable for validating a device's measurements.

Question Ib: If yes to question Ia, were the results of the inference test(s) used in the interpretation of the validity of the device?

This question inquires whether the outcomes of these hypothesis tests were used to draw conclusions about the device's validity, despite their inherent limitations for this purpose. This question highlights the importance of distinguishing between statistical significance and practical relevance in the context of device validation. Simply performing these tests (maybe because a reviewer requested it) will not inherently increase bias in the paper, but using them in the interpretation of the validity of the device is surely a poor choice.

II. Environmental Factors

Question IIa: Were environmental factors reported (e.g., temperature, humidity, altitude)?

This question assesses the reporting of environmental conditions during data collection, recognizing their potential impact on the performance of wearable devices. Detailed reporting of such factors may help identify possible confounding variables that could influence the validity of the test device under certain circumstances. In any case, understanding the conditions under which the device was validated can only improve the strength of the paper guide future generalization of the results.

III. Participant Biological Variability

Question IIIa: Were any steps taken to assess or control for participant biological variability, such as potential bilateral asymmetries in participants (differences between left and right sides) or other intrinsic biological variability?

This question is meant to evaluate whether the study being evaluated accounted for biological variability among participants that could potentially affect the measurement accuracy of the device. Due to the variability inherent in this aspect, recommendations for best practices in validity studies would be

difficult to establish. Therefore, it remains in the “Areas of Consideration”, and if further research establishes how to deal with participant biological variability, this tool may need to be updated to reflect the recommendations of researchers.

Reliability Checklist

Study Design and Methodology

Question 1a: Was reliability tested concurrently (using two devices at the same time) as opposed to sequentially (using two trials with one device)?

This question determines which approach to reliability testing was used, concurrent or sequential. Concurrent testing assesses reliability by using two devices simultaneously on a subject, whereas sequential testing uses the same device across multiple trials. This distinction is necessary for understanding the context in which reliability is assessed and directs the subsequent focus of the evaluation. Evaluators should answer “Yes” if there is clear evidence that the devices being tested concurrently were the exact same model (i.e., it is explicitly stated in the manuscript) or answer “Probably Yes” if there are indications the devices were the exact same model but it was not explicitly stated in the manuscript.

In-text instructions direct evaluators to answer specific questions, based on which testing methodology was used. Concurrent testing methodologies answer questions b-d, while sequential testing methodologies answer question b, and e-h.

Question 1b: Was device placement on the participant standardized and appropriate for each device?

This question checks to see if the placement of devices during testing was consistent and according to approved manufacturer protocols, ensuring that data comparability is not compromised by variations in

device positioning. As stated earlier, these devices should be used as the manufacturers designed them to be used, and to ensure that they were, researchers should report how they were used in their papers. With that being said, it should be noted that these devices are meant to be used by the general population, and most manufacturers allow for some level of variety in how they are used. As long as the placement and use are reported and in-line with manufacturer recommendations, this question should be answered "Yes". However, device placement standardization is particularly important for reliability testing, because as many variables as possible need to be the same between trials, or for each device. Therefore, device placement must be standardized and appropriate in reliability studies.

Question 1c: Were the devices the exact same model?

This question confirms that the devices used for concurrent testing were of the same make and model, eliminating variability that could arise from hardware differences. If the devices were not the exact same make and model, then that introduces serious bias into the reliability study.

Question 1d: Did the devices have the same software/firmware updates?

This question verifies that both devices were operating on the same software or firmware version, ensuring that any differences observed are not due to discrepancies in software functionality. As with many of the questions for the reliability testing section, researchers need to ensure that as many variables as possible are consistent between devices when testing concurrently.

Question 1e: Were steps taken to ensure consistent intensity and other testing parameters between trials?

This question examines whether measures were in place to maintain uniform testing conditions across trials, such as exercise intensity, environmental conditions, activity modality, among other testing parameters, to ensure the reliability of results. As the testing environment, activity, and other aspects of

the reliability studies can vary widely, evaluators must use their best judgement to decide whether enough effort was taken to control for potential confounding variables to ensure successful tests.

Question 1f: Was the software/firmware the same for each trial?

This question ensures that no software or firmware updates occurred between trials that could influence the comparability of data collected sequentially. This is particularly important if a significant amount of time has passed since the initial trial.

Question 1g: Was device placement on the participant the same for each trial?

This question checks to see if the placement of device during testing was consistent between trials and according to approved manufacturer protocols, ensuring that data comparability is not compromised by variations in device positioning. As stated earlier, these devices should be used as the manufacturers designed them to be used, and to ensure that they were, researchers should report how they were used in their papers. With that being said, it should be noted that these devices are meant to be used by the general population, and most manufacturers allow for some level of variety in how they are used. As long as the placement and use are properly reported and in-line with manufacturer recommendations, this question should be answered “Yes”. However, device placement standardization is particularly important for reliability testing, because as many variables as possible need to be the same between trials, or for each device. Therefore, device placement must be standardized and appropriate in reliability studies.

Question 1h: Was an appropriate amount of time given between tests?

This question examines the scheduling of sequential trials within a study, assessing whether the interval between them was appropriately chosen to mitigate carryover effects from prior activities while also being close enough to prevent any significant alterations in participants' physiological or physical states.

The optimal timing between tests is crucial to ensure that each measurement reflects the intended conditions without interference from previous tests or natural variations in participants' health or performance over time. Evaluators must consider several factors, including the intensity of any exercise or activity involved, the total time commitment required from participants, and the nature of the measurements being taken, among other potential variables. For example, high-intensity activities may necessitate longer recovery periods to return to baseline conditions, while assessments of more stable physiological markers might allow for shorter intervals. The appropriateness of the time interval is thus contingent upon a nuanced understanding of the study's design and objectives, requiring evaluators to apply their expertise and knowledge of the field to determine whether the chosen intervals were suitable for the study's aims.

Statistical Analysis Methods

Subsection 2: Data Processing

Question 2a: Were the data processing methods described appropriately and in a reproducible manner?

This question assesses whether the methodology section provides a detailed account of the data processing steps, and could include reporting of specific software tools, versions, and settings used. The aim is to determine if another researcher could replicate the study based on the information provided. Data cleaning is “the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset” (Tableau.com, nd). Data processing is specific to the device and measurement. If the measurement is taken several times throughout the data collection period (such as heart rate or accelerometry data), there may be missing data, or data that needs to be cleaned. However, if the data is cross-sectional, and only a single value is provided by the device (such as estimated VO_{2max}), then data cleaning is not necessary. Whether the methods were described

“appropriately” will require the evaluator to make a judgement call. The researchers should examine whether enough data was provided to accurately reproduce the data processing methods, and whether they believe anything was omitted from the processing methodology. It is important that the researchers using the WEAR-BOT be familiar with at least basic types of data processing that may be required when using wearable technology in research, as specific devices have specific data processing needs.

Question 2b: Was the amount of missing and/or cleaned data reported?

This question is pivotal to ensuring data processing, and specifically cleaning of the data, is done transparently and with integrity. In a reliability study, removing data should be done cautiously, and with good justification to prevent biased results. Researchers should report the volume and reasons for data exclusion, such as outliers, errors in the criterion data, or other reasons. In-text instructions to select N/A applies if the data was cross-sectional or if cleaning was deemed unnecessary due to the nature of the data collection method.

Question 2c: Was justification provided for any data removed?

This question seeks to ensure that any decision to exclude data from the analysis is transparent and justified in the report. It underscores the necessity for researchers to provide a clear and well-founded rationale for any data they decide to remove or exclude from their study. Such justifications are crucial for understanding the boundaries and conditions the device being tested can be found reliable or not reliable. The criteria for data removal can vary widely, but the explicit reporting of these criteria not only enhances the study's reproducibility but also allows for a critical assessment of its findings. In-text instructions to select N/A if no data was removed or if the removal was not explicitly reported are present.

Question 2d: If necessary, was it reported how missing data from the test devices and/or trials were handled?

This question probes the transparency and methodological rigor with which a study addresses the possible issue of missing data, a common challenge in research involving wearable technology. It assesses whether the researchers provided a clear account of the approaches used to manage gaps in the data, which could range from sophisticated imputation techniques that estimate missing values based on available information to straightforward exclusion criteria that remove incomplete observations from the analysis. The chosen strategy for handling missing data is pivotal, as it can significantly influence the study's findings and their reliability. Imputation is generally not recommended in reliability studies examining consumer-grade wearable technology, so if it was performed, good justification must be provided. By detailing these methods, a study ensures that other researchers can accurately replicate the analysis and assess the robustness of the conclusions drawn, thereby enhancing the credibility of the research.

Question 2e: Was any software used for analysis disclosed?

This question simply checks if the study provided detailed information on the software tools and versions used for data analysis, promoting transparency and reproducibility. Not doing so would introduce a risk of bias.

Subsection 3: Statistical Tests

Question 3a: Were multiple measures of reliability reported?

This question scrutinizes the depth of the reliability analysis conducted in the study by inquiring if a range of statistical measures were utilized to assess the reliability of a device. This question does not

recommend which tests should be performed, but is to ensure a complete assessment of the reliability was performed. A singular measure, while informative, might not fully capture the nuances of a device's performance across different conditions and metrics. For example, reporting both the Intraclass Correlation Coefficient (ICC) for relative reliability and the Standard Error of Measurement (SEM) for absolute reliability provides a more rounded view of the device's consistency and the precision of individual scores, respectively.

Question 3b: Was a test of absolute reliability reported (e.g. coefficient of variation, standard error of measurement)?

This question delves into whether the study reported measures of absolute reliability. Some examples of tests being the coefficient of variation (CV) and the standard error of measurement (SEM). Absolute reliability refers to the degree to which repeated measurements vary for individuals, emphasizing the importance of understanding the inherent measurement error and its impact on the precision of the device. The SEM provides a direct measure of this error in the same units as the measurements themselves, offering a clear indication of the expected range within which a measurement might vary due to random error. A device that is perfectly reliable would have a SEM of 0. On the other hand, the CV is the ratio of the standard deviation to the mean, and expresses the reliability of a device as a percentage, providing an easily understandable metric of reliability that can be used across measurements without the need for conversions. These tests provide context of the precision of the device in question. By reporting these measures of absolute reliability, researchers provide the readers with the necessary information to assess the reliability of the device they are testing.

Question 3c: Was a test of relative reliability reported (e.g. ICC)?

This question simply assesses whether the study reported a measure of relative reliability, possibly through the use of the Intraclass Correlation Coefficient (ICC). Relative reliability refers to the degree to

which individuals maintain their position in a sample over repeated measurements under varying conditions, highlighting the consistency and reproducibility of the measurements. The ICC is a versatile statistical tool used to evaluate this aspect of reliability, offering a quantifiable measure of the correlation between measurements taken at different times or under different conditions. By reporting a measure of relative reliability, researchers provide the readers with necessary information to assess the reliability of the device they are testing.

Question 3d: Were the reliability thresholds stated?

This question gets to the very heart of reliability studies, to answer the question of whether a device was reliable or not. As thresholds for reliability have not been widely established (as of the publication of this paper), it is left up to the individual researchers to determine whether the device meets their standards. There have been several authors who propose varying thresholds for reliability, some more conservative, and others more liberal. Whatever thresholds the researcher chooses should be established prior to data collection and reported in the published work.

Acknowledgements

Generative AI (OpenAI, ChatGPT 4.0) was used to assist in the writing of this manuscript. All aspects were reviewed and approved by the authors and are the result of the combined efforts of humans and computers working together.

Chapter 2 References

Aromataris, E., Fernandez, R., Godfrey, C. M., Holly, C., Khalil, H., & Tungpunkom, P. (2015). Summarizing systematic reviews: methodological development, conduct and reporting of an umbrella review approach. *JBI Evidence Implementation, 13*(3), 132-140.

Barker, T. H., Stone, J. C., Sears, K., Klugar, M., Tufanaru, C., Leonardi-Bee, J., Aromataris, E., & Munn, Z. (2023). The revised JBI critical appraisal tool for the assessment of risk of bias for randomized controlled trials. *JBI Evidence Synthesis, 21*(3), 494-506.

BUNN, J. A., Navalta, J. W., Fountaine, C. J., & REECE, J. D. (2018). Current state of commercial wearable technology in physical activity monitoring 2015–2017. *International Journal of Exercise Science, 11*(7), 503.

Campbell, J. M., Kulgar, M., Ding, S., Carmody, D. P., Hakonsen, S. J., Jadotte, Y. T., & Ws, C. (2020). Diagnostic test accuracy systematic reviews. *JBI Manual for Evidence Synthesis, 1*

Carrier, B., Barrios, B., Jolley, B. D., & Navalta, J. W. (2020). Validity and Reliability of Physiological Data in Applied Settings Measured by Wearable Technology: A Rapid Systematic Review. *Technologies, 8*(4), 70.

Carrier, B., Salatto, R. W., Davis, D. W., Sertic, J. V. L., Barrios, B., Cater, P., & Navalta, J. W. (2021). Assessing the Validity of Several Heart Rate Monitors in Wearable Technology While Mountain Biking. Paper presented at the *, 14*(1) 18.

Evenson, K. R., Goto, M. M., & Furberg, R. D. (2015). Systematic review of the validity and reliability of consumer-wearable activity trackers.(Report). *The International Journal of Behavioral Nutrition and Physical Activity, 12*(161)10.1186/s12966-015-0314-1

Gagnier, J. J., Lai, J., Mokkink, L. B., & Terwee, C. B. (2021). COSMIN reporting guideline for studies on measurement properties of patient-reported outcome measures. *Quality of Life Research, 30*, 2197-2218.

Keadle, S. K., Lyden, K. A., Strath, S. J., Staudenmayer, J. W., & Freedson, P. S. (2019). A framework to evaluate devices that assess physical behavior. *Exercise and Sport Sciences Reviews, 47*(4), 206-214.

Mokkink, L. B., Terwee, C. B., Knol, D. L., Stratford, P. W., Alonso, J., Patrick, D. L., Bouter, L. M., & De Vet, H. C. (2006). Protocol of the COSMIN study: COnsensus-based Standards for the selection of health Measurement INstruments. *BMC Medical Research Methodology, 6*, 1-7.

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., Bouter, L. M., & De Vet, H. C. (2010). The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Quality of Life Research, 19*, 539-549.

Munn, Z., Barker, T. H., Moola, S., Tufanaru, C., Stern, C., McArthur, A., Stephenson, M., & Aromataris, E. (2020). Methodological quality of case series studies: an introduction to the JBI critical appraisal tool. *JBI Evidence Synthesis, 18*(10), 2127-2133.

Patel, V., Orchanian-Cheff, A., & Wu, R. (2021). Evaluating the validity and utility of wearable technology for continuously monitoring patients in a hospital setting: systematic review. *JMIR mHealth and uHealth, 9*(8), e17411.

Prinsen, C. A., Mokkink, L. B., Bouter, L. M., Alonso, J., Patrick, D. L., De Vet, H. C., & Terwee, C. B. (2018). COSMIN guideline for systematic reviews of patient-reported outcome measures. *Quality of Life Research, 27*, 1147-1157.

Sterne, J. A., Hernán, M. A., Reeves, B. C., Savović, J., Berkman, N. D., Viswanathan, M., Henry, D., Altman, D. G., Ansari, M. T., & Boutron, I. (2016). ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *Bmj, 355*

Sterne, J. A., Savović, J., Page, M. J., Elbers, R. G., Blencowe, N. S., Boutron, I., Cates, C. J., Cheng, H., Corbett, M. S., & Eldridge, S. M. (2019). RoB 2: a revised tool for assessing risk of bias in randomised trials. *Bmj, 366*

Tableau.com.*Guide To Data Cleaning: Definition, Benefits, Components, And How To Clean Your Data.* Retrieved 03/03/2024, from https://www.tableau.com/learn/articles/what-is-data-cleaning

van Lier, H.,G., Pieterse, M. E., Garde, A., Postel, M. G., de Haan, H.,A., Vollenbroek-Hutten, M., Schraagen, J. M., & Noordzij, M. L. (2020). A standardized validity assessment protocol for physiological signals from wearable technology: Methodological underpinnings and an application to the E4 biosensor. *Behavior Research Methods, 52*, 607-629.

Welk, G. J., Bai, Y., Lee, J., Godino, J., Saint-Maurice, P. F., & Carr, L. (2019). Standardizing analytic methods and reporting in activity monitor validation studies. *Medicine and Science in Sports and Exercise, 51*(8), 1767.

Welk, G. J., McClain, J., & Ainsworth, B. E. (2012). Protocols for evaluating equivalency of accelerometry-based activity monitors. *Medicine and Science in Sports and Exercise, 44*(1 Suppl 1), 39.

Chapter 3 - The Risk of Bias in Validity and Reliability Studies Testing Physiological Variables using Consumer-Grade Wearable Technology: A Systematic Review and Meta-Analysis with WEAR-BOT Analysis

Abstract

INTRODUCTION: Wearable technology is a quickly evolving field, and new devices with new features to measure/estimate physiological variables are being released constantly. This technology is being used by recreational athletes, coaches, collegiate and professional athletes, military personnel, and researchers to quantify physiological variables during sport and exercise. Despite their use, the validity of the devices are largely unknown to the users or researchers, and the quality of the studies that do test validity and reliability vary widely.

PURPOSE: Therefore, the purpose of this systematic review and meta-analysis was to review the current validity and reliability literature concerning consumer-grade wearable technology measurements/estimates of physiological variables (e.g. heart rate, energy expenditure, etc.) during exercise. Additionally, we sought to perform risk of bias assessments utilizing the novel **WEA**rable technology **R**isk of **B**ias and **O**bjectivity **T**ool (WEAR-BOT), and perform meta-analytic calculations on the reported data.

METHODS: This review was conducted following PRISMA guidelines, searching three databases: Google Scholar, Scopus, and SPORTDiscus. Papers published between Jan 2020 and April 2023 were evaluated. After screening, 46 papers were identified that met the pre-determined criteria. Then data was extracted and risk of bias assessment performed by independent researchers. Descriptive statistics were calculated to describe the studies and their results, including counts of devices, exercise modalities, and statistical tests. Weighted averages of mean absolute percentage error (MAPE) and Pearson correlations were

calculated, weighted by sample size. Sample size statistics were performed utilizing the lower 95% confidence interval of the weighted correlation average.

RESULTS: Of the 46 papers reviewed, 44 performed validity testing, while nine performed reliability. Seventy different devices were evaluated across 34 manufacturers. The weighted average for MAPE was 12.48% for heart rate (HR) and 30.70% for energy expenditure (EE). The weighted average for Pearson correlations was 0.737 for HR and 0.672 for EE. Heart rate was the most common variable tested, with EE being second most. Walking, then running, then cycling were the three most common exercise modalities. Risk of bias assessment of validity studies resulted in 30/44 studies being classified as having a “High Risk of Bias”, and 14/44 having “Some Risk of Bias”. None had a “Low Risk of Bias”, according to the novel WEAR-BOT. For reliability studies, 7/9 were classified as “High Risk of Bias”, 2 as “Some Risk of Bias”, and 0 as “Low Risk of Bias”.

CONCLUSION: The risk of bias assessment and descriptive statistics paint a troubling picture of the overall state of validity and reliability studies. Statistical analyses, methods, and reporting vary excessively, as can be expected of an emerging field. This review and associated WEAR-BOT analysis can be used by researchers to help standardize methodology, analytics, and reporting of validation and reliability studies of consumer-grade wearable technology.

Introduction

As wearable technology continues to grow in popularity, use, and sophistication, validity and reliability studies have sought to determine just how accurate and reliable these devices are (Carrier et al., 2020a; Evenson et al., 2015; Fuller et al., 2020; Patel et al., 2021). This type of research into all wearable technology, but especially consumer-grade wearable technology, is important so that users can determine if the results can be trusted. These devices can be used by researchers, athletes, coaches, and the general population to help improve fitness metrics and better understand a person's physiology during sport or exercise. Wearable technology represents an untapped wealth of data regarding human physiology in real-world settings that can be utilized to improve our understanding of human behavior, physical activity, and physiology (Wright et al., 2017). However, without an understanding of the overall accuracy and reliability of these devices, use in research, athletics, military settings, or even recreational applications should be done cautiously, and possibly not at all.

As with any emerging field, there are growing pains in the beginning, as best practices are yet to be established. That, unfortunately, includes the validation and reliability studies done up to this point. These studies can vary widely in their methods, and especially their statistical analyses (Carrier et al., 2020a; Welk et al., 2019). Authors have proposed statistical tests to standardize the analytics when validating new devices, such as a test of error, test of linearity, and a test of equivalence, as well as bias graphically represented via a Bland-Altman plot (Carrier et al., 2020a; van Lier et al., 2020; Welk et al., 2019). As a result, we have seen improved consistency of analytics in papers, which has also allowed for greater comparisons across studies.

While the number of these types of studies have increased, systematic reviews for these studies have increased proportionately. However, when authors of systematic reviews for wearable technology attempt to evaluate the risk of bias in the studies they are analyzing – which is recommended practice

when performing systematic reviews – some authors note the inability to do so, as there is not a sufficient tool for this type of research (Carrier et al., 2020a). A risk of bias analysis is generally a checklist or list of questions the researcher will answer for the study in question, and once filled out, a ranking of low risk, some risk, or high risk of bias is generally provided. This is useful for those reading the systematic review, as they can quickly see the quality of the research done prior, and what level of bias there may be in that research. Additionally, risk of bias checklists can provide guidance to researchers as they seek to perform their own research. It can identify several aspects they should seek to include, report, or control for, in their experiments and data collections to minimize the risk of biasing their own research.

When performing systematic reviews and risk of bias analyses on wearable technology literature, some researchers will use a risk of bias tool not designed for this type of research, such as the Cochrane Risk of Bias (RoB) 2.0, Joanna Briggs risk of bias tools, or none at all (BUNN et al., 2018; Carrier et al., 2020a; Evenson et al., 2015; Volkova et al., 2023). They may even note their inability to properly evaluate the risk of bias in the published studies (Carrier et al., 2020a). Recently, authors have used a portion of the COSMIN risk of bias checklist, as it is somewhat more appropriate for evaluating these studies (Patel et al., 2021; Prill et al., 2021). Fortunately, a collaborative effort involving experts from multiple Universities and industry, including several authors of the current review with expertise in wearable technology testing, has led to the development of a risk of bias tool tailored for assessing studies on the validity or reliability of consumer-grade wearable technology. This tool, known as the **WEA**rable technology **R**isk of **B**ias and **O**bjectivity **T**ool (WEAR-BOT), is now established and ready for implementation (citation pending, awaiting publication).

As current wearable technology is a quickly evolving field, testing the validity and reliability of these devices is difficult to keep up with. Studies are continually being published as devices evolve. Thus, it is important to review the most current literature to evaluate the validity of the devices, and the practices

of the researchers. While wearable technology can measure a myriad of variables, both physiologic and physical, of specific interest to the authors of the current paper, is physiological variables. This systematic review is unique, because it will be the first ever systematic review to make use of the novel risk of bias tool for consumer-grade wearable technology. This will provide the clearest picture yet of the state of the literature, and the risk of bias that is in the validity and reliability studies. Therefore, it is the aim of this paper to perform a systematic review and meta-analysis to describe the methods and results of the wearable technology literature evaluating validity and reliability. In addition, we seek to determine the risk of bias of each published paper, using the novel risk of bias tool, the WEAR-BOT.

Methods

This review was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Page et al., 2021). Search and screening occurred between March 2023 and July 2023.

Inclusion/Exclusion Criteria

The inclusion/exclusion criteria for the papers were as follows: 1). Validity or reliability of consumer-grade wearable technology needed to be tested. 2). The measurements from the wearable technology needed to be regarding an individual's physiology and taken during sport or exercise. 3). Healthy or apparently healthy individuals were the population of interest. 4). The article needed to be peer reviewed, available in English, and published between January 2020 and April 2023. For the purposes of this review, consumer-grade wearable technology was loosely defined as, marketed to multiple groups of the population (athletes, general population, etc.) and was currently available to purchase online. If

there were any disagreements between members as to whether a device should be considered "consumer-grade", it was resolved through discussion with the research group. We adopted the American College of Sports Medicine definition of exercise for this review, namely "Exercise is a type of physical activity consisting of planned, structured, and repetitive bodily movement done to improve and/or maintain one or more components of physical fitness" (American College of Sports Medicine, 2013). Studies that examined only activities of daily living were not included, but if exercise was specifically analyzed in addition to activities of daily living, it was included. Walking was considered a type of exercise and included if only walking was performed in the study. Only physiological variables were considered (heart rate, core body temperature, energy expenditure, etc.), while physical variables were not (steps, repetitions, speed, vertical oscillation, etc.). The timeline going back to 2020 was chosen to represent the most current literature, while limiting the scope of the review to a manageable number of papers.

Search Strategy

Researchers searched three databases for this review. The included databases were Google Scholar, Scopus, and SPORTDiscus. Google Scholar was accessed directly through the website https://scholar.google.com/ (Google LLC., Mountain View, CA, USA). Scopus and SPORTDiscus databases were both accessed through the researcher's institution's library website. The same search combination was used for each database, which was: "wearable AND technology OR tracker OR monitor AND exercise OR fitness OR activity AND validity OR accuracy OR reliability". Search results were verified by the researchers prior to beginning the screening process, and results were the same for Google Scholar and Scopus databases, while SPORTDiscus had slightly different results for researchers depending on the date of the search (exact numbers below). Filters were applied to each database after pilot searching and

discussion with the research group to determine a sufficient scope for the review. Google Scholar had one filter applied, a time filter using the “Custom range” option, and 2020 – present was input as the timeline (second date for range left blank). Scopus filters had time, document type, publication stage, and language filters applied (Time: 2020 - Present; Document Type: Article; Publication Stage: Final; Language: English). SPORTDiscus had time, peer review status, publication type, and language filters applied (Time: 2020 - Present; Peer Reviewed; Publication Type: Academic Journal; Language: English). The search results for Google Scholar were 85,700 articles on April 6, 2023. However, Google Scholar limits the accessible search results to 1,000 articles, therefore, that was the number of articles screened from Google Scholar. The search results for Scopus were 800 articles on April 13, 2023. Search results for SPORTDiscus were 11,599 and 11,597 for two different researchers on April 14, 2023, and was subsequently split into assigned page numbers for separate teams to review, and the results increased to 11,739 by May 4, 2023.

Screening Process

Six researchers in pairs of two independent reviewers per team conducted three phases of screening across the three different databases. The screening process began with title screening, then abstract screening, and ended with a full-text review of the articles. Each individual in the team worked separately from their counterpart, only meeting to discuss possible issues as a whole research group to ensure independence across the screening of the articles. Discrepancies between researchers (when only one would choose to include the article) resulted in the article moving forward to the next screening process. Any discrepancies that persisted after full-text review was reviewed by a third reviewer for final decision on inclusion/exclusion. Figure 1 shows the flowchart of the screening process. There were 761 articles included after title screening, 330 articles after abstract review, and 46 articles

after full-text review. The most common reasons for excluding studies were that the device being studied was not consumer-grade (it may have been a novel device, proof of concept being tested, etc.), the wearable device was not being tested for validity or reliability, the study was actually testing an algorithm associated with wearable technology (rather than the device performance), the study did not contain an exercise or sport-related task (activities of daily living may have been tested, or sleep, etc.), among many others.

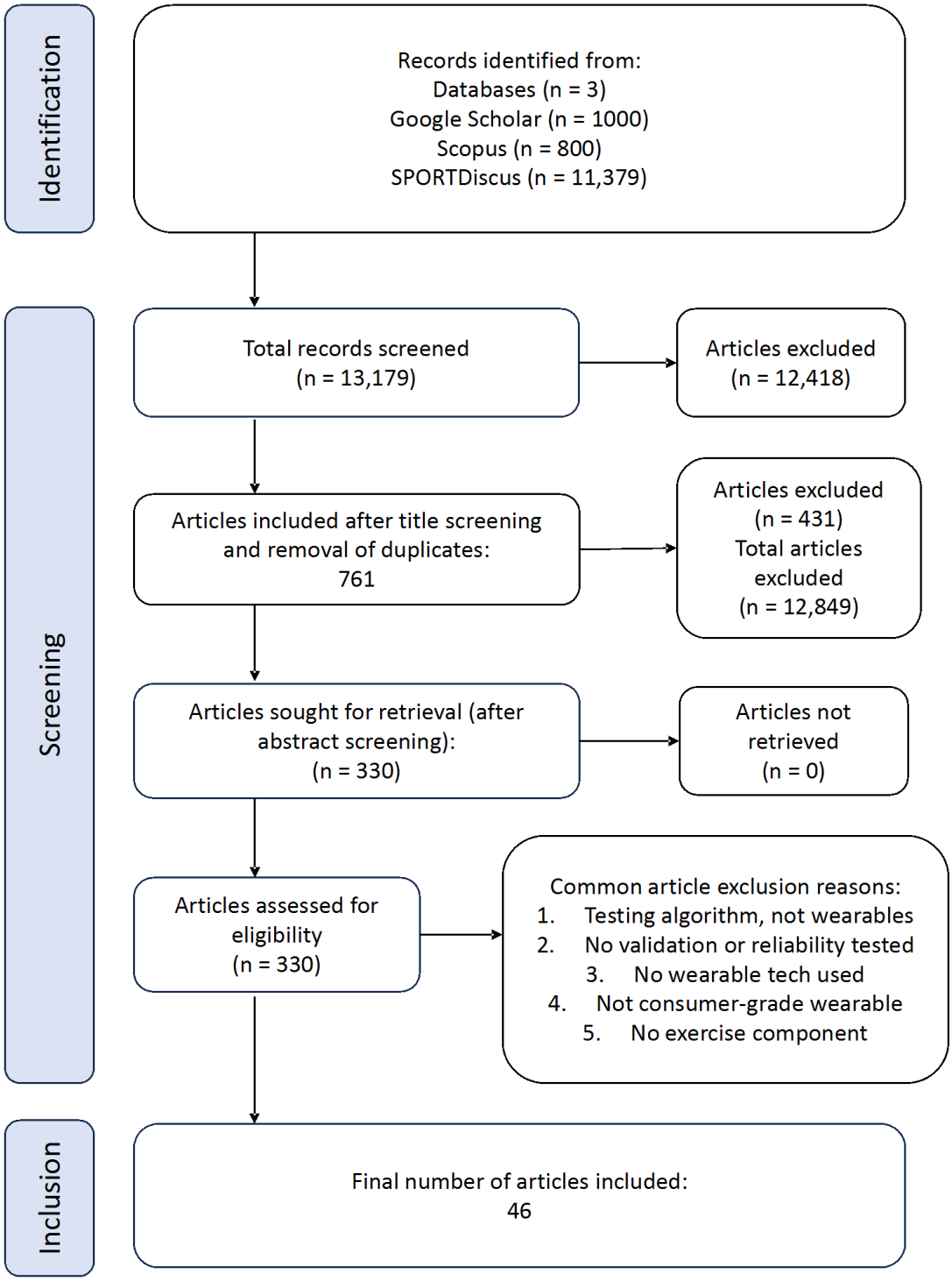

Figure 3.1. Flowchart of Search Strategy.

Flowchart includes identification, screening, and inclusion/exclusion process.

Data Extraction and Risk of Bias Assessment

Data extraction was completed after the screening process by pairs of two independent reviewers, per team. The research group created an Excel spreadsheet with the areas of interest for the researchers to extract. After extraction was completed by each researcher, team members met and resolved any discrepancies together to create a final data extraction file that was compiled and sent to the research group to help synthesize the papers succinctly for the current paper. In conjunction with the data extraction, a risk of bias assessment was provided for each study utilizing the novel WEAR-BOT.

Data Analysis

Summary statistics were calculated from the data extracted, including test variables, exercise modalities, counts of manufacturers and device models, average sample size, weighted average of overall Pearson correlations, weighted average of overall MAPE, and sample size statistics (based on weighted average of correlations). Summary statistics were performed in Google Sheets (Google LLC, Mountain View, CA, USA) and Excel (Microsoft Corporation, Redmond, WA, USA) with the MetaXL add-in for the meta-analysis and forest plot generation (EpiGear International Pty Ltd., Queensland, AUS). Counts of physiological variables tested are calculated per study and exercise modality. For example, if a study tested heart rate (HR) during walking, running, and biking, each would be added in the counts. Exercise modalities are calculated per study. The weighted correlation average was calculated by MetaXL while the weighted MAPE average was based on the sample size and the reported MAPE, per device and exercise modality. If a study reported Pearson correlation coefficients (or MAPE) for 20 individuals for walking and running for two different devices, there would be four coefficients (or MAPE values) taken into account in the weighted average, and the sample size would be 20 for each coefficient. MetaXL also produced 95% CI, and the lower bound was used as the effect size for sample size calculations. Sample

size statistics were performed using G*Power version 3.1.9.4 (Faul et al., 2009) using a correlation (bivariate normal model), and a desired power of 0.95, for HR, energy expenditure (EE), and VO2max variables. No other variables had a sufficient number of studies that reported Pearson's correlation to warrant analysis. In addition, if a specific variable did not have several data collection periods constituting the overall average, it was excluded from reporting in this paper.

Results

Of the 46 articles that were included in this review, 44 of them performed validation testing, while eight performed reliability testing. Only two studies performed reliability testing exclusively, the other six that performed reliability also performed validity testing. The complete list of included articles can be found in Table 1 (Alfonso et al., 2022; Baek et al., 2021; Bent et al., 2020; Budig et al., 2021; Carrier et al., 2020b; Chow & Yang, 2020; Climstein et al., 2020; Cosoli et al., 2022; Cosoli et al., 2023; Costello et al., 2022; Damasceno et al., 2022; Davarzani et al., 2020; de la Casa Pérez et al., 2022; Düking et al., 2020; Goods et al., 2023; Haddad et al., 2020; Hajj-Boutros et al., 2023; Hashimoto et al., 2022; Haveman et al., 2022; Hermand et al., 2021; Ho et al., 2022; Hopkins et al., 2020; Jachymek et al., 2021; Jagim et al., 2020; Kristiansson et al., 2023; Lucernoni et al., 2022; Martín-Escudero et al., 2023; Muggeridge et al., 2021; Navalta et al., 2020a; Navalta et al., 2020b; Nazari & MacDermid, 2020; Newton et al., 2023; Nissen et al., 2022; O'Driscoll et al., 2020; Paradiso et al., 2020; Reece et al., 2021; Rider et al., 2021; Rodin et al., 2022; Schams et al., 2022; Shumate et al., 2021; Snarr et al., 2021; Snyder et al., 2021; Stove & Hansen, 2022; Støve et al., 2020; Takahashi et al., 2022; Tokizawa et al., 2022).

Table 3.1. Complete List of Included Studies.

Number	Title	Author	Year
1	Agreement between two photoplethysmography-based wearable devices for monitoring heart rate during different physical activity situations: a new analysis methodology	Alfonso et al.	2022
2	Accuracy of wearable devices for measuring heart rate during conventional and Nordic walking	Baek, Ha, & Park	2021
3	Investigating sources of inaccuracy in wearable optical heart rate sensors	Bent, Goldstein, Kibbe, & Dunn	2020
4	Heart Rate and Distance Measurement of Two Multisport Activity Trackers and a Cellphone App in Different Sports: A Cross-Sectional Validation and Comparison Field Study	Budig et al.	2021
5	Validation of garmin fenix 3 HR fitness tracker biomechanics and metabolics (VO2max)	Carrier et al.	2020
6	Accuracy of optical heart rate sensing technology in wearable fitness trackers for young and older adults: Validation and comparison study	Chow & Yang	2020
7	Reliability of the polar vantage m sports watch when measuring heart rate at different treadmill exercise intensities	Climstein et al.	2020
8	Wearable Electrocardiography for Physical Activity Monitoring: Definition of Validation Protocol and Automatic Classification	Cosoli, Antognoli, & Scalise	2023
9	Accuracy and Precision of Wearable Devices for Real-Time Monitoring of Swimming Athletes	Cosoli, Antognoli, Veroli, & Scalise	2022
10	Isolated & combined wearable technology underestimate the total energy expenditure of professional young rugby league players; a doubly labelled water validation study	Costello et al.	2022
11	Criterion validity and accuracy of a heart rate monitor	Damasceno et al.	2022
12	Validity and reliability of StriveTM Sense3 for muscle activity monitoring during the squat exercise	Davarzani et al.	2020

13	Is the xiaomi mi band 4 an accuracy tool for measuring health-related parameters in adults and older people? an original validation study	de la Casa Pérez et al.	2022
14	Wrist-worn wearables for monitoring heart rate and energy expenditure while sitting or performing light-to-vigorous physical activity: validation study	Düking et al.	2020
15	Concurrent validity of the CORE wearable sensor with BodyCap temperature pill to assess core body temperature during an elite women's field hockey heat training camp	Goods et al.	2023
16	Ecological validation and reliability of hexoskin wearable body metrics tool in measuring pre-exercise and peak heart rate during shuttle run test in professional handball players	Haddad et al.	2020
17	Wrist-worn devices for the measurement of heart rate and energy expenditure: A validation study for the Apple Watch 6, Polar Vantage V and Fitbit Sense	Hajj-Boutros et al.	2022
18	Validation of Wearable Device Consisting of a Smart Shirt with Built-In Bioelectrodes and a Wireless Transmitter for Heart Rate Monitoring in Light to Moderate Physical Work	Hashimoto et al.	2022
19	Continuous monitoring of vital signs with wearable sensors during daily life activities: validation study	Haveman et al.	2022
20	Accuracy and reliability of pulse O2 saturation measured by a wrist-worn oximeter	Hermand, Coll, Richalet, & Lhuissier	2021
21	Accuracy of wrist-worn wearable devices for determining exercise intensity	Ho, Yang, & Li	2022
22	Consumer-grade biosensor validation for examining stress in healthcare professionals	Hopkins et al.	2020
23	Wristbands in Home-Based Rehabilitation Validation of Heart Rate Measurement	Jachymek, et al.	2021
24	The accuracy of fitness watches for the measurement of heart rate and energy expenditure during moderate intensity exercise	Jagim et al.	2020
25	Validation of Oura ring energy expenditure and steps in laboratory and free-living	Kristiansson et al.	2023
26	ActivPAL accuracy in determining metabolic rate during walking, running and cycling	Lucernoni, Kim, & Byrnes	2022

27	Are Activity Wrist-Worn Devices Accurate for Determining Heart Rate during Intense Exercise?	Martin-Escudero et al.	2023
28	Measurement of heart rate using the polar OH1 and Fitbit charge 3 wearable devices in healthy adults during light, moderate, vigorous, and sprint-based exercise: validation study	Muggeridge et al.	2021
29	Concurrent heart rate validity of wearable technology devices during trail running	Navalta, Montes et al.	2020
30	Validity and reliability of three commercially available smart sports bras during treadmill walking and running	Navalta, Ramirez et al.	2020
31	Reliability of zephyr bioHarness respiratory rate at rest, during the modified Canadian aerobic fitness test and recovery	Nazari & MacDermid	2020
32	The Validity of a Novel Low-Cost, Wearable Physical Activity Monitor in a Laboratory Setting: Direct Original Research	Newton, Glickman, & Barkley	2023
33	Heart rate measurement accuracy of fitbit charge 4 and samsung galaxy watch active2: Device evaluation study	Nissen et al.	2022
34	The validity of two widely used commercial and research-grade activity monitors, during resting, household and activity behaviours	O'Driscoll et al.	2020
35	The validity and reliability of the mi band wearable device for measuring steps and heart rate	Paradiso, Colino, & Liu	2020
36	Assessing heart rate using consumer technology association standards	Reece et al.	2021
37	Examining the accuracy of the polar A360 monitor	Rider et al.	2021
38	An accurate wearable hydration sensor: Real-world evaluation of practical use	Rodin et al.	2022
39	Validation of a smart shirt for heart rate variability measurements at rest and during exercise	Schams et al.	2022
40	Validity of the Polar Vantage M watch when measuring heart rate at different exercise intensities	Shumate et al.	2021
41	Validity of Wearable Electromyographical Compression Shorts to Predict Lactate Threshold During Incremental Exercise in Healthy Subjects	Snarr, Tolusso, Hallmark, & Esco	2021

42	Comparison of the Polar V800 and the Garmin Forerunner 230 to predict VO2max	Snyder, Willoughby, & Smith	2021
43	Accuracy of the Apple Watch Series 6 and the Whoop Band 3.0 for assessing heart rate during resistance exercises	Støve et al.	2022
44	Measurement latency significantly contributes to reduced heart rate measurement accuracy in wearable devices	Støve et al.	2020
45	Accuracy of Heart Rate and Respiratory Rate Measurements Using Two Types of Wearable Devices	Takahashi et al.	2022
46	Validity of a wearable core temperature estimation system in heat using patch-type sensors on the chest	Tokizawa et al.	2022

Complete list of included studies based on inclusion/exclusion criteria stated above, from 3 databases

(Google Scholar, Scopus, and SPORTDiscus).

Study Characterization Results

In this analysis, the studies reviewed included 70 different wearable devices across 34 different manufacturers. A list of all devices and the studies that tested them can be found in Table 2. There were 14 different physiological variables tested, with HR and energy expenditure (EE) being the top 2 (see Table 3). The average sample size was 29.80 participants. The weighted Pearson correlation average was 0.81, 0.73, 0.83 for HR, EE, and VO2max, respectively (see Figures 2-4). The combined sample size for HR is 2,780, across 89 data collection periods (modalities and devices per study), 2,178 (across 61 data collection periods) for EE, and 61 (across 3 data collection periods) for VO2max (see Table 4). While there were 14 physiological variables evaluated, only data for six are aggregated and included in Table 4, because there was not enough data to provide a meaningful aggregation in all variables tested. The data for the weighted average can be found in the appendix (Table A.1). The lower bound of the 95% confidence interval (CI) for HR, EE, and VO2max is 0.77, 0.68, and 0.70. The minimum sample size needed to reach a power of 0.95 is 13, 18, and 17, for HR, EE, and VO2max, respectively. The weighted average for MAPE was 12.48%, 34.13%, 19.29%, 2.42%, and 6.21% for HR, EE, respiratory rate (RR), oxygen saturation (OS), and skin temperature (ST), respectively. The combined sample size for HR is 4,084 (across 154 data collection periods), 1,108 for EE (across 32 data collection periods), 120 for RR (across 6 data collection periods), 80 for OS (across 4 data collection periods), and 120 for ST (across 6 data collection periods) (see Table 4). The data for the weighted average can be found in the appendix (Table A.2).

Of the 44 studies that tested validity, 33 (75%) of them utilized some form of correlation analysis, with 21 (47.73%) using Pearson's, seven (15.91%) using Lin's Concordance Correlation Coefficient (CCC), seven (15.91%) using Intraclass Correlation Coefficient (ICC), and two (4.55%) using Spearman's. Some studies utilized multiple correlation tests (e.g. Pearson's and Lin's). Additionally, 32 (72.73%) tested error in some manner, with 23 (52.27%) using mean absolute percentage error (MAPE), 12 (27.27%) using mean

absolute error (MAE), and 11 (25%) using root mean square error (RMSE). There were 36 (81.82%) that plotted Bland-Altman plots, and only two (4.55%) that utilized any type of equivalence testing.

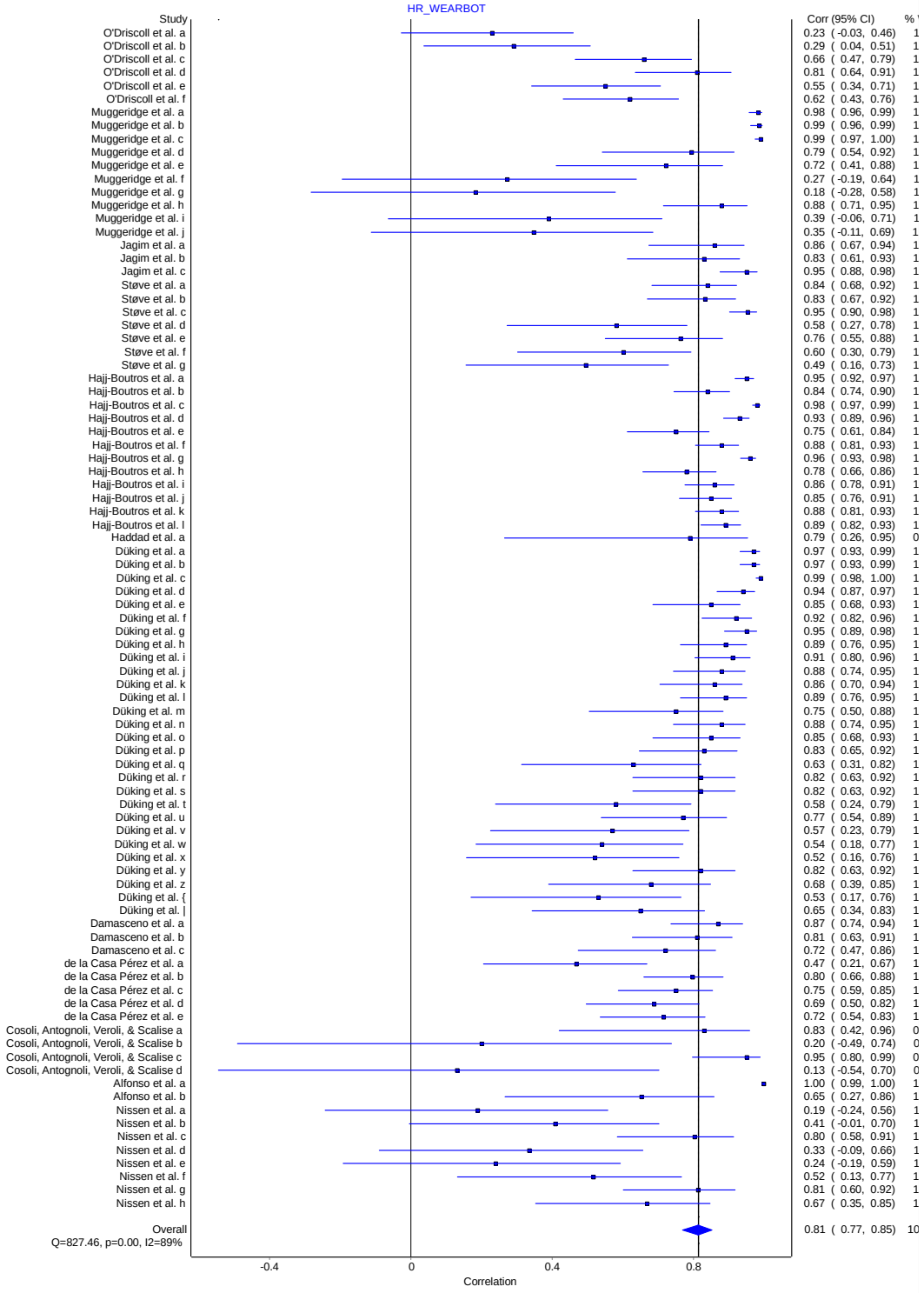

Figure 3.2. Forest Plot for Correlation Studies that Examined HR.

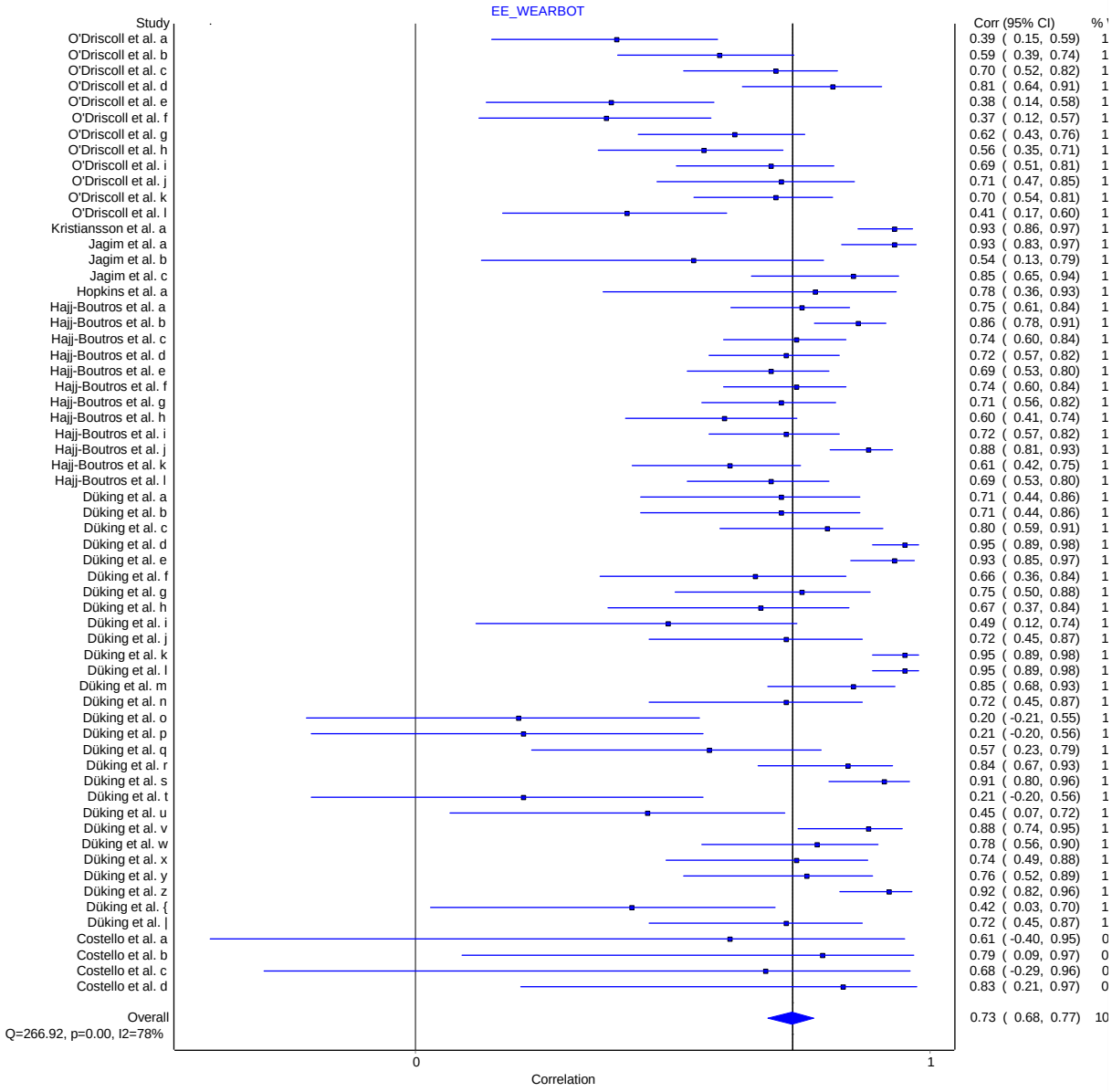

Figure 3.3. Forest Plot for Studies that Examined EE.

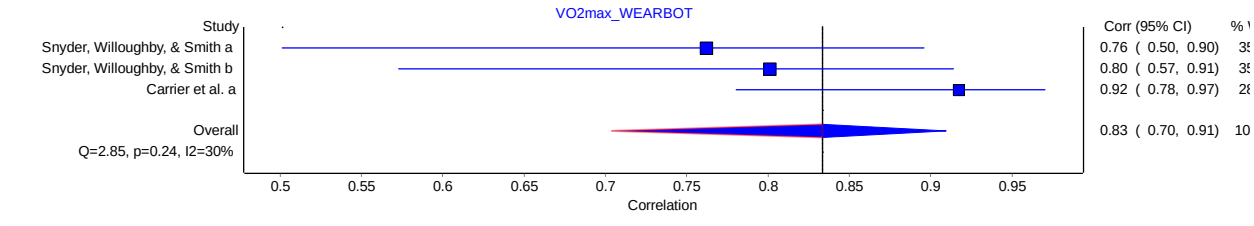

Figure 3.4. Forest Plot for Studies that Examined VO2max.

Table 3.2. Wearable Technology Tested, by Study

Manufacturer	Model	Author	Year
Adidas	Smart Sports Bra	Navalta, Ramirez et al.	2020
Ambiotex	Ambiotex Smart Shirt	Schams et al.	2022
Apple	Apple Watch (model not specified)	Martin-Escudero et al.	2023
Apple	Apple Watch Series 2	Støve et al.	2020
Apple	Apple Watch Series 4	Bent, Goldstein, Kibbe, & Dunn	2020
Apple	Apple Watch Series 4	Düking et al.	2020
Apple	Apple Watch Series 4	Reece et al.	2021
Apple	Apple Watch Series 6	Ho, Yang, & Li	2022
Apple	Apple Watch Series 6	Støve et al.	2022
Apple	Apple Watch Series 6	Alfonso et al.	2022
Apple	Apple Watch Series 6	Hajj-Boutros et al.	2022
Berlei	Sports Bra	Navalta, Ramirez et al.	2020
Biovotion AG	Everion	Haveman et al.	2022
BodyMedia	SenseWear Pro3	Costello et al.	2022
Fitbit	Charge	Martin-Escudero et al.	2023
Fitbit	Charge 2	Baek, Ha, & Park	2021
Fitbit	Charge 2	Bent, Goldstein, Kibbe, & Dunn	2020
Fitbit	Charge 2	O'Driscoll et al.	2020
Fitbit	Charge 3	Haveman et al.	2022
Fitbit	Charge 3	Muggeridge et al.	2021
Fitbit	Charge 4	Jachymek, et al.	2021

Fitbit	Charge 4	Nissen et al.	2022
Fitbit	Sense	Hajj-Boutros et al.	2022
Fitbit	Versa	Düking et al.	2020
Fitbit	Versa	Jagim et al.	2020
Garmin	fēnix 3 HR	Carrier et al.	2020
Garmin	fēnix 5	Düking et al.	2020
Garmin	fēnix 5	Navalta, Montes et al.	2020
Garmin	fēnix 5	Düking et al.	2020
Garmin	Forerunner 230	Snyder, Willoughby, & Smith	2021
Garmin	Forerunner 235	Støve et al.	2020
Garmin	Forerunner 245	Hermand, Coll, Richalet, & Lhuissier	2021
Garmin	Forerunner 735 XT	Reece et al.	2021
Garmin	Forerunner 735XT	Damasceno et al.	2022
Garmin	Forerunner 945	Budig et al.	2021
Garmin	Forerunner 945	Ho, Yang, & Li	2022
Garmin	Venu Sq	Cosoli, Antognoli, Veroli, & Scalise	2022
Garmin	Vivosmart 3	Bent, Goldstein, Kibbe, & Dunn	2020
Garmin	Vivosmart HR+	Chow & Yang	2020
Goldwin	C3fit IN-pulse	Hashimoto et al.	2022
greenTEG	CORE	Goods et al.	2023
Hexoskin	Smart Shirt	Haddad et al.	2020
Jabra	Elite Sport Earbuds	Navalta, Montes et al.	2020
Jabra	Elite Sport Earbuds	Reece et al.	2021
Mad Apparel	Athos	Snarr, Tolusso, Hallmark, & Esco	2021

MediBioSense	VitalPatch	Haveman et al.	2022
Motiv	Motiv Ring	Navalta, Montes et al.	2020
Movband	Movband 3	Newton, Glickman, & Barkley	2023
Movband	Movband 4	Newton, Glickman, & Barkley	2023
Murata	Moni-Patch	Tokizawa et al.	2022
Oura	Gen2	Kristiansson et al.	2023
PAL Technologies	ActivPAL	Lucernoni, Kim, & Byrnes	2022
Polar	A360	Rider et al.	2021
Polar	Ignite	Budig et al.	2021
Polar	Ignite	Jagim et al.	2020
Polar	OH1	Muggeridge et al.	2021
Polar	H7	Baek, Ha, & Park	2021
Polar	TeamPro Sensor	Jagim et al.	2020
Polar	Vantage M	Climstein et al.	2020
Polar	Vantage M	Shumate et al.	2021
Polar	Vantage M2	Alfonso et al.	2022
Polar	Vantage V	Düking et al.	2020
Polar	Vantage V	Hajj-Boutros et al.	2022
Polar	Vantage V2	Cosoli, Antognoli, Veroli, & Scalise	2022
Polar	Vantage V3	Cosoli, Antognoli, Veroli, & Scalise	2022
Samsung	Galaxy Watch 3	Cosoli, Antognoli, & Scalise	2023
Samsung	Galaxy Watch Active 2	Nissen et al.	2022
Samsung	Gear S2	Martin-Escudero et al.	2023
Scosche	Rhythm 24	Reece et al.	2021

Scosche	Rhythm+	Navalta, Montes et al.	2020
Sensewear	Armband Mini	O'Driscoll et al.	2020
Sensoria Fitness	Biometric Sports Bra	Navalta, Ramirez et al.	2020
SpectroPhon	Dehydration Body Monitor (DBM) Paired with Samsung Gear Fit2	Rodin et al.	2022
SpectroPhon	Dehydration Body Monitor (DBM) Paired with Samsung Gear S2	Rodin et al.	2022
Spire Health	Stone	Takahashi et al.	2022
Striv	Sense3	Davarzani et al.	2020
Suunto	Spartan Sport Watch + Chest Strap	Navalta, Montes et al.	2020
TDK	Silmee W22	Takahashi et al.	2022
TomTom	Runner Cardio	Martin-Escudero et al.	2023
Vital	Scout	Hopkins et al.	2020
Whoop	Band 3.0	Støve et al.	2022
Xiaomi	Mi Band 2	Chow & Yang	2020
Xiaomi	Mi Band 2	Paradiso, Colino, & Liu	2020
Xiaomi	Mi Band 3	Bent, Goldstein, Kibbe, & Dunn	2020
Xiaomi	Mi Band 4	de la Casa Pérez et al.	2022
Xiaomi	Mi Band 5	Jachymek, et al.	2021
Zephr	Bioharness	Nazari & MacDermid	2020

Chapter 3 Table 2. Complete list of manufacturers and models tested, and what study tested them.

Table 3.3. Total Variables Tested and Exercise Modalities Used

Variables	Count	Modalities	Count
Heart Rate	88	Walking	30
Energy Expenditure	16	Running	26
Core Body Temperature	6	Cycling	15
Respiratory Rate	5	Sprinting	2
Fluid Loss	2	Swimming	2
Oxygen Saturation	2	Resistance Training	2
Skin Temperature	2	Squatting	2
VO2max	2	Stairs	2
Moves (correlated to HR)	1	Trail Running	1
Moves (correlated to VO2)	1	Hockey	1
R-R Interval	1	Rugby	1
Lactate Threshold	1	Arm Ergometer	1
Oxygen Consumption (VO2)	1		
Muscle Activation	1		

Counts of variables testes and modalities used in all included studies.

Table 3.4. Weighted Averages for Correlation and MAPE Values

	Heart Rate	Energy Expenditure	VO2max	Respiratory Rate	O2 Saturation	Skin Temperature
Pearson Correlation Weighted Average	0.74	0.67	0.82			
Lower 95% CI	0.70	0.64	0.82			
Upper 95% CI	0.77	0.70	0.82			
Standard Deviation	0.23	0.19	0.08			
Sample Size	2,780	2,178	61			
MAPE Weighted Average	12.48%	30.70%		19.08%	2.40%	6.13%
Lower 95% CI	4.75%	27.27%		18.88%	2.38%	6.05%
Upper 95% CI	20.20%	34.13%		19.29%	2.42%	6.21%
Standard Deviation	51.85%	38.65%		9.18%	1.20%	3.66%
Sample Size	4,084	1,108		120	80	120

Statistics for weighted averages based on studies that reported either Pearson correlation coefficient or mean absolute percentage error (MAPE). Weighted average is weighted by sample size per data collection period. CI = confidence interval.

Validity Risk of Bias Results

Of the 44 studies that tested validity, 14 (31.82%) were classified overall as having “Some Risk of Bias”, and 30 (68.18%) were classified as having a “High Risk of Bias” (see Table 5). None of the published studies were classified overall as having a “Low Risk of Bias”. The areas that pose the greatest risk of introducing bias in validation studies are in the “Data Processing” and “Statistical Tests” sections of the WEAR-BOT (specifically “Statistical Tests – Continuous Variables”). Some areas generate an N/A if the study does not address that section, so while a total of 44 studies were tested, some sections have fewer total results. For the “Data Processing” section, 26/44 (59.09%) studies examining validity had a “High Risk of Bias”, 14/44 (31.82%) had “Some Risk of Bias” rating, and 4/44 (9.09%) had a “Low Risk of Bias”. This section asks questions such as, “Was the data processing methods described appropriately and reproducible?”, “Was the amount of missing and/or cleaned data reported?”, “If necessary, was the method of aligning data reported and reasonable (e.g. aligned on timestamp, elapsed time)?”, among others. The next section that introduced the greatest risk of bias into studies was the “Statistical Tests” section, with 14/43 (32.56%) and 29/43 (67.44%) studies being classified as “High Risk of Bias” and “Some Risk of Bias”, respectively. 0 studies produced a classification of “Low Risk of Bias” for the “Statistical Tests” section.

The areas where studies performed the best were in the “Test Variables” and “Test Protocols” sections. The “Test Variables” section had 43/44 (97.73%) studies classified as “Low Risk of Bias”, 1/44 (2.27%) as “Some Risk of Bias”, and 0 as “High Risk of Bias”. The “Test Protocols” section had 42/44 (95.45%) studies as “Low Risk of Bias”, 2/44 (0.455%) studies as “Some Risk of Bias”, and 0 as “High Risk of Bias”. The “Test Variables” portion of the WEAR-BOT is a single question that asks, “Are the units of measurement (or estimated values) between test device and criterion the same?”. The “Test Protocols” section asks questions such as, “Were measurements between the criterion and test device taken concurrently?”, “If tested sequentially, was the test order randomized?”, among others.

Table 3.5. Risk of Bias Analysis for All Validation Studies Reviewed

Article Information		Category Risk of Bias Results:									Overall Result
Author	Year	Test Variables	Criterion Device	Test Devices	Test Protocols	Participants	Test Environment	Data Processing	Statistical Tests - Continuous Variables	Statistical Tests - Categorical Variables	
Alfonso et al.	2022	Low Risk of Bias	Some Risk of Bias	Some Risk of Bias	Low Risk of Bias	Some Risk of Bias	Some Risk of Bias	High Risk of Bias	High Risk of Bias	N/A	High Risk of Bias
Baek, Ha, & Park	2021	Low Risk of Bias	Some Risk of Bias	Some Risk of Bias	Low Risk of Bias	Some Risk of Bias	Low Risk of Bias	High Risk of Bias	Some Risk of Bias	N/A	High Risk of Bias
Bent, Goldstein, Kibbe, & Dunn	2020	Low Risk of Bias	Low Risk of Bias	Some Risk of Bias	Low Risk of Bias	Low Risk of Bias	Some Risk of Bias	Some Risk of Bias	High Risk of Bias	N/A	High Risk of Bias
Budig et al.	2021	Low Risk of Bias	Some Risk of Bias	High Risk of Bias	Low Risk of Bias	Some Risk of Bias	Some Risk of Bias	High Risk of Bias	High Risk of Bias	N/A	High Risk of Bias
Carrier et al.	2020	Low Risk of Bias	Some Risk of Bias	Low Risk of Bias	Low Risk of Bias	Some Risk of Bias	Low Risk of Bias	High Risk of Bias	Some Risk of Bias	N/A	High Risk of Bias
Chow & Yang	2020	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	High Risk of Bias	Some Risk of Bias	N/A	High Risk of Bias
Cosoli, Antognoli, & Scalise	2023	Low Risk of Bias	Some Risk of Bias	Low Risk of Bias	Low Risk of Bias	Some Risk of Bias	Low Risk of Bias	High Risk of Bias	N/A	Some Risk of Bias	High Risk of Bias

Cosoli, Antognoli, Veroli, & Scalise	2022	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	Some Risk of Bias	Low Risk of Bias	High Risk of Bias	Some Risk of Bias	N/A	High Risk of Bias
Costello et al.	2022	Low Risk of Bias	Some Risk of Bias	Some Risk of Bias	Low Risk of Bias	Some Risk of Bias	Some Risk of Bias	High Risk of Bias	Some Risk of Bias	N/A	High Risk of Bias
Damasceno et al.	2022	Low Risk of Bias	Some Risk of Bias	Some Risk of Bias	Low Risk of Bias	Some Risk of Bias	Low Risk of Bias	High Risk of Bias	Some Risk of Bias	Some Risk of Bias	High Risk of Bias
Davarzani et al.	2020	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	Some Risk of Bias	Low Risk of Bias	Low Risk of Bias	High Risk of Bias	N/A	High Risk of Bias
de la Casa Pérez et al.	2022	Low Risk of Bias	Some Risk of Bias	High Risk of Bias	Some Risk of Bias	Some Risk of Bias	Low Risk of Bias	High Risk of Bias	High Risk of Bias	N/A	High Risk of Bias
Düking et al.	2020	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	Some Risk of Bias	Some Risk of Bias	Some Risk of Bias	Some Risk of Bias	N/A	Some Risk of Bias
Goods et al.	2023	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	Some Risk of Bias	Low Risk of Bias	Low Risk of Bias	Some Risk of Bias	N/A	Some Risk of Bias
Haddad et al.	2020	Low Risk of Bias	Some Risk of Bias	Low Risk of Bias	Low Risk of Bias	Some Risk of Bias	Low Risk of Bias	High Risk of Bias	Some Risk of Bias	N/A	High Risk of Bias
Hajj-Boutros et al.	2022	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	Some Risk of Bias	Low Risk of Bias	High Risk of Bias	Some Risk of Bias	N/A	High Risk of Bias
Hashimoto et al.	2022	Low Risk of Bias	Low Risk of Bias	Low Risk	Low Risk of Bias	Some Risk of Bias	Low Risk of Bias	High Risk of Bias	Some Risk of Bias	N/A	High Risk

				of Bias							of Bias
Haveman et al.	2022	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	Some Risk of Bias	Some Risk of Bias	N/A	Some Risk of Bias
Hermand, Coll, Richalet, & Lhuissier	2021	Low Risk of Bias	High Risk of Bias	High Risk of Bias	Low Risk of Bias	Some Risk of Bias	Low Risk of Bias	High Risk of Bias	High Risk of Bias	N/A	High Risk of Bias
Ho, Yang, & Li	2022	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	Some Risk of Bias	Low Risk of Bias	High Risk of Bias	Some Risk of Bias	N/A	High Risk of Bias
Hopkins et al.	2020	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	Some Risk of Bias	Some Risk of Bias	High Risk of Bias	High Risk of Bias	N/A	High Risk of Bias
Jachymek, et al.	2021	Low Risk of Bias	Low Risk of Bias	Some Risk of Bias	Low Risk of Bias	Some Risk of Bias	Low Risk of Bias	Some Risk of Bias	Some Risk of Bias	N/A	Some Risk of Bias
Jagim et al.	2020	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	Some Risk of Bias	Low Risk of Bias	Some Risk of Bias	Some Risk of Bias	N/A	Some Risk of Bias
Kristiansson et al.	2023	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	Some Risk of Bias	Low Risk of Bias	Some Risk of Bias	Some Risk of Bias	N/A	Some Risk of Bias
Lucernoni, Kim, & Byrnes	2022	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	Some Risk of Bias	High Risk of Bias	N/A	High Risk of Bias
Martin-Escudero et al.	2023	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	Some Risk of Bias	Low Risk of Bias	High Risk of Bias	High Risk of Bias	N/A	High Risk of Bias

Muggeridge et al.	2021	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	Some Risk of Bias	Low Risk of Bias	High Risk of Bias	Some Risk of Bias	N/A	High Risk of Bias
Navalta, Montes et al.	2020	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	High Risk of Bias	Some Risk of Bias	N/A	High Risk of Bias
Navalta, Ramirez et al.	2020	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	High Risk of Bias	Some Risk of Bias	N/A	High Risk of Bias
Newton, Glickman, & Barkley	2023	Some Risk of Bias	Low Risk of Bias	Some Risk of Bias	Low Risk of Bias	Some Risk of Bias	Low Risk of Bias	High Risk of Bias	High Risk of Bias	N/A	High Risk of Bias
Nissen et al.	2022	Low Risk of Bias	Low Risk of Bias	Some Risk of Bias	Low Risk of Bias	Low Risk of Bias	Some Risk of Bias	Low Risk of Bias	Some Risk of Bias	N/A	Some Risk of Bias
O'Driscoll et al.	2020	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	Some Risk of Bias	Low Risk of Bias	Some Risk of Bias	Some Risk of Bias	N/A	Some Risk of Bias
Paradiso, Colino, & Liu	2020	Low Risk of Bias	Some Risk of Bias	Some Risk of Bias	Low Risk of Bias	Some Risk of Bias	Low Risk of Bias	High Risk of Bias	High Risk of Bias	N/A	High Risk of Bias
Reece et al.	2021	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	Some Risk of Bias	Low Risk of Bias	Some Risk of Bias	Some Risk of Bias	N/A	Some Risk of Bias
Rider et al.	2021	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	Some Risk of Bias	Low Risk of Bias	Some Risk of Bias	High Risk of Bias	Some Risk of Bias	High Risk of Bias
Rodin et al.	2022	Low Risk of Bias	Low Risk of Bias	Low Risk	Low Risk of Bias	Some Risk of Bias	Low Risk of Bias	Low Risk of Bias	Some Risk of Bias	N/A	Some Risk

				of Bias							of Bias
Schams et al.	2022	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	Some Risk of Bias	Low Risk of Bias	Some Risk of Bias	Some Risk of Bias	N/A	Some Risk of Bias
Shumate et al.	2021	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	Some Risk of Bias	Low Risk of Bias	Some Risk of Bias	Some Risk of Bias	N/A	Some Risk of Bias
Snarr, Tolusso, Hallmark, & Esco	2021	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	High Risk of Bias	High Risk of Bias	N/A	High Risk of Bias
Snyder, Willoughby, & Smith	2021	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	Some Risk of Bias	Some Risk of Bias	Low Risk of Bias	Some Risk of Bias	Some Risk of Bias	N/A	Some Risk of Bias
Støve et al.	2020	Low Risk of Bias	Some Risk of Bias	Some Risk of Bias	Low Risk of Bias	Some Risk of Bias	Low Risk of Bias	High Risk of Bias	High Risk of Bias	N/A	High Risk of Bias
Støve et al.	2022	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	Low Risk of Bias	Some Risk of Bias	Low Risk of Bias	High Risk of Bias	Some Risk of Bias	N/A	High Risk of Bias
Takahashi et al.	2022	Low Risk of Bias	Low Risk of Bias	Some Risk of Bias	Low Risk of Bias	Some Risk of Bias	Low Risk of Bias	High Risk of Bias	Some Risk of Bias	N/A	High Risk of Bias
Tokizawa et al.	2022	Low Risk of Bias	Low Risk of Bias	Some Risk of Bias	Low Risk of Bias	Some Risk of Bias	Low Risk of Bias	Some Risk of Bias	Some Risk of Bias	N/A	Some Risk of Bias

Risk of Bias results for all validation studies reviewed.

Reliability Risk of Bias Results

Of the eight studies that tested reliability, 6/8 (75%) are classified overall as having a “High Risk of Bias”, 2/8 (25%) studies are classified overall as having “Some Risk of Bias”, and zero studies overall classified as having “Low Risk of Bias” (see Table 6). The reliability risk of bias analysis in the WEAR-BOT only has three overall sections, “Study Design and Methodology”, “Data Processing”, and “Statistical Tests”. The area that researchers performed the best on, was “Study Design and Methodology”, with 6/8 (75%) studies being classified as “Low Risk of Bias”, 1/8 (12.5%) as “Some Risk of Bias”, and 1/8 (12.5%) as “High Risk of Bias”. The area that researchers have the most room for improvement is “Data Processing”, where 6/8 (75%) are listed as “High Risk of Bias”, 1/8 (12.5%) as “Some Risk of Bias”, and 1/8 (12.5%) as “Low Risk of Bias”. As there is significantly less research examining reliability in these devices, descriptive statistics and weighted averages are not presented here due to the lack of available data.

Table 3.6. Risk of Bias Analysis for All Reliability Studies Reviewed

Article Information		Category Risk of Bias Results:			Overall Result
Author	Year	Study Design and Methodology	Data Processing	Statistical Tests	
Haddad et al.	2020	Low Risk of Bias	High Risk of Bias	Some Risk of Bias	High Risk of Bias
Davarzani et al.	2020	Low Risk of Bias	Some Risk of Bias	Some Risk of Bias	Some Risk of Bias
Climstein et al.	2020	Low Risk of Bias	High Risk of Bias	Some Risk of Bias	High Risk of Bias
Hermand, Coll, Richalet, & Lhuissier	2021	High Risk of Bias	High Risk of Bias	Some Risk of Bias	High Risk of Bias
de la Casa Pérez et al.	2022	Some Risk of Bias	High Risk of Bias	High Risk of Bias	High Risk of Bias
Navalta, Ramirez et al.	2020	Low Risk of Bias	High Risk of Bias	Low Risk of Bias	High Risk of Bias
Nazari & MacDermid	2020	Low Risk of Bias	Low Risk of Bias	Some Risk of Bias	Some Risk of Bias
Paradiso, Colino, & Liu	2020	Low Risk of Bias	High Risk of Bias	High Risk of Bias	High Risk of Bias

Chapter 3 Table 6. Risk of Bias results for all reliability studies reviewed.

Discussion

With overall MAPE averages ranging from 2.4% to 30.7%, and Pearson correlation averages ranging from 0.67, to 0.82, there are highly variable results in overall validity in current wearable technology. Thus, those seeking to utilize these devices should first seek to determine their validity, either by testing themselves, or reviewing published literature. The tables in the appendix of this paper may help researchers, coaches, or other users to evaluate the appropriate use-case of the devices they are looking to use. However, without established thresholds for validity, users will be required to make their own decisions on whether the device has appropriate validity for their specific use-case.

The Consumer Technology Association has previously recommended a sample size of 20 for validity studies (Consumer Technology Association, 2016). This analysis shows that 20 individuals should be sufficient for studying the validity of devices estimating or measuring HR, EE, or VO2max, given that their sample is also properly diverse. Sample diversity can affect sensor readings, such as skin type, tattoos, or body fat percentage for light-based PPG sensors (Consumer Technology Association, 2018). However, as the VO2max weighted average is only from three data collection periods (studies), and 61 participants, the minimum sample size of 17 should be cautiously considered, and efforts should be made to reach at least the recommended sample size of 20 until more is known about testing this variable. It is also important to consider that the weighted average was for many different exercise modalities. Certain modalities may require additional sample sizes, but these calculations can give researchers a starting point for planning studies.

As can be seen in the risk of bias analysis, the validation and reliability studies looking at physiological variables, published between January 2020 and April 2023, vary widely. This is despite several attempts to standardize the practices for validating consumer-grade wearable devices. There is substantial risk of bias in the majority of the reviewed studies. The biggest areas of concern are in the statistical tests and

data processing areas. There is less risk of bias in the study methodology. This means that, while the statistical analyses performed may not be easily compared across studies, and may be inappropriate at times, the heart of the studies (data collection methodology) being performed is generally performed properly.

Data processing is sometimes referred to as cleaning the data. This is needed only during repeated measures, and the amount of data processing necessary is generally proportional to the granularity of the signal. Raw data will need the most cleaning to be useful, and sensors that are more susceptible to motion artifacts or other noise in the signal will need more processing. Depending on what the signal to noise ratio is, researchers may need to perform additional cleaning, possibly even applying algorithms to clean the data. However, consumer-grade wearable technology does not generally make the raw data available to the user. Therefore, the initial processing is generally performed by proprietary algorithms, unknown to the user or researcher, rather than applying their own algorithms to clean the raw data. It is likely, however, that the data will still require cleaning if the signal processing time interval (epoch) accessible to the researcher is still relatively granular (e.g. second-by-second). If the data is aggregated into a larger epoch (e.g. 5-min, 10-min), then less cleaning is generally required, and possibly none at all. The risk for bias occurs if cleaning is done to alter the dataset that is not reported or not reasonable. This was the biggest concern for the published literature. Studies would not report how much missing data there was, what and how much data the researchers removed, and why and how it was done. Researchers should strive to properly describe their efforts to clean their data, and reviewers should ensure it is properly described before allowing the study to be published. Another possibility to reduce bias in the literature is for wearable technology companies to make a greater amount of raw, or semi-raw data available to users and researchers (encouraging agnostic data). Having the data prior to any proprietary cleaning or smoothing algorithms performed by the company will allow the devices to be evaluated more completely and increase our understanding of their appropriate use cases. The WEAR-

BOT asks questions to evaluate the amount of data processing that was performed, why, and if it was justified and appropriate. It also seeks to evaluate the alignment of data. When comparing two devices, there may be a lag in the data (such as a PPG test device vs an ECG criterion device when testing HR). Statistical measures have been established to identify lags in the signal, such as cross-correlations. The use of these tests in validation of consumer wearables have been proposed by researchers (van Lier et al., 2020), yet they are extremely rare in the literature. In addition to more advanced statistical tests, even basic aspects of data processing were missing from several papers reviewed, such as the epoch used for data aggregation, if any. Thus, the WEAR-BOT seeks to guide researchers to a better practice of performing and reporting their data processing efforts.

As has been reported in previous literature, the statistical tests performed for validity and reliability literature are far from uniform across papers. The most common test is a test of linearity, in the form of correlation (Pearson, Lin's, Spearman, etc.), though regression is also appropriate (Carrier et al., 2020a; Welk et al., 2019). The next most common type of test is a test of error (MAPE, RMSE, MAE, etc.). Finally, Bland-Altman plots are commonly reported, however, mean bias and limits of agreement are frequently not reported alongside the plots. Previous literature has proposed correlational tests, error tests, Bland-Altman, and equivalence tests to properly evaluate the validity of a device (Carrier et al., 2020a; Welk et al., 2019). The WEAR-BOT suggests three types of tests, (1) a test of error, (2) linearity, and (3) equivalence, in addition to a graphical representation of bias. While correlation tests for assessing linearity are discussed above, regression models for assessing linearity may be less well known. There are specific regression techniques best suited for method comparison, including Deming regression (parametric) and Passing-Bablok regression (non-parametric). While the FDA does not have specific guidance on validating consumer-grade wearable technology, they do recommend establishing linearity through regression in other method verification procedures (Office of Regulatory Affairs (ORA) Laboratory Manual Volume II, 2020). For validation purposes, reporting the y-intercept and regression

slope would be appropriate for Deming regression, while adding residual sum of squares, and the R^2 would be appropriate after performing simple linear regression. This is all assuming that the variables are continuous (which most are). However, if the variables in question are categorical, such as human activity recognition or physical activity classification, these tests are not appropriate. For validation studies with categorical variables, the WEAR-BOT proposes diagnostic tests (accuracy, specificity, sensitivity, AUC), reporting the classification table (confusion matrix), and a test of association appropriate for the level of measurement (Cohen's Kappa, rank-based correlations, etc.). While there are validation studies that performed some type of regression (Amitrano et al., 2020; Donisi et al., 2021; Hinde et al., 2021; Sen-Gupta et al., 2019), the vast majority of studies reviewed did not. For reliability testing, most performed either a test of absolute reliability (coefficient of variation) or relative reliability (Intraclass Correlation Coefficient). Thus, the results for the statistical tests for reliability performed much better than the validity testing. As stated previously, there is significantly less research examining reliability in these devices, thus additional insights, including descriptive statistics and weighted averages are not presented here due to the lack of available data.

The WEAR-BOT goes one step further than previous works, evaluating and recommending methodology and statistical tests for validity and reliability studies. However, there still is no universal acceptance on validity and reliability thresholds. As has been proposed in previous literature, establishing appropriate thresholds for validity and reliability are important to establish appropriate use cases (Carrier et al., 2021). While wearable technology is being used in research (Coughlin & Stewart, 2016; Mansi et al., 2021; Park & Jayaraman, 2003; Xiang et al., 2022), and even medical research (Burnham et al., 2018; Greiwe & Nyenhuis, 2020; Iqbal et al., 2016; Wu & Luo, 2019), the validity of the devices is often not known. Wearable technology has the potential to revolutionize biomedical and physiological research (Carrier et al., 2020a; Wright et al., 2017), but only if the appropriate devices are used in the appropriate scenarios. In the absence of regulation from a governing entity, the responsibility of testing these devices

falls to independent researchers. It is likely that tiered thresholds will need to be established as the accuracy thresholds for recreational use will not be as stringent as those necessary for research, professional/collegiate athletics, or military scenarios. Thus, we encourage researchers to seek to establish such thresholds.

As stated previously, the work of validating these devices is important. However, as these devices are meant for the general population, and the studies are oftentimes behind a paywall, there is a need for an easily accessible database for consumers, coaches, athletes, and researchers to access. Such a database would be a valuable resource for many. To compile the results of years, and soon-to-be decades of research is a difficult feat. Something no one person or lab group could reasonably accomplish. An organization would need to dedicate serious resources to compile and host this database, while continually monitoring it for the world to use. But the need for this will only continue to grow, as wearable technology grows. Therefore, we encourage the development of collaborative efforts to establish such a database and make it free for the world to use.

Limitations

This review was done with a version of the WEAR-BOT that was slightly different than the WEAR-BOT that has been published since. Minor grammatical changes, as well as question differences were made between the time this analysis was performed and the tool was published. The largest difference between the published WEAR-BOT and the WEAR-BOT used in this analysis is that the “Test Protocols” and “Test Environment” sections have been combined into a single section, entitled “Test Protocols and Parameters”. Readers should take this into consideration when reviewing this paper and the risk of bias analysis results. In addition, the weighted average only takes into account the reported participant sample size, and does not take into account the sampling rate of individual studies. For example, if two

studies examined HR during 20 minutes of running, and one study took HR measurements cross-sectionally after 20 minutes of running for 20 individuals, and the other recorded the repeated measures for the entire time and aggregated an overall average, they would both have the same weight. As almost all studies did not report the overall count of their repeated measures, a more appropriate weighted average (based on individual measurement observations, rather than participants) was not possible. Therefore, readers should use the results of the weighted average analysis carefully. We would also like to direct the reader to some limitations stated earlier, regarding the calculated sample size statistics and reported (or lack of reported) individual observations.

Conclusion

This systematic review looked at the risk of bias in validity and reliability research using consumer-grade wearable technology using the novel **WEA**rable Technology **R**isk **o**f **B**ias and **O**bjectivity **T**ool (WEAR-BOT). We found that every study evaluated from January 2020 to April 2023 had either "Some Risk of Bias" or "High Risk of Bias", overall. No study that was evaluated was classified as "Low Risk of Bias", overall. While some sections of the WEAR-BOT had some or the majority of the studies classified as "Low Risk of Bias", every study had at least one section that introduced risk of bias into the research. This most often came from the "Data Processing" or "Statistical Tests" sections. In addition, sample size calculations based on weighted averages of Pearson correlations from previous studies show a minimum sample size of 13, 19, and 17, for heart rate, energy expenditure, and VO2max validation studies, respectively. Therefore, the recommended sample size from the CTA of 20 is supported as being sufficiently powered for validity and reliability studies when studying HR, EE, or VO2max. In conclusion, we encourage those performing validity or reliability research into wearable technology to utilize the

WEAR-BOT checklist to ensure they reduce the amount of bias in their studies, and to improve the standardization across studies of methodology, analysis, and reporting.

Chapter 3 References

Alfonso, C., Garcia-Gonzalez, M. A., Parrado, E., Gil-Rojas, J., Ramos-Castro, J., & Capdevila, L. (2022). Agreement between two photoplethysmography-based wearable devices for monitoring heart rate during different physical activity situations: A new analysis methodology. Scientific Reports, 12(1), 15448.

American College of Sports Medicine. (2013). ACSM's health-related physical fitness assessment manual. Lippincott Williams & Wilkins.

Amitrano, F., Coccia, A., Ricciardi, C., Donisi, L., Cesarelli, G., Capodaglio, E. M., & D'Addio, G. (2020). Design and validation of an e-textile-based wearable sock for remote gait and postural assessment. Sensors, 20(22), 6691.

Baek, S., Ha, Y., & Park, H. (2021). Accuracy of wearable devices for measuring heart rate during conventional and Nordic walking. Pm&r, 13(4), 379-386.

Bent, B., Goldstein, B. A., Kibbe, W. A., & Dunn, J. P. (2020). Investigating sources of inaccuracy in wearable optical heart rate sensors. NPJ Digital Medicine, 3(1), 18.

Budig, M., Keiner, M., Stoohs, R., Hoffmeister, M., & Höltke, V. (2021). Heart rate and distance measurement of two multisport activity trackers and a cellphone app in different sports: a cross-sectional validation and comparison field study. Sensors, 22(1), 180.

BUNN, J. A., Navalta, J. W., Fountaine, C. J., & REECE, J. D. (2018). Current state of commercial wearable technology in physical activity monitoring 2015–2017. International Journal of Exercise Science, 11(7), 503.

Burnham, J. P., Lu, C., Yaeger, L. H., Bailey, T. C., & Kollef, M. H. (2018). Using wearable technology to predict health outcomes: a literature review. Journal of the American Medical Informatics Association, 25(9), 1221-1227.

Carrier, B., Barrios, B., Jolley, B. D., & Navalta, J. W. (2020a). Validity and Reliability of Physiological Data in Applied Settings Measured by Wearable Technology: A Rapid Systematic Review. Technologies, 8(4), 70.

Carrier, B., Creer, A., Williams, L. R., Holmes, T. M., Jolley, B. D., Dahl, S., Weber, E., & Standifird, T. (2020b). Validation of Garmin Fenix 3 HR fitness tracker biomechanics and metabolics (VO2max). Journal for the Measurement of Physical Behaviour, 3(4), 331-337.

Carrier, B., Salatto, R. W., Davis, D. W., Sertic, J. V. L., Barrios, B., Cater, P., & Navalta, J. W. (2021). Assessing the Validity of Several Heart Rate Monitors in Wearable Technology While Mountain Biking. Paper presented at the , 14(1) 18.

Chow, H., & Yang, C. (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(4), e14707.

Climstein, M., Alder, J. L., Brooker, A. M., Cartwright, E. J., Kemp-Smith, K., Simas, V., & Furness, J. (2020). Reliability of the polar vantage m sports watch when measuring heart rate at different treadmill exercise intensities. Sports, 8(9), 117.

Consumer Technology Association. (2016). Physical Activity Monitoring for Step Counting

ANSI/CTA-2056. ().Consumer Technology Association.

Consumer Technology Association. (2018). Physical Activity Monitoring for Heart Rate, ANSI/CTA-2065.

Cosoli, G., Antognoli, L., & Scalise, L. (2023). Wearable Electrocardiography for Physical Activity Monitoring: Definition of Validation Protocol and Automatic Classification. Biosensors, 13(2), 154.

Cosoli, G., Antognoli, L., Veroli, V., & Scalise, L. (2022). Accuracy and precision of wearable devices for real-time monitoring of swimming athletes. Sensors, 22(13), 4726.

Costello, N., Deighton, K., Cummins, C., Whitehead, S., Preston, T., & Jones, B. (2022). Isolated & combined wearable technology underestimate the total energy expenditure of professional young rugby league players; a doubly labelled water validation study. Journal of Strength and Conditioning Research, 36(12), 3398-3403.

Coughlin, S. S., & Stewart, J. (2016). Use of consumer wearable devices to promote physical activity: a review of health intervention studies. Journal of Environment and Health Sciences, 2(6)

Damasceno, V., Costa, A., Campello, M., Souza, D., Gonçalves, R., Campos, E., & Santos, T. (2022). Criterion validity and accuracy of a heart rate monitor. Human Movement, 23(1), 60-68.

Davarzani, S., Helzer, D., Rivera, J., Saucier, D., Jo, E., Chander, H., Strawderman, L., Ball, J. E., Smith, B. K., & Luczak, T. (2020). Validity and reliability of StriveTM Sense3 for muscle activity monitoring during the squat exercise. International Journal of Kinesiology & Sports Science, 8(4), 1.

de la Casa Pérez, A., Latorre Román, P. Á, Muñoz Jiménez, M., Lucena Zurita, M., Laredo Aguilera, J. A., Párraga Montilla, J. A., & Cabrera Linares, J. C. (2022). Is the Xiaomi Mi Band 4 an accuracy tool for measuring health-related parameters in adults and older People? An original validation study. International Journal of Environmental Research and Public Health, 19(3), 1593.

Donisi, L., Pagano, G., Cesarelli, G., Coccia, A., Amitrano, F., & D'Addio, G. (2021). Benchmarking between two wearable inertial systems for gait analysis based on a different sensor placement using several statistical approaches. Measurement, 173, 108642.

Düking, P., Giessing, L., Frenkel, M. O., Koehler, K., Holmberg, H., & Sperlich, B. (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(5), e16716.

Evenson, K. R., Goto, M. M., & Furberg, R. D. (2015). Systematic review of the validity and reliability of consumer-wearable activity trackers. International Journal of Behavioral Nutrition and Physical Activity, 12(1), 159.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. Behavior Research Methods, 41(4), 1149-1160.

Fuller, D., Colwell, E., Low, J., Orychock, K., Tobin, M. A., Simango, B., Buote, R., Van Heerden, D., Luan, H., & Cullen, K. (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.

Goods, P. S., Maloney, P., Miller, J., Jennings, D., Fahey-Gilmour, J., Peeling, P., & Galna, B. (2023). Concurrent validity of the CORE wearable sensor with BodyCap temperature pill to assess core body temperature during an elite women's field hockey heat training camp. European Journal of Sport Science, , 1-9.

Greiwe, J., & Nyenhuis, S. M. (2020). Wearable technology and how this can be implemented into clinical practice. Current Allergy and Asthma Reports, 20, 1-10.

Haddad, M., Hermassi, S., Aganovic, Z., Dalansi, F., Kharbach, M., Mohamed, A. O., & Bibi, K. W. (2020). Ecological validation and reliability of hexoskin wearable body metrics tool in measuring pre-exercise and peak heart rate during shuttle run test in professional handball players. Frontiers in Physiology, 11, 957.

Hajj-Boutros, G., Landry-Duval, M., Comtois, A. S., Gouspillou, G., & Karelis, A. D. (2023). Wrist-worn devices for the measurement of heart rate and energy expenditure: A validation study for the Apple Watch 6, Polar Vantage V and Fitbit Sense. European Journal of Sport Science, 23(2), 165-177.

Hashimoto, Y., Sato, R., Takagahara, K., Ishihara, T., Watanabe, K., & Togo, H. (2022). Validation of Wearable Device Consisting of a Smart Shirt with Built-In Bioelectrodes and a Wireless Transmitter for Heart Rate Monitoring in Light to Moderate Physical Work. Sensors, 22(23), 9241.

Haveman, M. E., van Rossum, M. C., Vaseur, R. M., van der Riet, C., Schuurmann, R. C., Hermens, H. J., de Vries, J. P., & Tabak, M. (2022). Continuous monitoring of vital signs with wearable sensors during daily life activities: validation study. JMIR Formative Research, 6(1), e30863.

Hermand, E., Coll, C., Richalet, J., & Lhuissier, F. J. (2021). Accuracy and reliability of pulse O2 saturation measured by a wrist-worn oximeter. International Journal of Sports Medicine, 42(14), 1268-1273.

Hinde, K., White, G., & Armstrong, N. (2021). Wearable devices suitable for monitoring twenty four hour heart rate variability in military populations. Sensors, 21(4), 1061.

Ho, W., Yang, Y., & Li, T. (2022). Accuracy of wrist-worn wearable devices for determining exercise intensity. Digital Health, 8, 20552076221124393.

Hopkins, L., Stacey, B., Robinson, D. B., James, O. P., Brown, C., Egan, R. J., Lewis, W. G., & Bailey, D. M. (2020). Consumer-grade biosensor validation for examining stress in healthcare professionals. Physiological Reports, 8(11), e14454.

Iqbal, M. H., Aydin, A., Brunckhorst, O., Dasgupta, P., & Ahmed, K. (2016). A review of wearable technology in medicine. Journal of the Royal Society of Medicine, 109(10), 372-380.

Jachymek, M., Jachymek, M. T., Kiedrowicz, R. M., Kaźmierczak, J., Płońska-Gościniak, E., & Peregud-Pogorzelska, M. (2021). Wristbands in Home-Based Rehabilitation—Validation of Heart Rate Measurement. Sensors, 22(1), 60.

Jagim, A. R., Koch-Gallup, N., Camic, C. L., Kroening, L., Nolte, C., Schroeder, C., Gran, L., & Erickson, J. L. (2020). The accuracy of fitness watches for the measurement of heart rate and energy expenditure during moderate intensity exercise. The Journal of Sports Medicine and Physical Fitness, 61(2), 205-211.

Kristiansson, E., Fridolfsson, J., Arvidsson, D., Holmäng, A., Börjesson, M., & Andersson-Hall, U. (2023). Validation of Oura ring energy expenditure and steps in laboratory and free-living. BMC Medical Research Methodology, 23(1), 1-11.

Lucernoni, K. M., Kim, S., & Byrnes, W. C. (2022). ActivPAL accuracy in determining metabolic rate during walking, running and cycling. Journal of Sports Sciences, 40(5), 591-599.

Mansi, S. A., Barone, G., Forzano, C., Pigliautile, I., Ferrara, M., Pisello, A. L., & Arnesano, M. (2021). Measuring human physiological indices for thermal comfort assessment through wearable devices: A review. Measurement, 183, 109872.

Martín-Escudero, P., Cabanas, A. M., Dotor-Castilla, M. L., Galindo-Canales, M., Miguel-Tobal, F., Fernández-Pérez, C., Fuentes-Ferrer, M., & Giannetti, R. (2023). Are Activity Wrist-Worn Devices Accurate for Determining Heart Rate during Intense Exercise? Bioengineering, 10(2), 254.

Muggeridge, D. J., Hickson, K., Davies, A. V., Giggins, O. M., Megson, I. L., Gorely, T., & Crabtree, D. R. (2021). Measurement of heart rate using the polar OH1 and Fitbit charge 3 wearable devices in healthy adults during light, moderate, vigorous, and sprint-based exercise: validation study. JMIR mHealth and uHealth, 9(3), e25313.

Navalta, J. W., Montes, J., Bodell, N. G., Salatto, R. W., Manning, J. W., & DeBeliso, M. (2020a). Concurrent heart rate validity of wearable technology devices during trail running. Plos One, 15(8), e0238569.

Navalta, J. W., Ramirez, G. G., Maxwell, C., Radzak, K. N., & McGinnis, G. R. (2020b). Validity and reliability of three commercially available smart sports bras during treadmill walking and running. Scientific Reports, 10(1), 7397.

Nazari, G., & MacDermid, J. C. (2020). Reliability of zephyr bioHarness respiratory rate at rest, during the modified canadian aerobic fitness test and recovery. The Journal of Strength & Conditioning Research, 34(1), 264-269.

Newton, A., Glickman, E., & Barkley, J. (2023). The Validity of a Novel Low-Cost, Wearable Physical Activity Monitor in a Laboratory Setting: Direct Original Research. Research Directs in Health Sciences, 3(1)

Nissen, M., Slim, S., Jäger, K., Flaucher, M., Huebner, H., Danzberger, N., Fasching, P. A., Beckmann, M. W., Gradl, S., & Eskofier, B. M. (2022). Heart rate measurement accuracy of fitbit charge 4 and samsung galaxy watch active2: Device evaluation study. JMIR Formative Research, 6(3), e33635.

O'Driscoll, R., Turicchi, J., Hopkins, M., Gibbons, C., Larsen, S. C., Palmeira, A. L., Heitmann, B. L., Horgan, G. W., Finlayson, G., & Stubbs, R. J. (2020). The validity of two widely used commercial and research-grade activity monitors, during resting, household and activity behaviours. Health and Technology, 10, 637-648.

Office of Regulatory Affairs (ORA) Laboratory Manual Volume II. (2020). Methods, Method Verification and Validation. (). https://www.fda.gov/media/73920/download

Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., & Brennan, S. E. (2021). PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. Bmj, 372

Paradiso, C., Colino, F., & Liu, S. (2020). The validity and reliability of the mi band wearable device for measuring steps and heart rate. International Journal of Exercise Science, 13(4), 689.

Park, S., & Jayaraman, S. (2003). Enhancing the quality of life through wearable technology. IEEE Engineering in Medicine and Biology Magazine, 22(3), 41-48.

Patel, V., Orchanian-Cheff, A., & Wu, R. (2021). Evaluating the validity and utility of wearable technology for continuously monitoring patients in a hospital setting: systematic review. JMIR mHealth and uHealth, 9(8), e17411.

Prill, R., Walter, M., Królikowska, A., & Becker, R. (2021). A systematic review of diagnostic accuracy and clinical applications of wearable movement sensors for knee joint rehabilitation. Sensors, 21(24), 8221.

Reece, J. D., Bunn, J. A., Choi, M., & Navalta, J. W. (2021). Assessing heart rate using consumer technology association standards. Technologies, 9(3), 46.

Rider, B. C., Conger, S. A., Ditzenberger, G. L., Besteman, S. S., Bouret, C. M., & Coughlin, A. M. (2021). Examining the accuracy of the polar A360 monitor. The Journal of Strength & Conditioning Research, 35(8), 2165-2169.

Rodin, D., Shapiro, Y., Pinhasov, A., Kreinin, A., & Kirby, M. (2022). An accurate wearable hydration sensor: Real-world evaluation of practical use. PloS One, 17(8), e0272646. 10.1371/journal.pone.0272646

Schams, P., Feodoroff, B., Zacher, J., Eibl, A., & Froböse, I. (2022). Validation of a smart shirt for heart rate variability measurements at rest and during exercise. Clinical Physiology and Functional Imaging, 42(3), 190-199. 10.1111/cpf.12746

Sen-Gupta, E., Wright, D. E., Caccese, J. W., Wright Jr, J. A., Jortberg, E., Bhatkar, V., Ceruolo, M., Ghaffari, R., Clason, D. L., & Maynard, J. P. (2019). A pivotal study to validate the performance of a novel wearable sensor and system for biometric monitoring in clinical and remote environments. Digital Biomarkers, 3(1), 1-13.

Shumate, T., Link, M., Furness, J., Kemp-Smith, K., Simas, V., & Climstein, M. (2021). Validity of the Polar Vantage M watch when measuring heart rate at different exercise intensities. PeerJ (San Francisco, CA), 9, e10893. 10.7717/peerj.10893

Snarr, R. L., Tolusso, D. V., Hallmark, A. V., & Esco, M. R. (2021). Validity of Wearable Electromyographical Compression Shorts to Predict Lactate Threshold During Incremental Exercise in Healthy Subjects. Journal of Strength and Conditioning Research, 35(3), 702-708. 10.1519/JSC.0000000000002721

Snyder, N. C., Willoughby, C. A., & Smith, B. K. (2021). Comparison of the Polar V800 and the Garmin Forerunner 230 to Predict V̇o 2max. Journal of Strength and Conditioning Research, 35(5), 1403-1409. 10.1519/JSC.0000000000002931

Stove, M. P., & Hansen, E. C. K. (2022). Accuracy of the Apple Watch Series 6 and the Whoop Band 3.0 for assessing heart rate during resistance exercises. Journal of Sports Sciences, 40(23), 2639-2644. 10.1080/02640414.2023.2180160

Støve, M. P., Holm, R. S., Kjaersgaard, A. S., Duncker, K., Jensen, M. R., & Larsen, B. T. (2020). Measurement latency significantly contributes to reduced heart rate measurement accuracy in wearable devices. Journal of Medical Engineering & Technology, 44(3), 125-132. 10.1080/03091902.2020.1753836

Takahashi, Y., Okura, K., Minakata, S., Watanabe, M., Hatakeyama, K., Chida, S., Saito, K., Matsunaga, T., & Shimada, Y. (2022). Accuracy of Heart Rate and Respiratory Rate Measurements Using Two Types of Wearable Devices. Progress in Rehabilitation Medicine, 7, 20220016. 10.2490/prm.20220016

Tokizawa, K., Shimuta, T., & Tsuchimoto, H. (2022). Validity of a wearable core temperature estimation system in heat using patch-type sensors on the chest. Journal of Thermal Biology, 108, 103294. 10.1016/j.jtherbio.2022.103294

van Lier, H. G., Pieterse, M. E., Garde, A., Postel, M. G., de Haan, H. A., Vollenbroek-Hutten, M. M., Schraagen, J. M., & Noordzij, M. L. (2020). A standardized validity assessment protocol for physiological signals from wearable technology: Methodological underpinnings and an application to the E4 biosensor. Behavior Research Methods, 52, 607-629.

Volkova, V. G., Räisänen, A., Benson, L. C., Ferber, R., & Kenny, S. J. (2023). Systematic review of methods used to measure training load in dance. BMJ Open Sport & Exercise Medicine, 9(3), e001484.

Welk, G. J., Bai, Y., Lee, J., Godino, J., Saint-Maurice, P. F., & Carr, L. (2019). Standardizing analytic methods and reporting in activity monitor validation studies. Medicine and Science in Sports and Exercise, 51(8), 1767.

Wright, S. P., Hall Brown, T. S., Collier, S. R., & Sandberg, K. (2017). How consumer physical activity monitors could transform human physiology research. American Journal of Physiology-Regulatory, Integrative and Comparative Physiology, 312(3), R358-R367.

Wu, M., & Luo, J. (2019). Wearable technology applications in healthcare: a literature review. Online J.Nurs.Inform, 23(3)

Xiang, L., Wang, A., Gu, Y., Zhao, L., Shim, V., & Fernandez, J. (2022). Recent machine learning progress in lower limb running biomechanics with wearable technology: A systematic review. Frontiers in Neurorobotics, 16, 913052.

Chapter 4 - Validation of Aerobic Capacity (VO2max) and Pulse Oximetry in Wearable Technology

Abstract

Introduction: Wearable technology continues to grow in popularity and sophistication, and the need for independent validation of these devices is needed to determine their overall accuracy and possible use-cases. Therefore, the purpose of this study was to evaluate the accuracy (validity) of maximal oxygen consumption (VO2max) estimates and blood oxygen saturation (BOS) measured via pulse oximetry using the Garmin fēnix 6 with a general population participant pool.

Methods: We recruited apparently healthy individuals (both active and sedentary) for VO2max (n=19) and pulse oximetry testing (n=22). VO2max was assessed through a graded exercise test and an outdoor run, comparing results from the Garmin fēnix 6 to a criterion measurement obtained from a metabolic system. Pulse oximetry involved comparing fēnix 6 readings under normoxic and hypoxic conditions against a medical-grade pulse oximeter. Data analysis included descriptive statistics, error analysis, correlation analysis, equivalence testing, and bias assessment, with validation criteria set at a concordance correlation coefficient (CCC) > 0.7 and mean absolute percentage error (MAPE) < 10%.

Results: The Garmin fēnix 6 provided accurate VO2max estimates, closely aligning with the 15-sec and 30-sec averaged laboratory data (MAPE for 30-sec avg = 7.05%, CCC for 30-sec avg = 0.73). However, it failed to accurately measure BOS under any condition or combined analysis (MAPE for combined conditions BOS = 4.29%, CCC for combined conditions BOS = 0.10).

Conclusion: While the Garmin fēnix 6 shows promise for estimating VO2max, reflecting its utility for both individuals and researchers, it falls short in accurately measuring BOS, limiting its application for monitoring acclimatization and managing pulmonary diseases. This research underscores the importance

of validating wearable technology devices to leverage their full potential in enhancing personal health and advancing public health research.

Introduction

Wearable technology (WT) has continued to grow in popularity and sophistication each year, with WT reaching the #1 spot in worldwide surveys of fitness trends in seven of the last nine years and being in the top three for the other two years (2018 and 2021) (Newsome et al., 2024; Thompson, 2015; Thompson, 2019; Thompson, 2022; Thompson, 2016; Thompson, 2017; Thompson, 2018; Thompson, 2021; Thompson, 2023). According to recent surveys, almost one in three Americans uses a wearable device to track their health and exercise, and around 70% of people own at least one wearable or plan to buy one in the next year (Clark, 2019; Dhingra et al., 2023). This prevalence of WT may represent a revolutionary change in physiology and public health research, simply due to the vast pool of potential data that may become available to researchers. Also an important aspect is the constant monitoring of physiological metrics that these devices perform, which will provide granular details into a person's physiology that could transform human physiology research (Carrier et al., 2020a; Wright et al., 2017). However, this transformation may only come to be realized if WT devices are found to be accurate in their measurements and estimates. As these consumer grade wearable devices are not subject to any type of regulation, there is no governing body ensuring accuracy. Thus, if researchers, athletes/coaches, public health officials, and health-care professionals hope to continue to utilize these devices, an understanding of their accuracy and when they can appropriately be used is necessary. This underpins the importance of independent validation of WT devices by researchers to further several scientific fields.

Among the many variables WT can estimate or measure, maximal aerobic capacity (or VO2max) and blood oxygen saturation (BOS) measured via pulse oximetry are important for a variety of health and fitness related purposes. VO2max represents the maximal amount of oxygen an individual can transport

from the environment into their lungs, diffuse into the blood, and extract at the muscles and organs to produce energy, or ATP. It represents a measure of cardiorespiratory fitness (CRF) and has a strong inverse relation with all-cause mortality and cardiovascular diseases (Harber et al., 2017; Lee et al., 2011; Qiu et al., 2021). VO2max also has an important relationship to endurance performance among athletes, being the most important single factor in predicting race performance (Bassett & Howley, 2000; Joyner & Coyle, 2008; Kenney et al., 2021). Pulse oximeters can non-invasively measure the amount of oxygen bound to hemoglobin based on how light reflects off the blood cells when broadcast from the device. Devices with pulse oximeters to measure BOS can also be used to monitor cardiorespiratory functions, especially in people with pulmonary diseases. It can also be useful for athletes looking to travel to altitude for an event or competition who wish to monitor their acclimatization process (Dünnwald et al., 2021; Sinex & Chapman, 2015). Therefore, the purpose of this study was to evaluate the accuracy (validity) of VO2max estimates and blood oxygen saturation measured via pulse oximetry using the Garmin fēnix 6 with a general population participant pool.

Methods

Prior to data collection occurring for this study, the protocols were approved by the University of Nevada, Las Vegas Institutional Review Board (IRB). All participants signed an informed consent and filled out pre-assessment documents prior to completing the study. While the VO2max and pulse oximetry testing were completed separately, some participants completed both, and are included in each dataset. As the participant pool for both VO2max and pulse oximetry testing are different, demographic data is provided for each group.

VO2max Testing

For VO2max testing, 19 apparently healthy, active and sedentary individuals were recruited to participate (25.50±5.26 years, 11 male, 8 female, 173.63±9.08 cm, 74.08±14.16 kg, BMI=24.42±3.21, 22.14±6.06% fat mass, 36.87±4.58% muscle mass, 25.07±23.65 km run per week, all reported as mean±SD). Data collection occurred over two separate days. The first day participants completed a graded exercise test on a treadmill utilizing progressive increases in speed and grade every two minutes until volitional exhaustion to determine VO2max. Maximal oxygen consumption was measured using the ParvoMedics TrueOne 2400 metabolic system (ParvoMedics Inc, Salt Lake City, UT, USA). VO2max was determined by taking the highest average oxygen consumption during the graded exercise test for a set timeframe. Aggregated VO2max values for 4-breath, 15-second, 30-second, and 1-minute averaged timeframes were obtained by the metabolic system and served as the criterion measure for comparisons to the WT device. The second day consisted of an outdoor run, guided by the wearable device (Garmin fēnix 6®, Garmin Ltd, Olathe, KS, USA) to generate an estimated VO2max value. Participants were asked to come back between two and seven days from the first visit (5.06±3.96 days). Researchers performed a factory reset on the watch prior to each subject to prevent data from previous participants influencing the measurements and estimates of the current subject. Participants then put on the associated heart rate monitor (Garmin HRM-Run®) for the outdoor run. The outdoor run involved a 10-15 minute run at an intensity above 70% of the participants estimated max HR, according to manufacturer guidelines. This provided the device enough data to estimate VO2max, using a linear extrapolation of heart rate (HR) and running speed (*Aerobic Fitness Level (VO_2max) Estimation – Firstbeat White Paper.* 2017). The outdoor run was performed in one of two places, the University track, or a flat area of campus, depending on logistics and track availability. Participants ran laps until researchers told them to stop within the time window. Five participants completed the testing at the track, and 15 participants completed testing on campus. The altitude was ~686m, and the average temperature during outdoor testing was 20.67±12.62

°C, as measured by local weather readings. The average distance, time, pace, and HR were 2.13±0.17 km, 12.91±1.42 min, 6.33±1.49 min/km, and 153.50±11.45 bpm, respectively, as measured by the device.

Pulse Oximetry Testing

For pulse oximetry testing, 22 apparently healthy individuals were recruited to participate (25.48±6.02 years, 13 male, 9 female, 173.27±7.70 cm, 68.88±9.10 kg, BMI=22.91±2.40, 18.55±7.05% fat mass, 38.73±3.61% muscle mass). Participants began by putting on the fēnix 6 on their left wrist and were instructed to have the strap tension secure but comfortable. Researchers then placed a medical grade pulse oximeter (Roscoe Medical Fingertip Pulse Oximeter, Model: POX-ROS, Roscoe Medical Inc., Middleburg Heights, OH, USA), on the right index finger of the participant. Participants completed testing under four conditions (normoxia/hypoxia, anterior/posterior watch placement). All participants were seated for all pulse oximetry tests. The first testing condition was under normoxic (normal oxygen concentration) conditions, with the watch head placed on the posterior wrist. Researchers performed the necessary steps on the watch to generate a BOS level by the fēnix 6 and recorded the value from the fingertip oximeter at the same time the watch generated a value. Afterwards, the watch was then placed on the anterior wrist, and the process repeated. After both normoxic conditions were completed, participants performed hypoxic (low oxygen concentration) testing of the pulse oximeter. Participants were connected to an altitude simulator machine (Hypoxico Everest Summit II, Hypoxico Inc., New York, New York, USA), for a minimum of five minutes to allow for blood oxygen levels to stabilize prior to testing. The machine was set to an altitude of 3657.6 m (12,000 ft) as the default for participants. However, if participants got lightheaded or uncomfortable at that simulated altitude, it was lowered to an altitude better tolerated by the individual and the five-minute waiting period reset, with the possibility of returning to normoxia for as long as needed before restarting at a lower simulated altitude.

All participants were seated for all pulse oximetry tests. Participants were instructed to control their breathing rate and breathed in and out in synchronization with the altitude simulator bursts of air. This corresponded to a breathing rate of 12.5 breaths per minute. Blood oxygen saturation testing under hypoxia was tested with the watch on the anterior and posterior left wrist, as was performed prior in the normoxic testing condition. The average time under hypoxia was 9.18±1.05 min. If the fēnix 6 was unable to generate a measurement of BOS for any trial, the researchers retried up to three times for each trial that the watch did not generate a value on the first attempt. If it was still unable to generate a measurement after three tries, no further attempts were made. Once values were obtained from the watch and the fingertip oximeter, the pulse oximetry testing was concluded.

Data Analysis

VO2max values for each timeframe (4-breath, 15-sec, 30-sec, and 1-min) and BOS values for each condition (anterior/posterior placement, normoxia/hypoxia) were input into Google Sheets (Alphabet Inc., Mountain View, CA, USA). Pulse oximetry values were compared by condition as well as combined dataset. All granular calculations were completed within Google Sheets. All summary statistics, validation measures, and figures were completed and generated in jamovi (jamovi project, version 2.2, https://www.jamovi.org/). Descriptive statistics, error analysis (mean absolute percentage error), correlation analysis (Pearson's r, Lin's Concordance Correlation Coefficient [CCC]), equivalence testing (TOST Paired Samples Test), and bias assessment (Bland-Altman analysis) were also performed. TOST test upper and lower bounds were set at +0.5 and -0.5 Cohen's D for each test. Data analysis for VO2max was completed by comparing the fēnix 6 estimates of VO2max to each laboratory aggregated timeframe. Determination of validation was pre-determined, and any condition that produced a CCC > 0.7, and a MAPE < 10%, the device was considered valid.

Results

VO2max

The 19 participants used for this analysis had an average VO2max of 48.9ml/kg/min and an average VO2max percentile of 83.37±21.14%, based on the 30-sec averaged VO2max values. Error analysis showed that the fēnix 6 VO2max estimate had a MAPE of less 10% for the 15-sec, 30-sec, and 1-min averaged timeframes (see Table 1). Correlation analysis produced a CCC > 0.7 for both the 15-sec and 30-sec averaged timeframes (see Table 1). Equivalence testing via the TOST test produced no equivalent results, with equivalence conditions being violated for the 4-breath, 15-sec, 30-sec, and 1-min averaged times (see Table 1). Bland-Altman bias values and 95% confidence intervals can be found in Table 1 and associated plots can be found for all time parameters in Figure 1.

Table 4.1. Validity Statistics for Garmin VO_2max Estimate

	fēnix 6 VO2max Estimate	Lab VO2max – 4 breath avg	Lab VO2max – 15 sec avg	Lab VO2max – 30 sec avg	Lab VO2max – 1 min avg
Mean (ml/kg/min)	49.68	54.54	49.95	48.94	47.91
Standard Deviation	4.61	7.28	7.04	6.67	6.76
MAPE		10.70%	**7.23%**	**7.05%**	**8.53%**
Pearson Correlation		0.73	0.78	0.78	0.76
Lin's Concordance		0.49	**0.71**	**0.73**	0.68
Bland-Altman Bias		-4.87 (-7.30, -2.44)	-0.26 (-2.45, 1.92)	0.75 (-1.28, 2.78)	1.77 (-0.35, 3.89)
TOST Test p-value (Upper)		<0.001	0.80	0.45	0.10
TOST Test p-value (Lower)		<0.972	0.01	0.09	0.34

VO2max descriptive and validation statistics results, n=20. MAPE = Mean Absolute Percentage Error, TOST Test = Two One-Sided T-Tests. Bland-Altman bias values and 95% confidence intervals provided. Values that met the predetermined validation criteria are bolded.

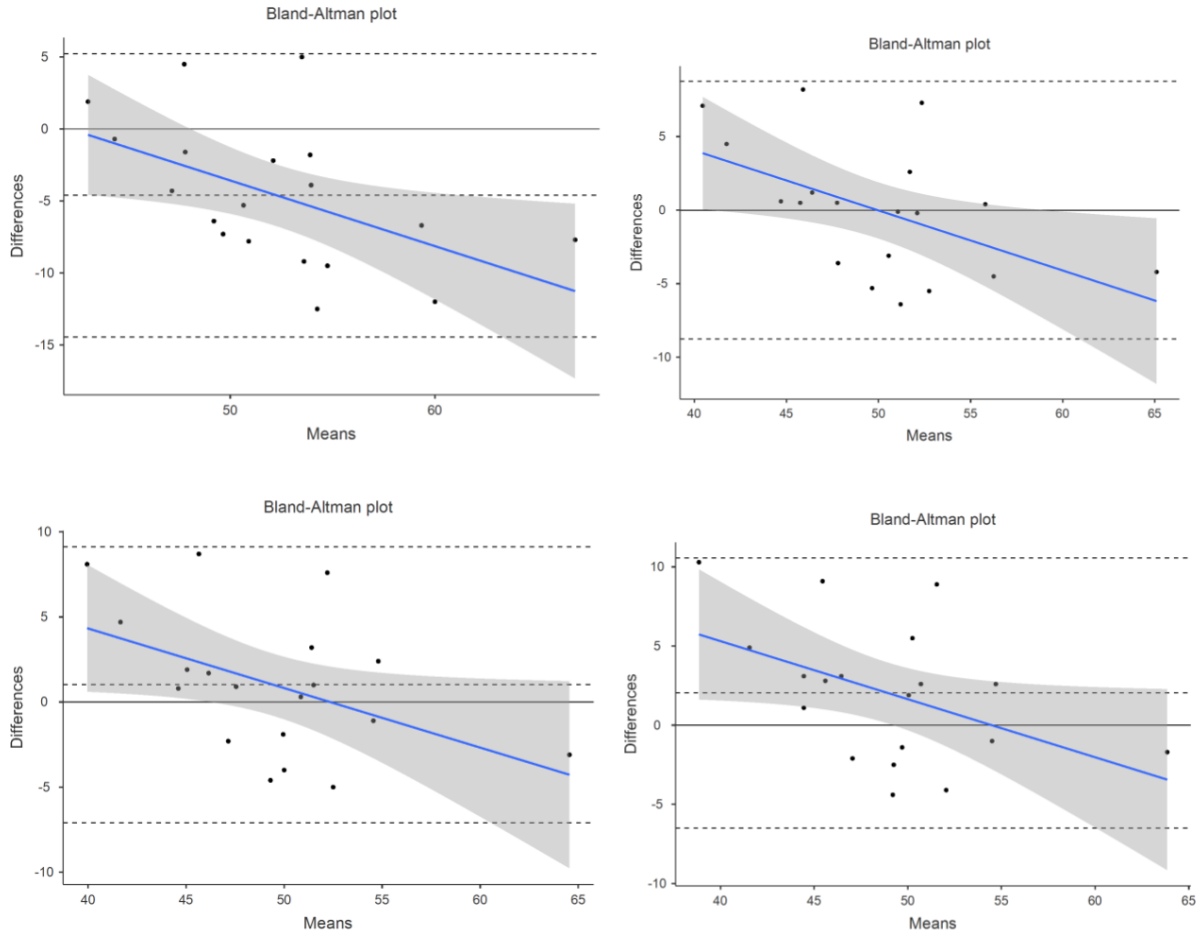

Figure 4.1. VO2 Bland-Altman Plot of fēnix 6 Compared to Laboratory VO2max Values.

4-sec average in top left, 15-sec average in top right, 30-sec average in bottom left, 1-min average in bottom right. X-axis units of measurement are VO2max (ml/kg/min) and Y-axis units is the difference between the two measurements (test device vs criterion).

Pulse Oximetry

Error analysis showed that the fēnix 6 BOS values had a MAPE of less than 10% for all four conditions and the combined data (see Table 2 and appendix Table A.3). Correlation analysis did not produce a CCC > 0.7 for any conditions, including the combined data (see Table 2 and appendix Table 3). Equivalence testing via TOST test was violated for all four conditions but was met for the combined data (see Table 2 and appendix Table A.3). Bland-Altman bias values and 95% confidence intervals can be found in Table 2 for the combined data and appendix files for individual conditions. The associated plots can be found for the combined data in Figure 3. The total number of measurements the fēnix 6 generated was 52, for a total success rate (or data availability rate) of 59%. This means that when prompted for a blood oxygen saturation measurement, it only provided data 59% of the time.

Table 4.2. Validity Statistics for Garmin Blood Oxygen Saturation Estimates

	fēnix 6 Blood Oxygen Saturation Measurement (%)	Criterion Blood Oxygen Saturation Measurement (%)
Mean	95.44%	92.06%
Standard Deviation	1.60%	8.17%
MAPE		**4.29%**
Pearson Correlation		0.18
Lin's Concordance		0.10
Bland-Altman Bias		1.12 (-0.34, 2.57)
TOST Test p-value (Upper)		0.13
TOST Test p-value (Lower)		0.02

Blood oxygen saturation measurements, measured via pulse oximetry in Garmin fēnix 6 and criterion device. Descriptive and validation statistics results for n=22 (52 distinct fēnix 6 values from all conditions and participants). Bland-Altman bias values and 95% confidence intervals provided. Values that met the predetermined validation criteria are bolded.

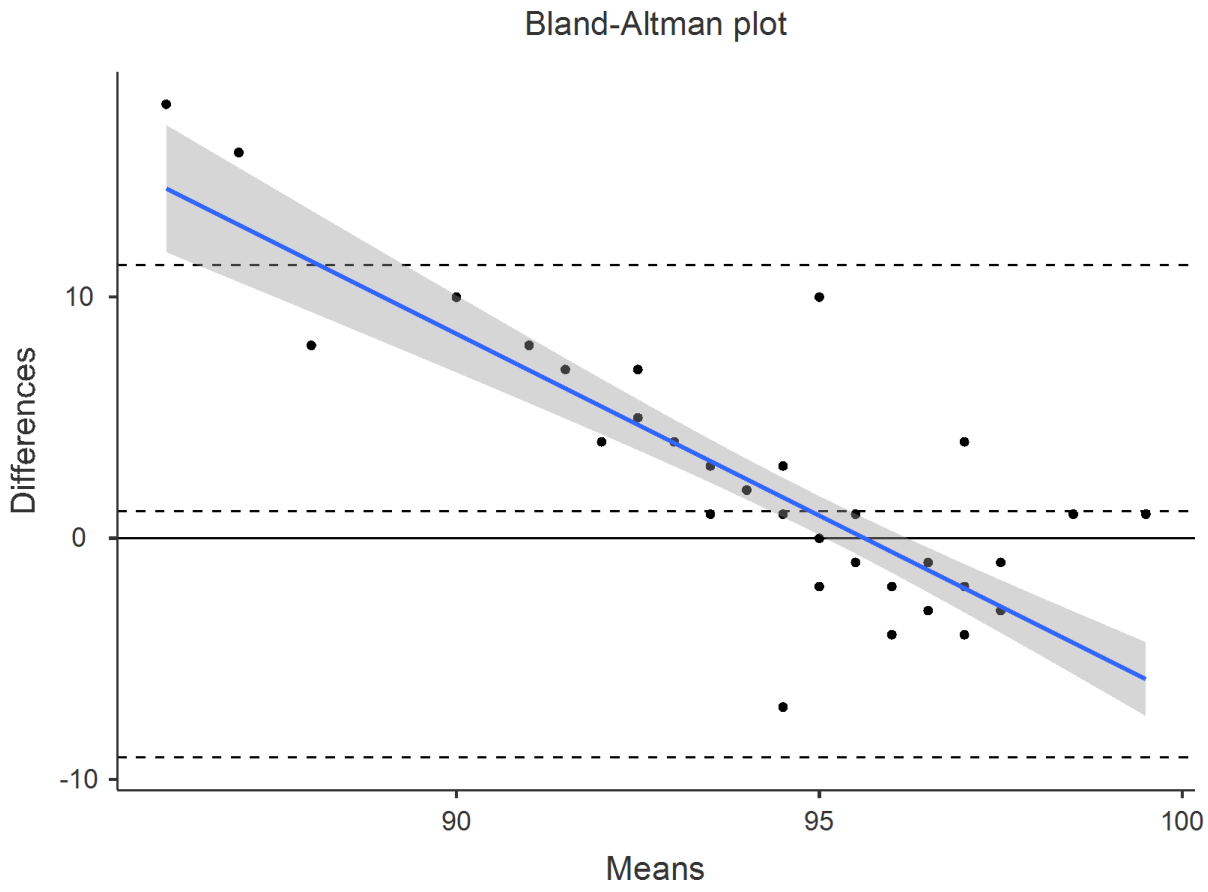

Figure 4.2. Bland-Altman Plots for the Combined Pulse Oximetry Data.

Data in figure contains all four conditions hypoxia/normoxia and anterior/posterior watch placement. X-axis unit of measurement is percent oxygen saturation (%) and Y-axis units is the difference between the two measurements (test device vs criterion).

Discussion

In this study, the validity of VO2max estimates and BOS values measured via pulse oximetry in WT was compared to gold standard measurements. Based on the pre-established validation criteria, the fēnix 6 has acceptable accuracy in its estimation of VO2max and corresponds closely to the 15-sec and 30-sec averaged timeframes. The measurements of BOS via pulse oximetry did not have acceptable accuracy for any condition or the combined data. As the appropriate use-cases of these devices is discussed, it is important to note that these are consumer-grade devices, not medical devices. Thus, they are not subject to FDA regulation (or any other governing body) in terms of accuracy and effectiveness. VO2max and pulse oximeters have an important role in monitoring the health of an individual, including general health and fitness levels and those with potential cardiovascular disease (CVD) and pulmonary diseases. While these devices are being used for measuring variables in diseased populations, they are not intended for that purpose. Despite this, researchers, health-care professionals, and public health officials are utilizing WT to track these metrics for scientific, policy, and health-care related purposes (Burnham et al., 2018; Greiwe & Nyenhuis, 2020; Iqbal et al., 2016; Ming et al., 2020; Park & Jayaraman, 2003; Phillips et al., 2018; Qaddoori et al., 2023). This illustrates the need for independent evaluation of these devices, in terms of validity and reliability, compared to gold-standard measurements. Wearable technology has the potential to revolutionize public health and physiology research, due to its constant monitoring and widespread availability (Carrier et al., 2020a; Wright et al., 2017). Thus, researchers, health-care professionals, public health officials, and scientific journals should be invested in the independent validation of these devices to further several scientific fields.

Wearable technology can generate an estimate of VO2max through HR, as the linear relationship between HR and VO2 is well established (*Aerobic Fitness Level (VO_2max) Estimation – Firstbeat White*

Paper. 2017). The fēnix 6 measures HR and speed and utilizes a linear extrapolation up to the estimated max HR, based an individual's age, to determine VO2max. While this can be accomplished simply with the watch and built in photoplethysmography (PPG) based HR monitor, an accessory HR monitor that is placed on the chest and utilizes ECG technology to determine HR can also be used. The PPG sensors common in many watch-based wearable devices have been shown to be much less accurate at reading HR during exercise than ECG-based HR monitors, mainly due to the PPG sensors susceptibility to motion artifacts during movement (Carrier et al., 2021; Chow & Yang, 2020; Estepp et al., 2014; Lu & Yang, 2009; Terbizan et al., 2002). ECG-based HR monitors have been recommended for use during exercise, which was observed in the current investigation. While WT represents an improvement in availability in tracking physiological metrics, such as VO2max, field-based maximal and submaximal tests to estimate VO2max have been around for decades (Zwiren et al., 1991). A meta-analysis detailing the performance of these submaximal predictive equations compared to gold-standard testing found that they have a correlation range of r = 0.57 to 0.92 (Evans et al., 2015). The current investigation found an r value of 0.78 for the 15-sec and 30-sec timeframes. Previous studies have found the Garmin fēnix 3 to have correlations of up to 0.92 (Carrier et al., 2020b), equal to the best submaximal equations that have been developed, in terms of correlation values. Although comparing these devices solely based on correlation gives an imperfect view of their validity, accuracy, and reliability, they do offer some comparative value.

Having an accurate estimate of VO2max can be very useful, as it represents an important metric to determine a person's health status. VO2max is a reliable predictor for overall cardiorespiratory fitness (CRF), which is an independent risk factor for all-cause and disease-specific mortality (Harber et al., 2017; Lee et al., 2011; Qiu et al., 2021). Meaning, an individual with a low VO2max value will be at a higher risk of mortality due only to that metric, regardless of any other health metrics. The American Heart Association has released a lengthy review and position statement endorsing regular measurement

of CRF in clinical practice. They state, “A growing body of epidemiological and clinical evidence demonstrates not only that CRF is a potentially stronger predictor of mortality than established risk factors such as smoking, hypertension, high cholesterol, and type 2 diabetes mellitus, but that the addition of CRF to traditional risk factors significantly improves the reclassification of risk for adverse outcomes” (Ross et al., 2016). As assessment of CRF is ideally performed through a maximal exercise test and measurement of oxygen consumption and carbon dioxide production through a metabolic system. Unfortunately, that is not possible for many people who cannot complete a maximal exercise test (those with CVD, musculoskeletal diseases, pulmonary diseases, etc.) or those who cannot afford the cost of laboratory measurements. Wearable technology has the potential to evaluate VO2max through a relatively light exercise bout (as is the case with the current device being tested) or even at rest (as is the case with other wearable devices). Thus, an accurate estimate of VO2max has the potential to influence personal health measures, as well as provide greater insights into the public health status for researchers and policy makers. As the fēnix 6 was found to generate accurate estimates of VO2max, individual recreational users, and possibly researchers, public health officials, and health-care professionals can trust the values generated by the device. However, researchers and health-care workers may want to utilize a more stringent validation threshold than what has been employed in the current investigation.

In addition to the role of VO2max in personal health, it is also an important measure for endurance athletes. VO2max is among the most important single measure to determine performance in an endurance event and is considered by many to be the single most important metric in determining performance (Bassett & Howley, 2000; Joyner & Coyle, 2008; Kenney et al., 2021). Having the ability to know an athletes VO2max allows for improved training programs to be developed that are tailored to the athlete’s specific fitness level. As gold-standard methods of determining VO2max can be expensive and time consuming, they are not a practical option for many recreational athletes or teams. Wearable

technology can represent a cost-effective method of determining aerobic capacity for individuals, as well as teams. These devices can also generate a VO2max value during the course of normal training, eliminating the need to take a day off from training for testing purposes. It also has the added benefit of constant monitoring, allowing small changes in aerobic capacity to influence the training protocol.

Measuring BOS via pulse oximetry is a well-established, and widely used method in clinical settings. The introduction of pulse oximetry into smart watches and other wearable devices is a recent advancement. Pulse oximeters measure BOS by broadcasting pulses of light and measuring the reflection via PPG sensors to monitor changes in blood oxygen concentration. This technology may prove to be an important way to monitor a person's disease status and health metrics, especially those with pulmonary diseases, such as asthma, emphysema, and chronic obstructive pulmonary disease (COPD). However, independent validation of these devices will need to be completed in order to trust the measures. It can also be useful for athletes who travel to altitude to monitor the acclimatization process, such as hikers, mountaineers, or other athletes travelling to higher altitudes than their current altitude (Luks & Swenson, 2011). While the device tested in the current investigation performed poorly, especially during the hypoxic conditions, it may be of interest to future researchers to test the ability to accurately measure BOS throughout the day, rather than on-demand. However, as we have mentioned previously, PPG sensors are susceptible to motion artifacts, and could have similar issues with accuracy when measured throughout the day. Some research has demonstrated that desaturations below 50% can be observed when patients are moving during testing (Chan et al., 2013). With the severe limitations in terms of the accuracy of this device, especially during hypoxic conditions, those looking to use this device to measure acclimatization when at altitude should look elsewhere for accurate measurements.

For the current investigation, we have used the generally accepted thresholds of MAPE < 10% and CCC > 0.7. However, universal agreement for thresholds or even analytical tests to determine validity have not been established. As we recruited from the general population for this study, the fairly liberal thresholds of 10% and 0.7 seemed appropriate. However, those looking to use this device in higher level athletics, public health and/or physiology research, and healthcare may seek more conservative thresholds to determine appropriate use-cases. In the future, a tiered threshold system could be established, to better understand the appropriate use-cases of these devices. In terms of analytical tests, we have decided only to use MAPE and CCC in the determination of validity. However, we have also included bias assessments (Bland-Altman analysis) and equivalence testing (TOST test). These have all been suggested as appropriate analytical techniques to determine validity, though are not always common in other validation literature (Carrier et al., 2020a; van Lier et al., 2020; Welk et al., 2019). For instance, equivalence testing is especially absent from much of the validation literature. We have included all for the benefit of the reader and because we view them as appropriate tests to determine validity. However, because thresholds have not been established for these additional tests, we have not included them in our validity thresholds.

Conclusion

In this study, we tested the Garmin fēnix 6 VO2max estimate and blood oxygen saturation values measured via pulse oximetry for accuracy compared to gold-standard, laboratory measurements. The fēnix 6 showed acceptable accuracy for VO2max, and most closely aligned with the 15-sec and 30-sec timeframes. The fēnix 6 did not show acceptable accuracy for blood oxygen levels for any condition, or the combined analysis. Therefore, the Garmin fēnix 6 may reasonably be expected to generate an accurate estimate of an individual's VO2max based on 15-sec or 30-sec aggregated data if more accurate,

laboratory tests are not available. In addition, the fēnix 6 will not generate an accurate estimate of an individual's blood oxygen levels, either in normoxia/hypoxia, or anterior/posterior watch placement on the wrist.

Chapter 4 References

Aerobic Fitness Level (VO_2max) Estimation – Firstbeat White Paper. (2017, 6/30/). https://www.firstbeat.com. Retrieved 07/14/2022, from https://www.firstbeat.com/en/aerobic-fitness-level-vo%e2%82%82max-estimation-firstbeat-white-paper-2/

Bassett, D. R., & Howley, E. T. (2000). Limiting factors for maximum oxygen uptake and determinants of endurance performance. Medicine and Science in Sports and Exercise, 32(1), 70-84.

Burnham, J. P., Lu, C., Yaeger, L. H., Bailey, T. C., & Kollef, M. H. (2018). Using wearable technology to predict health outcomes: a literature review. Journal of the American Medical Informatics Association, 25(9), 1221-1227.

Carrier, B., Barrios, B., Jolley, B. D., & Navalta, J. W. (2020a). Validity and Reliability of Physiological Data in Applied Settings Measured by Wearable Technology: A Rapid Systematic Review. Technologies, 8(4), 70.

Carrier, B., Creer, A., Williams, L. R., Holmes, T. M., Jolley, B. D., Dahl, S., Weber, E., & Standifird, T. (2020b). Validation of garmin fenix 3 HR fitness tracker biomechanics and metabolics (VO2max). Journal for the Measurement of Physical Behaviour, 3(4), 331-337.

Carrier, B., Salatto, R. W., Davis, D. W., Sertic, J. V. L., Barrios, B., Cater, P., & Navalta, J. W. (2021). Assessing the Validity of Several Heart Rate Monitors in Wearable Technology While Mountain Biking. Paper presented at the , 14(1) 18.

Chan, E. D., Chan, M. M., & Chan, M. M. (2013). Pulse oximetry: understanding its basic principles facilitates appreciation of its limitations. Respiratory Medicine, 107(6), 789-799.

Chow, H., & Yang, C. (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(4), e14707.

Clark, E. (2019, Oct. 17,). Is Your Fitness Tracker Helping or Hurting Your Health? themanifest.com. Retrieved 03/08/2024, from https://themanifest.com/app-development/fitness-tracker-helping-hurting-health

Dhingra, L. S., Aminorroaya, A., Oikonomou, E. K., Nargesi, A. A., Wilson, F. P., Krumholz, H. M., & Khera, R. (2023). Use of Wearable Devices in Individuals With or at Risk for Cardiovascular Disease in the US, 2019 to 2020. JAMA Network Open, 6(6), e2316634.

Dünnwald, T., Kienast, R., Niederseer, D., & Burtscher, M. (2021). The use of pulse oximetry in the assessment of acclimatization to high altitude. Sensors, 21(4), 1263.

Estepp, J. R., Blackford, E. B., & Meier, C. M. (2014). Recovering pulse rate during motion artifact with a multi-imager array for non-contact imaging photoplethysmography. Paper presented at the 2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 1462-1469.

Evans, H. J., Ferrar, K. E., Smith, A. E., Parfitt, G., & Eston, R. G. (2015). A systematic review of methods to predict maximal oxygen uptake from submaximal, open circuit spirometry in healthy adults. Journal of Science and Medicine in Sport, 18(2), 183-188.

Greiwe, J., & Nyenhuis, S. M. (2020). Wearable technology and how this can be implemented into clinical practice. Current Allergy and Asthma Reports, 20, 1-10.

Harber, M. P., Kaminsky, L. A., Arena, R., Blair, S. N., Franklin, B. A., Myers, J., & Ross, R. (2017). Impact of cardiorespiratory fitness on all-cause and disease-specific mortality: advances since 2009. Progress in Cardiovascular Diseases, 60(1), 11-20.

Iqbal, M. H., Aydin, A., Brunckhorst, O., Dasgupta, P., & Ahmed, K. (2016). A review of wearable technology in medicine. Journal of the Royal Society of Medicine, 109(10), 372-380.

Joyner, M. J., & Coyle, E. F. (2008). Endurance exercise performance: the physiology of champions. The Journal of Physiology, 586(1), 35-44.

Kenney, W. L., Wilmore, J. H., & Costill, D. L. (2021). Physiology of sport and exercise. Human kinetics.

Lee, D., Sui, X., Artero, E. G., Lee, I., Church, T. S., McAuley, P. A., Stanford, F. C., Kohl III, H. W., & Blair, S. N. (2011). Long-term effects of changes in cardiorespiratory fitness and body mass index on all-cause and cardiovascular disease mortality in men: the Aerobics Center Longitudinal Study. Circulation, 124(23), 2483-2490.

Lu, G., & Yang, F. (2009). Limitations of oximetry to measure heart rate variability measures. Cardiovascular Engineering, 9(3), 119-125.

Luks, A. M., & Swenson, E. R. (2011). Pulse oximetry at high altitude. High Altitude Medicine & Biology, 12(2), 109-119.

Ming, D. K., Sangkaew, S., Chanh, H. Q., Nhat, P. T., Yacoub, S., Georgiou, P., & Holmes, A. H. (2020). Continuous physiological monitoring using wearable technology to inform individual management of infectious diseases, public health and outbreak responses. International Journal of Infectious Diseases, 96, 648-654.

Newsome, A., Reed, R., Sansone, J., Batrakoulis, A., McAvoy, C., & Parrott, M.,W. (2024). 2024 ACSM Worldwide Fitness Trends: Future Directions of the Health and Fitness Industry. ACSM's Health & Fitness Journal, 28(1) https://journals.lww.com/acsm-healthfitness/fulltext/2024/01000/2024_acsm_worldwide_fitness_trends__future.7.aspx

Park, S., & Jayaraman, S. (2003). Enhancing the quality of life through wearable technology. IEEE Engineering in Medicine and Biology Magazine, 22(3), 41-48.

Phillips, S. M., Cadmus-Bertram, L., Rosenberg, D., Buman, M. P., & Lynch, B. M. (2018). Wearable technology and physical activity in chronic disease: opportunities and challenges. American Journal of Preventive Medicine, 54(1), 144-150.

Qaddoori, S. L., Fathi, I., Hammoudy, M. A., & Ali, Q. I. (2023). Advancing Public Health Monitoring through Secure and Efficient Wearable Technology. International Journal of Safety & Security Engineering, 13(6)

Qiu, S., Cai, X., Sun, Z., Wu, T., & Schumann, U. (2021). Is estimated cardiorespiratory fitness an effective predictor for cardiovascular and all-cause mortality? A meta-analysis. Atherosclerosis, 330, 22-28.

Ross, R., Blair, S. N., Arena, R., Church, T. S., Després, J., Franklin, B. A., Haskell, W. L., Kaminsky, L. A., Levine, B. D., & Lavie, C. J. (2016). Importance of assessing cardiorespiratory fitness in clinical practice: a case for fitness as a clinical vital sign: a scientific statement from the American Heart Association. Circulation, 134(24), e653-e699.

Sinex, J. A., & Chapman, R. F. (2015). Hypoxic training methods for improving endurance exercise performance. Journal of Sport and Health Science, 4(4), 325-332.

Terbizan, D. J., Dolezal, B. A., & Albano, C. (2002). Validity of seven commercially available heart rate monitors. Measurement in Physical Education and Exercise Science, 6(4), 243-247.

Thompson, W. R. (2015). Worldwide survey of fitness trends for 2016. ACSM's Health & Fitness Journal, 19(6), 9-18.

Thompson, W. R. (2019). Worldwide survey of fitness trends for 2020. ACSM's Health & Fitness Journal, 23(6), 10-18.

Thompson, W. R. (2022). Worldwide survey of fitness trends for 2022. ACSM's Health & Fitness Journal, 26(1), 11-20.

Thompson, W. R. (2016). WORLDWIDE SURVEY OF FITNESS TRENDS FOR 2017. ACSM's Health & Fitness Journal, 20(6), 8-17. 10.1249/FIT.0000000000000252

Thompson, W. R. (2017). WORLDWIDE SURVEY OF FITNESS TRENDS FOR 2018: The CREP Edition. ACSM's Health & Fitness Journal, 21(6), 10-19. 10.1249/FIT.0000000000000341

Thompson, W. R. (2018). WORLDWIDE SURVEY OF FITNESS TRENDS FOR 2019. ACSM's Health & Fitness Journal, 22(6), 10-17. 10.1249/fit.0000000000000438

Thompson, W. R. (2021). Worldwide Survey of Fitness Trends for 2021. ACSM's Health & Fitness Journal, 25(1), 10-19. 10.1249/FIT.0000000000000631

Thompson, W. R. (2023). Worldwide Survey of Fitness Trends for 2023. ACSM's Health & Fitness Journal, 27(1) https://journals.lww.com/acsm-healthfitness/fulltext/2023/01000/worldwide_survey_of_fitness_trends_for_2023.6.aspx

van Lier, H.,G., Pieterse, M. E., Garde, A., Postel, M. G., de Haan, H.,A., Vollenbroek-Hutten, M., Schraagen, J. M., & Noordzij, M. L. (2020). A standardized validity assessment protocol for physiological signals from wearable technology: Methodological underpinnings and an application to the E4 biosensor. Behavior Research Methods, 52, 607-629.

Welk, G. J., Bai, Y., Lee, J., Godino, J., Saint-Maurice, P. F., & Carr, L. (2019). Standardizing analytic methods and reporting in activity monitor validation studies. Medicine and Science in Sports and Exercise, 51(8), 1767.

Wright, S., Brown, T., Collier, S., & Sandberg, K. (2017). How consumer physical activity monitors could transform human physiology research. American Journal of Physiology-Regulatory Integrative and Comparative Physi, 312(3), R358-R367. 10.1152/ajpregu.00349.2016

Zwiren, L. D., Freedson, P. S., Ward, A., Wilke, S., & Rippe, J. M. (1991). Estimation of VO2max: a comparative analysis of five exercise tests. Research Quarterly for Exercise and Sport, 62(1), 73-78.

Chapter 5 - Conclusion

This dissertation marks a significant advancement in the field of wearable technology, addressing critical gaps and setting new standards for the evaluation and application of wearable technology devices in various contexts. The dissertation consisted of three primary works, the first established a risk of bias tool to evaluate validity and reliability studies that utilize wearable technology, known as the **WEA**rable technology **R**isk of **B**ias and **O**bjectivity **T**ool (WEAR-BOT). The second was a systematic review and meta-analysis looking at physiological variables measured during exercise via wearable technology, with WEAR-BOT analysis. The final piece was a validation study as an example of the type of research that the WEAR-BOT can evaluate, or guide researchers through a methodological design for their own studies.

As stated earlier in the dissertation, wearable technology may be viewed as an emerging field of science, as it has only been in the last ~10 years that the technology has progressed enough to be genuinely useful to the user. The technology now has broad adoption, which will only continue as the technology advances and costs are reduced. This has led to some inappropriate applications and analyses of wearable technology, especially in terms of validity and reliability studies which was examined earlier. Despite this, it continues to be an important means of collecting data and informing decisions for individual users and researchers. As the technology matures, and adoption continues to spread to many different fields, we can expect to improve our understanding of human physiology, human behavior, and many other areas that the technology will infiltrate. Research will continue to be proposed and performed using wearable technology, some possibilities for future research can be seen below.

The foundation laid by this dissertation paves the way for several future projects aimed at further refining and expanding our understanding and application of wearable technology. Planned systematic reviews using the WEAR-BOT tool will extend into areas such as physical variables and special populations, broadening the scope of validated wearable technology applications. These reviews are

crucial for identifying gaps in the current research landscape and setting the agenda for future investigations.

Furthermore, the need to establish appropriate thresholds for validity, likely utilizing a tiered approach for different use-cases, signifies a pioneering step towards a personalized and context-specific application of wearable technologies. By differentiating between general population use, recreational athletics, collegiate and professional athletics, military use, and research use, this endeavor will cater to the varying demands for accuracy and reliability across different sectors, thereby enhancing the utility and adoption of wearable devices.

Incorporating wearable technology into public health research opens up vast possibilities for exploring the relationship between physiological or physical variables and health or disease status. This approach not only contributes to the body of knowledge on disease prevention and management but also leverages wearable technology as a tool for public health advancement.

There is also a need to establish an easily accessible database, compiling results from different validation and reliability studies, which would represent a strategic move towards transparency and accessibility in wearable technology research. Such a database would serve as a valuable resource for researchers, practitioners, and consumers alike, facilitating informed decision-making and fostering collaborative efforts across disciplines.

In conclusion, this dissertation not only enriches the scientific quality of wearable technology research but also charts a course for future inquiries and applications. By addressing current challenges and setting forth ambitious future projects, this work contributes meaningfully to the field, enhancing the reliability, validity, and utility of wearable technology. The implications of this research are far-reaching, promising to impact not only future scientific investigations but also the integration of wearable technology in everyday life, public health initiatives, and specialized professional practices.

Appendix

Table A.1. Compiled MAPE Results

Manufacturer	Model	Author	Exercise Modality	Exercise Subgroup	Population	Sample Size	HR	EE	RR	O2 Saturation	Skin Temp	VO2 max	Fluid Loss
SpectroPhon	Dehydration Body Monitor (DBM) Paired with Samsung Gear S2	Rodin et al.	Walking		Male	120							10.16%
SpectroPhon	Dehydration Body Monitor (DBM) Paired with Samsung Gear S2	Rodin et al.	Walking		Female	120							8.96%
SpectroPhon	Dehydration Body Monitor (DBM) Paired with Samsung Gear Fit2	Rodin et al.	Walking		Male	120							10.32%
SpectroPhon	Dehydration Body Monitor (DBM) Paired with	Rodin et al.	Walking		Female	120							9.52%

	Samsung Gear Fit2												
Jabra	Elite Sport Earbuds	Reece et al.	Walking		General	23	7.91 %						
Jabra	Elite Sport Earbuds	Reece et al.	Walking	Fast Walking	General	23	10.80%						
Jabra	Elite Sport Earbuds	Reece et al.	Running		General	23	7.91 %						
Jabra	Elite Sport Earbuds	Reece et al.	Cycling		General	23	7.15 %						
Scosche	Rhythm 24	Reece et al.	Walking		General	23	2.33 %						
Scosche	Rhythm 24	Reece et al.	Walking	Fast Walking	General	23	1.54 %						
Scosche	Rhythm 24	Reece et al.	Running		General	23	1.24 %						
Scosche	Rhythm 24	Reece et al.	Cycling		General	23	0.90 %						
Apple	Apple Watch Series 4	Reece et al.	Walking		General	23	1.50 %						
Apple	Apple Watch Series 4	Reece et al.	Walking	Fast Walking	General	23	1.17 %						
Apple	Apple Watch Series 4	Reece et al.	Running		General	23	0.92 %						
Apple	Apple Watch Series 4	Reece et al.	Cycling		General	23	0.62 %						

Garmin	Forerunner 735 XT	Reece et al.	Walking		General	23	8.75%						
Garmin	Forerunner 735 XT	Reece et al.	Walking	Fast Walking	General	23	12.63%						
Garmin	Forerunner 735 XT	Reece et al.	Running		General	23	11.22%						
Garmin	Forerunner 735 XT	Reece et al.	Cycling		General	23	7.56%						
Xiaomi	Mi Band 2	Paradiso, Colino, & Liu	Cycling	Light Cycling	General	14	16.80%						
Xiaomi	Mi Band 2	Paradiso, Colino, & Liu	Cycling	Light Cycling	General	14	14.20%						
Xiaomi	Mi Band 2	Paradiso, Colino, & Liu	Cycling	Light Cycling	General	14	21.90%						
Xiaomi	Mi Band 2	Paradiso, Colino, & Liu	Cycling	Light Cycling	General	14	17.40%						
Xiaomi	Mi Band 2	Paradiso, Colino, & Liu	Cycling	Moderate Cycling	General	14	32.20%						
Xiaomi	Mi Band 2	Paradiso, Colino, & Liu	Cycling	Moderate Cycling	General	14	27.30%						
Xiaomi	Mi Band 2	Paradiso, Colino, & Liu	Cycling	Vigorous Cycling	General	14	38.10%						
Xiaomi	Mi Band 2	Paradiso, Colino, & Liu	Cycling	Vigorous Cycling	General	14	37.40%						

Xiaomi	Mi Band 2	Paradiso, Colino, & Liu	Stairs		General	14	8.00%						
Xiaomi	Mi Band 2	Paradiso, Colino, & Liu	Stairs		General	14	7.80%						
Fitbit	Charge 2	O'Driscoll et al.	Walking		General	14	25.00%	44.00%					
Fitbit	Charge 2	O'Driscoll et al.	Walking	Incline Walking	General	14	17.00%	31.00%					
Fitbit	Charge 2	O'Driscoll et al.	Running		General	14	8.00%	15.00%					
Fitbit	Charge 2	O'Driscoll et al.	Running	Incline Running	General	14	5.00%	12.00%					
Fitbit	Charge 2	O'Driscoll et al.	Cycling	Light Cycling	General	14	12.00%	40.00%					
Fitbit	Charge 2	O'Driscoll et al.	Cycling	Moderate Cycling	General	14	16.00%	39.00%					
Sensewear	Armband Mini	O'Driscoll et al.	Walking		General	14		14.00%					
Sensewear	Armband Mini	O'Driscoll et al.	Walking	Incline Walking	General	14		13.00%					
Sensewear	Armband Mini	O'Driscoll et al.	Running		General	14		15.00%					
Sensewear	Armband Mini	O'Driscoll et al.	Running	Incline Running	General	14		15.00%					
Sensewear	Armband Mini	O'Driscoll et al.	Cycling	Light Cycling	General	14		31.00%					
Sensewear	Armband Mini	O'Driscoll et al.	Cycling	Moderate Cycling	General	14		35.00%					
Adidas	Smart Sports Bra	Navalta, Ramirez et al.	Run		General	24	13.57%						

Adidas	Smart Sports Bra	Navalta, Ramirez et al.	Walk		General	24	9.56 %						
Berlei	Sports Bra	Navalta, Ramirez et al.	Run		General	24	0.58 %						
Berlei	Sports Bra	Navalta, Ramirez et al.	Walk		General	24	0.61 %						
Sensoria Fitness	Biometric Sports Bra	Navalta, Ramirez et al.	Run		General	24	4.00 %						
Sensoria Fitness	Biometric Sports Bra	Navalta, Ramirez et al.	Walk		General	24	1.91 %						
Suunto	Spartan Sport Watch + Chest Strap	Navalta, Montes et al.	Running	Trail Running (Uphill)	General	21	1.50 %						
Suunto	Spartan Sport Watch + Chest Strap	Navalta, Montes et al.	Running	Trail Running (Downhill)	General	21	2.20 %						
Garmin	Fenix 5	Navalta, Montes et al.	Running	Trail Running (Uphill)	General	21	13.70%						
Garmin	Fenix 5	Navalta, Montes et al.	Running	Trail Running (Downhill)	General	21	13.40%						
Jabra	Elite Sport Earbuds	Navalta, Montes et al.	Running	Trail Running (Uphill)	General	21	24.50%						
Jabra	Elite Sport Earbuds	Navalta, Montes et al.	Running	Trail Running	General	21	20.60%						

				(Downhill)									
Motiv	Motiv Ring	Navalta, Montes et al.	Running	Trail Running (Uphill)	General	21	16.40%						
Motiv	Motiv Ring	Navalta, Montes et al.	Running	Trail Running (Downhill)	General	21	15.40%						
Scosche	Rhythm+	Navalta, Montes et al.	Running	Trail Running (Uphill)	General	21	6.20%						
Scosche	Rhythm+	Navalta, Montes et al.	Running	Trail Running (Downhill)	General	21	3.80%						
Biovotion AG	Everion	Havemann et al.	Walking		General	20	16.00%		22.90%	3.80%	7.80%		
Biovotion AG	Everion	Havemann et al.	Cycling		General	20	3.00%		27.10%	1.50%	8.20%		
Biovotion AG	Everion	Havemann et al.	Walking		General	20	23.40%		22.70%	3.00%	8.80%		
Biovotion AG	Everion	Havemann et al.	Cycling		General	20	2.90%		26.80%	1.30%	9.10%		
MediBioSense	VitalPatch	Havemann et al.	Walking		General	20	13.40%		8.30%		1.20%		
MediBioSense	VitalPatch	Havemann et al.	Cycling		General	20	2.30%		6.70%		1.70%		
Fitbit	Charge 3	Havemann et al.	Walking		General	20	20.20%						
Fitbit	Charge 3	Havemann et al.	Cycling		General	20	6.10%						
Oura	Gen2	Kristiansson et al.	Walking	Slow Walking	General	32		14.90%					

Oura	Gen2	Kristiansson et al.	Walking	Fast Walking	General	32		47.00%					
Oura	Gen2	Kristiansson et al.	Running	Slow Running	General	32		6.00%					
Oura	Gen2	Kristiansson et al.	Running	Fast Running	General	32		81.70%					
Oura	Gen2	Kristiansson et al.	Running	Sprinting	General	32		225.60%					
Fitbit	Versa	Jagim et al.	Running	Graded Exercise Test	General	20	11.60%	9.60%					
Polar	Ignite	Jagim et al.	Running	Graded Exercise Test	General	20	11.00%	16.70%					
Polar	TeamPro Sensor	Jagim et al.	Running	Graded Exercise Test	General	20	3.00%	13.80%					
Fitbit	Charge 4	Jachymek, et al.	Running	Graded Exercise Test	General	31	10.19%						
Xiaomi	Mi Band 5	Jachymek, et al.	Running	Graded Exercise Test	General	31	6.89%						
Xiaomi	Mi Band 2	Chow & Yang	Mixed		General	20	7.69%						
Xiaomi	Mi Band 2	Chow & Yang	Mixed		Elderly	20	6.04%						
Garmin	Vivosmart HR+	Chow & Yang	Mixed		General	20	3.77%						
Garmin	Vivosmart HR+	Chow & Yang	Mixed		Elderly	20	4.73%						

Apple	Apple Watch Series 6	Støve et al.	Strength Training	Barbell Exercises	General	29	6.30%						
Apple	Apple Watch Series 7	Støve et al.	Strength Training	Barbell Exercises	General	29	4.00%						
Apple	Apple Watch Series 8	Støve et al.	Strength Training	Recovery	General	29	1.90%						
Apple	Apple Watch Series 9	Støve et al.	Strength Training	Barbell Exercises	General	29	5.70%						
Apple	Apple Watch Series 10	Støve et al.	Strength Training	Barbell Exercises	General	29	5.70%						
Apple	Apple Watch Series 11	Støve et al.	Strength Training	Recovery	General	29	3.00%						
Apple	Apple Watch Series 12	Støve et al.	Strength Training	Dumbbell Exercises	General	29	10.40%						
Apple	Apple Watch Series 13	Støve et al.	Strength Training	Dumbbell Exercises	General	29	14.00%						
Apple	Apple Watch Series 14	Støve et al.	Strength Training	Recovery	General	29	3.50%						
Apple	Apple Watch Series 15	Støve et al.	Strength	Machine	General	29	5.50%						

			Training	Exercises									
Apple	Apple Watch Series 16	Støve et al.	Strength Training	Machine Exercises	General	29	3.10%						
Apple	Apple Watch Series 17	Støve et al.	Strength Training	Recovery	General	29	1.60%						
Apple	Apple Watch Series 18	Støve et al.	Strength Training	Bodyweight Exercises	General	29	11.90%						
Apple	Apple Watch Series 19	Støve et al.	Strength Training	Bodyweight Exercises	General	29	13.20%						
Apple	Apple Watch Series 20	Støve et al.	Strength Training	Recovery	General	29	2.80%						
Whoop	Band 3.0	Støve et al.	Strength Training	Barbell Exercises	General	29	12.10%						
Whoop	Band 3.1	Støve et al.	Strength Training	Barbell Exercises	General	29	14.70%						
Whoop	Band 3.2	Støve et al.	Strength Training	Recovery	General	29	5.60%						
Whoop	Band 3.3	Støve et al.	Strength Training	Barbell Exercises	General	29	9.90%						

Whoop	Band 3.4	Støve et al.	Strength Training	Barbell Exercises	General	29	13.80%						
Whoop	Band 3.5	Støve et al.	Strength Training	Recovery	General	29	8.10%						
Whoop	Band 3.6	Støve et al.	Strength Training	Dumbbell Exercises	General	29	8.20%						
Whoop	Band 3.7	Støve et al.	Strength Training	Dumbbell Exercises	General	29	12.60%						
Whoop	Band 3.8	Støve et al.	Strength Training	Recovery	General	29	4.40%						
Whoop	Band 3.9	Støve et al.	Strength Training	Machine Exercises	General	29	12.50%						
Whoop	Band 3.10	Støve et al.	Strength Training	Machine Exercises	General	29	14.80%						
Whoop	Band 3.11	Støve et al.	Strength Training	Recovery	General	29	6.50%						
Whoop	Band 3.12	Støve et al.	Strength Training	Bodyweight Exercises	General	29	10.80%						
Whoop	Band 3.13	Støve et al.	Strength	Bodyweight	General	29	13.30%						

			Training	Exercises									
Whoop	Band 3.14	Støve et al.	Strength Training	Recovery	General	29	6.20 %						
Apple	Apple Watch Series 6	Ho, Yang, & Li	Cycling	Light Cycling	General	30	1.00 %						
Apple	Apple Watch Series 6	Ho, Yang, & Li	Cycling	Moderate Cycling	General	30	0.92 %						
Apple	Apple Watch Series 6	Ho, Yang, & Li	Cycling	Vigorous Cycling	General	30	0.91 %						
Garmin	Forerunner 945	Ho, Yang, & Li	Cycling	Light Cycling	General	30	1.16 %						
Garmin	Forerunner 945	Ho, Yang, & Li	Cycling	Moderate Cycling	General	30	1.26 %						
Garmin	Forerunner 945	Ho, Yang, & Li	Cycling	Vigorous Cycling	General	30	1.39 %						
Goldwin	C3fit IN-pulse	Hashimoto et al.	Walking		General	8	0.49 %						
Goldwin	C3fit IN-pulse	Hashimoto et al.	Running		General	8	0.67 %						
Apple	Apple Watch Series 6	Hajj-Boutros et al.	Walking		General	60	2.30 %	24.10 %					
Apple	Apple Watch Series 6	Hajj-Boutros et al.	Running		General	60	2.90 %	14.90 %					

Apple	Apple Watch Series 6	Hajj-Boutros et al.	Strength Training		General	60	1.40 %	20.00 %					
Apple	Apple Watch Series 6	Hajj-Boutros et al.	Cycling		General	60	2.20 %	17.70 %					
Polar	Vantage V	Hajj-Boutros et al.	Walking		General	60	5.50 %	15.56 %					
Polar	Vantage V	Hajj-Boutros et al.	Running		General	60	4.00 %	15.70 %					
Polar	Vantage V	Hajj-Boutros et al.	Strength Training		General	60	3.50 %	34.60 %					
Polar	Vantage V	Hajj-Boutros et al.	Cycling		General	60	5.70 %	16.40 %					
Fitbit	Sense	Hajj-Boutros et al.	Walking		General	60	4.40 %	45.10 %					
Fitbit	Sense	Hajj-Boutros et al.	Running		General	60	3.80 %	17.80 %					
Fitbit	Sense	Hajj-Boutros et al.	Strength Training		General	60	5.30 %	34.10 %					
Fitbit	Sense	Hajj-Boutros et al.	Cycling		General	60	5.20 %	26.60 %					
Garmin	fēnix 3 HR	Carrier et al.	Running		General	17						8.05%	

Garmin	Forerunner 735XT	Damasceno et al.	Mixed	Light Exercise	General	28	2.90 %						
Garmin	Forerunner 735XT	Damasceno et al.	Mixed	Moderate Exercise	General	28	3.00 %						
Garmin	Forerunner 735XT	Damasceno et al.	Mixed	Vigorous Exercise	General	28	2.90 %						
Polar	Vantage V2	Cosoli, Antognoli, Veroli, & Scalise	Mixed	Walking and Running	General	10	8.29 %						
Polar	Vantage V2	Cosoli, Antognoli, Veroli, & Scalise	Swimming		General	10	29.78%						
Garmin	Venu Sq	Cosoli, Antognoli, Veroli, & Scalise	Mixed	Walking and Running	General	10	3.60 %						
Garmin	Venu Sq	Cosoli, Antognoli, Veroli, & Scalise	Swimming		General	10	58.94%						
Polar	Ignite	Budig et al.	Running		General	36	2.57 %						
Polar	Ignite	Budig et al.	Swimming		General	36	8.61 %						
Polar	Ignite	Budig et al.	Swimming		General	36	51.08%						
Garmin	Forerunner 945	Budig et al.	Running		General	36	1.23 %						

Garmin	Forerunner 945	Budig et al.	Swimming		General	36	3.29%						
Garmin	Forerunner 945	Budig et al.	Swimming		General	36	55.32%						
Polar	Polar H7	Baek, Ha, & Park	Walking		General	15	1.27%						
Polar	Polar H7	Baek, Ha, & Park	Walking	Nordic Walking	General	15	1.28%						
Fitbit	Charge 2	Baek, Ha, & Park	Walking		General	15	3.73%						
Fitbit	Charge 2	Baek, Ha, & Park	Walking	Nordic Walking	General	15	5.73%						
Apple	Apple Watch Series 2	Støve et al.	Cycling	Light Cycling	General	30	4.90%						
Apple	Apple Watch Series 2	Støve et al.	Cycling	Moderate Cycling	General	30	3.10%						
Apple	Apple Watch Series 2	Støve et al.	Cycling	Moderate Cycling	General	30	2.30%						
Apple	Apple Watch Series 2	Støve et al.	Walking		General	30	3.60%						
Apple	Apple Watch Series 2	Støve et al.	Running		General	30	3.40%						
Apple	Apple Watch Series 2	Støve et al.	Running		General	30	3.60%						
Garmin	Forerunner 235	Støve et al.	Cycling	Light Cycling	General	30	26.90%						

Garmin	Forerunner 235	Støve et al.	Cycling	Moderate Cycling	General	30	55.70%						
Garmin	Forerunner 235	Støve et al.	Cycling	Moderate Cycling	General	30	78.10%						
Garmin	Forerunner 235	Støve et al.	Walking		General	30	9.80%						
Garmin	Forerunner 235	Støve et al.	Running		General	30	5.60%						
Garmin	Forerunner 235	Støve et al.	Running		General	30	4.60%						
Fitbit	Charge 4	Nissen et al.	Mixed		General	23	9.76%						
Fitbit	Charge 4	Nissen et al.	Walking	Slow Walking	General	23	8.12%						
Fitbit	Charge 4	Nissen et al.	Walking	Fast Walking	General	23	6.364						
Fitbit	Charge 4	Nissen et al.	Stairs		General	23	7.61%						
Fitbit	Charge 4	Nissen et al.	Strength Training	Bodyweight Exercies	General	23	11.98%						
Samsung	Galaxy Watch Active 2	Nissen et al.	Mixed		General	23	9.42%						
Samsung	Galaxy Watch Active 2	Nissen et al.	Walking	Slow Walking	General	23	9.05%						
Samsung	Galaxy Watch Active 2	Nissen et al.	Walking	Fast Walking	General	23	6.29%						

Samsung	Galaxy Watch Active 2	Nissen et al.	Stairs		General	23	6.16 %						
Samsung	Galaxy Watch Active 2	Nissen et al.	Strength Training	Bodyweight Exercies	General	23	5.51 %						

Reported mean absolute percentage errors (MAPEs) from each study reviewed.

Table A.2. Compiled Pearson Correlation Results

Manufacturer	Model	Author	Exercise Modality	Exercise Subgroup	Population	Sample Size	HR	EE	VO2max	RR	VO2	Core Body Temp	Fluid Loss
SpectroPhon	Dehydration Body Monitor (DBM) Paired with Samsung Gear S2	Rodin et al.	Walking		General	240							0.9314
SpectroPhon	Dehydration Body Monitor (DBM) Paired with Samsung Gear Fit2	Rodin et al.	Walking		General	240							0.9259
Fitbit	Charge 2	O'Driscoll et al.	Walking		General	59	0.23	0.39					
Fitbit	Charge 3	O'Driscoll et al.	Walking	Incline Walking	General	59	0.29	0.59					
Fitbit	Charge 4	O'Driscoll et al.	Running		General	49	0.66	0.7					
Fitbit	Charge 5	O'Driscoll et al.	Running	Incline Running	General	30	0.81	0.81					
Fitbit	Charge 6	O'Driscoll et al.	Cycling	Light Cycling	General	59	0.55	0.38					
Fitbit	Charge 7	O'Driscoll et al.	Cycling	Moderate Cycling	General	58	0.62	0.37					
Sensewear	Armband Mini	O'Driscoll et al.	Walking		General	59		0.62					

Sensewear	Armband Mini	O'Driscoll et al.	Walking	Incline Walking	General	59		0.56					
Sensewear	Armband Mini	O'Driscoll et al.	Running		General	49		0.69					
Sensewear	Armband Mini	O'Driscoll et al.	Running	Incline Running	General	30		0.71					
Sensewear	Armband Mini	O'Driscoll et al.	Cycling	Light Cycling	General	59		0.7					
Sensewear	Armband Mini	O'Driscoll et al.	Cycling	Moderate Cycling	General	58		0.41					
Polar	OH1	Muggeridge et al.	Cycling	Light Cycling	General	20	0.983						
Polar	OH2	Muggeridge et al.	Cycling	Vigorous Cycling	General	20	0.985						
Polar	OH3	Muggeridge et al.	Running		General	20	0.99						
Polar	OH4	Muggeridge et al.	Running	Intense Cycling	General	20	0.794						
Polar	OH5	Muggeridge et al.	Running	Sprinting	General	20	0.722						
Fitbit	Charge 3	Muggeridge et al.	Cycling	Light Cycling	General	20	0.272						
Fitbit	Charge 4	Muggeridge et al.	Cycling	Vigorous Cycling	General	20	0.183						
Fitbit	Charge 5	Muggeridge et al.	Running		General	20	0.879						
Fitbit	Charge 6	Muggeridge et al.	Running	Intense Cycling	General	20	0.39						
Fitbit	Charge 7	Muggeridge et al.	Running	Sprinting	General	20	0.348						
Oura	Gen2	Kristiansson et al.	Running		General	32		0.93					

Fitbit	Versa	Jagim et al.	Mixed	Walking, running	General	20	0.86	0.93					
Polar	Ignite	Jagim et al.	Mixed	Walking, running	General	20	0.83	0.54					
Polar	TeamPro Sensor	Jagim et al.	Mixed	Walking, running	General	20	0.95	0.85					
Apple	Apple Watch Series 6	Støve et al.	Strength Training	Barbell Exercises	General	29	0.839						
Apple	Apple Watch Series 6	Støve et al.	Strength Training	Barbell Exercises	General	29	0.832						
Apple	Apple Watch Series 6	Støve et al.	Strength Training	Machine Exercises	General	29	0.953						
Apple	Apple Watch Series 6	Støve et al.	Strength Training	Bodyweight Exercises	General	29	0.581						
Whoop	Band 3.0	Støve et al.	Strength Training	Barbell Exercises	General	29	0.763						
Whoop	Band 3.1	Støve et al.	Strength Training	Barbell Exercises	General	29	0.602						
Whoop	Band 3.3	Støve et al.	Strength Training	Bodyweight Exercises		29	0.4945						
Garmin	Forerunner 230	Snyder, Willoughby, & Smith	Running		Male	22			0.762				
Garmin	Forerunner 230	Snyder, Willoughby, & Smith	Running		Female	22			0.801				

Vital	Scout	Hopkins et al.	Cycling		General	12		0.776		0.744	0.777		
Apple	Watch 6	Hajj-Boutros et al.	Walking		General	60	0.95	0.75					
Apple	Watch 6	Hajj-Boutros et al.	Running		General	60	0.84	0.86					
Apple	Watch 6	Hajj-Boutros et al.	Strength Training		General	60	0.98	0.74					
Apple	Watch 6	Hajj-Boutros et al.	Cycling		General	60	0.93	0.72					
Polar	Vantage V	Hajj-Boutros et al.	Walking		General	60	0.75	0.69					
Polar	Vantage V	Hajj-Boutros et al.	Running		General	60	0.88	0.74					
Polar	Vantage V	Hajj-Boutros et al.	Strength Training		General	60	0.96	0.71					
Polar	Vantage V	Hajj-Boutros et al.	Cycling		General	60	0.78	0.6					
Fitbit	Sense	Hajj-Boutros et al.	Walking		General	60	0.86	0.72					
Fitbit	Sense	Hajj-Boutros et al.	Running		General	60	0.85	0.88					
Fitbit	Sense	Hajj-Boutros et al.	Strength Training		General	60	0.88	0.61					

Fitbit	Sense	Hajj-Boutros et al.	Cycling		General	60	0.89	0.69					
Hexoskin	Smart Shirt	Haddad et al.	Running		General	9	0.79						
greenTEG	CORE	Goods et al.	Hockey	Hockey Practice	Athletes	11						0.89	
greenTEG	CORE	Goods et al.	Hockey	Hockey Match	Athletes	8						0.81	
greenTEG	CORE	Goods et al.	Hockey	Hockey Match	Athletes	7						0.88	
greenTEG	CORE	Goods et al.	Hockey	Hockey Match	Athletes	6						0.84	
Apple	Apple Watch Series 4	Düking et al.	Mixed	Walking, running	General	25	0.97	0.71					
Apple	Apple Watch Series 4	Düking et al.	Mixed	Walking, running	General	25	0.97	0.71					
Apple	Apple Watch Series 4	Düking et al.	Mixed	Walking, running	General	25	0.99	0.8					
Apple	Apple Watch Series 4	Düking et al.	Mixed	Walking, running	General	25	0.94	0.95					
Apple	Apple Watch Series 4	Düking et al.	Mixed	Walking, running	General	25	0.85	0.93					
Apple	Apple Watch Series 4	Düking et al.	Mixed	Walking, running	General	25	0.92	0.66					
Apple	Apple Watch Series 4	Düking et al.	Mixed	Walking, running	General	25	0.95	0.75					

Polar	Vantage V	Düking et al.	Mixed	Walking, running	General	25	0.89	0.67					
Polar	Vantage V	Düking et al.	Mixed	Walking, running	General	25	0.91	0.49					
Polar	Vantage V	Düking et al.	Mixed	Walking, running	General	25	0.88	0.72					
Polar	Vantage V	Düking et al.	Mixed	Walking, running	General	25	0.86	0.95					
Polar	Vantage V	Düking et al.	Mixed	Walking, running	General	25	0.89	0.95					
Polar	Vantage V	Düking et al.	Mixed	Walking, running	General	25	0.75	0.85					
Polar	Vantage V	Düking et al.	Mixed	Walking, running	General	25	0.88	0.72					
Garmin	fēnix 5	Düking et al.	Mixed	Walking, running	General	25	0.85	0.2					
Garmin	fēnix 5	Düking et al.	Mixed	Walking, running	General	25	0.83	0.21					
Garmin	fēnix 5	Düking et al.	Mixed	Walking, running	General	25	0.63	0.57					
Garmin	fēnix 5	Düking et al.	Mixed	Walking, running	General	25	0.82	0.84					
Garmin	fēnix 5	Düking et al.	Mixed	Walking, running	General	25	0.82	0.91					
Garmin	fēnix 5	Düking et al.	Mixed	Walking, running	General	25	0.58	0.21					
Garmin	fēnix 5	Düking et al.	Mixed	Walking, running	General	25	0.77	0.45					
Fitbit	Versa	Düking et al.	Mixed	Walking, running	General	25	0.57	0.88					
Fitbit	Versa	Düking et al.	Mixed	Walking, running	General	25	0.54	0.78					

Fitbit	Versa	Düking et al.	Mixed	Walking, running	General	25	0.52	0.74					
Fitbit	Versa	Düking et al.	Mixed	Walking, running	General	25	0.82	0.76					
Fitbit	Versa	Düking et al.	Mixed	Walking, running	General	25	0.68	0.92					
Fitbit	Versa	Düking et al.	Mixed	Walking, running	General	25	0.53	0.42					
Fitbit	Versa	Düking et al.	Mixed	Walking, running	General	25	0.65	0.72					
Garmin	Forerunner 735XT	Damasceno et al.	Mixed	Light Exercise	General	28	0.87						
Garmin	Forerunner 735XT	Damasceno et al.	Mixed	Moderate Exercise	General	28	0.81						
Garmin	Forerunner 735XT	Damasceno et al.	Mixed	Vigorous Exercise	General	28	0.72						
Xiaomi	Mi Band 4	de la Casa Pérez et al.	Stairs		General	46	0.468						
Xiaomi	Mi Band 5	de la Casa Pérez et al.	Stairs		General	46	0.797						
Xiaomi	Mi Band 6	de la Casa Pérez et al.	Stairs		General	46	0.75						
Xiaomi	Mi Band 7	de la Casa Pérez et al.	Stairs		General	46	0.688						
Xiaomi	Mi Band 8	de la Casa	Stairs		General	46	0.715						

		Pérez et al.											
Garmin	fēnix 3 HR	Carrier et al.	Running		General	17			0.917				
BodyMedia	SenseWear Pro3 (Band Only)	Costello et al.	Rugby	Rugy Pre-Season	Athletes	6		0.61					
BodyMedia	SenseWear Pro3 (Band Only)	Costello et al.	Rugby	Rugy In-Season	Athletes	7		0.79					
BodyMedia	SenseWear Pro3 (Band + Optimeye)	Costello et al.	Rugby	Rugy Pre-Season	Athletes	6		0.68					
BodyMedia	SenseWear Pro3 (Band + Optimeye)	Costello et al.	Rugby	Rugy In-Season	Athletes	7		0.83					
Polar	Vantage V2	Cosoli, Antognoli, Veroli, & Scalise	Mixed	Walking, Running	General	10	0.83						
Polar	Vantage V2	Cosoli, Antognoli, Veroli, & Scalise	Swimming		General	10	0.2						
Garmin	Venu Sq	Cosoli, Antognoli, Veroli, & Scalise	Mixed	Walking, Running	General	10	0.95						
Garmin	Venu Sq	Cosoli, Antognoli, Veroli, & Scalise	Swimming		General	10	0.13						

Apple	Watch 6	Alfonso et al.	Walking		General	18	0.998						
Polar	Vantage M2	Alfonso et al.	Walking		General	18	0.652						
Fitbit	Charge 4	Nissen et al.	Walking	Slow Walking	General	23	0.188						
Fitbit	Charge 4	Nissen et al.	Walking	Fast Walking	General	23	0.408						
Fitbit	Charge 4	Nissen et al.	Stairs	Stairs	General	23	0.803						
Fitbit	Charge 4	Nissen et al.	Strength Training	Bodyweight Exercise	General	23	0.335						
Samsung	Galaxy Watch Active 2	Nissen et al.	Walking	Slow Walking	General	23	0.24						
Samsung	Galaxy Watch Active 2	Nissen et al.	Walking	Fast Walking	General	23	0.516						
Samsung	Galaxy Watch Active 2	Nissen et al.	Stairs	Stairs	General	23	0.812						
Samsung	Galaxy Watch Active 2	Nissen et al.	Strength Training	Bodyweight Exercise	General	23	0.668						

Reported Pearson correlation coefficients from each study reviewed.

Table A.3. Individual Condition Test Results for Garmin Watch

	fēnix Anterior Normoxia	Criterion Anterior Normoxia	fēnix Posterior Normoxia	Criterion Posterior Normoxia	fēnix Anterior Hypoxia	Criterion Anterior Hypoxia	fēnix Posterior Hypoxia	Criterion Posterior Hypoxia
Mean (%)	95.67%	97.10%	95.74%	97.05%	94.90%	88.00%	95.18%	85.90%
Standard Deviation	2.35%	1.74%	1.41%	2.24%	0.32%	8.29%	1.66%	9.55%
MAPE		1.80%		2.95%		6.93%		6.92%
Pearson Correlation		0.44		-0.47		0.00		0.29
Lin's Concordance		0.37		-0.33		0.00		0.10
Bland-Altman Bias		-1.08 (-2.53, 0.36)		-1.37 (-2.95, 0.22)		4.90 (0.56, 9.24)		4.36 (0.30, 8.43)
TOST Test (Upper)		0.003		< 0.001		0.82		0.76
TOST Test (Lower)		0.47		0.36		0.001		0.001

Validity statistics for each individual condition (normoxia/hypoxia, anterior/posterior) for the Garmin watch compared to the criterion.

Curriculum Vitae

Bryson Carrier

brysoncarrier@gmail.com

Education

University of Nevada, Las Vegas, Las Vegas NV

- Degree: Doctor of Philosophy, Interdisciplinary Health Sciences
    - Dissertation Title (3 Primary Works): Development and Usage of Analytic Tools and Recommendations for Validation and Reliability Studies Using Consumer-Grade Wearable Technology
        - The WEAR-BOT Checklist: A Risk of Bias Tool for Evaluating Validity and Reliability Research in Wearable Technology
        - The Risk of Bias in Validity and Reliability Studies Testing Physiological Variables in Consumer-Grade Wearable Technology Assessed via the WEAR-BOT: A Systematic Review and Meta Analysis
        - Validation of Aerobic Capacity Estimate (VO2max) and Pulse Oximetry in Wearable Technology
    - Concentrations: Exercise Physiology, Statistics
- Expected Graduation Date: May 2024
- Current GPA: 3.93

University of Nevada, Las Vegas, Las Vegas NV

- Degree: Master of Science, Kinesiology

- Thesis Title: Assessing the Validity and Reliability of Several Heart Rate Monitors in Wearable Technology While Mountain Biking
- Concentration: Exercise Physiology

- Graduation Date: May 15, 2021

  GPA: 4.0

Utah Valley University, Orem UT

- Degree: Bachelor of Science, Biology
  - Undergraduate Senior Thesis Title: Phylogenetic Analysis of the Baetidae Family
- Graduation Date: Aug. 11, 2017
- GPA: 3.75

Research Publications

Published Manuscripts

1. **Carrier, B.**; Salatto, R.W.; Davis, D.W.; Sertic, J.V.L.; Barrios, B.; McGinnis, G.R.; Girouard, T.J.; Burroughs, B.; Navalta, J.W. Assessing the Validity of Several Heart Rate Monitors in Wearable Technology While Mountain Biking. Int. J. Exerc. Sci. 2023, Vol. 16 : Iss. 7, Pages 1440 - 1450. Available at: https://digitalcommons.wku.edu/ijes/vol16/iss7/9
2. Navalta, J.W.; Davis, D.W.; **Carrier, B.**; Malek, E.M.; Vargas, N.; Rodriguez, J.P.; Weyers, B.; Carlos, K.; Peck, M. Validity and Reliability of Wearable Devices during Self-Paced Walking, Jogging, and Overground Skipping. Sport Mont 2023, 21, 23-29. doi: 10.26773/smj.231004
3. Navalta, J.W.; Davis, D.; Malek, E.; **Carrier, B.**; Bodell, N.; Manning, J.; Cowley, J.; Funk, M.; Lawrence, M.; DeBeliso, M. Effect of Heart Rate Data Processing on Reliability and Validity Decisions in Wearable Technology Devices. Sci. Rep. 2023, 13, 11736. doi: 10.1038/s41598-023-38329-w

4. Gardner, C.; Navalta, J.W.; **Carrier, B.**; Aguilar, C.; Rodriguez, J.P. Training Impulse and Its Impact on Load Management in Collegiate and Professional Soccer Players. Technologies 2023, 11, 79. doi: 10.3390/technologies11030079
5. **Carrier, B.**; Helm, M.M.; Cruz, K.; Barrios, B.; Navalta, J.W. Validation of Aerobic Capacity (VO2max) and Lactate Threshold in Wearable Technology for Athletic Populations. Technologies 2023, 11, 71. doi: 10.3390/technologies11030071
6. Davis, D.W.; **Carrier, B.**; Cruz, K.; Barrios, B.; Landers, M.R.; Navalta, J.W. A Systematic Review of the Effects of Meditative and Mindful Walking on Mental and Cardiovascular Health. Int. J. Exerc. Sci. 2022, 15, 1692-1734
7. **Carrier, B.**; Navalta, J.W. Data Analysis Processes and Techniques for Validation of Wearable Technology: An Example. Topics in Exerc. Sci. Kinesiol. 2022, 3(1), Article 10. Available at: https://digitalscholarship.unlv.edu/scholarship_kin/vol3/iss1/10
8. Navalta, J.W.; Davis, D.W.; **Carrier, B.**; Sertic, J.V.L.; Cater, P. Teaching Applied Exercise Physiology Using a Prototype Energy Expenditure Measurement Device. Interdiscip. J. Probl.-Based Learn. 2021, 15(2). doi: 10.14434/ijpbl.v15i2.31525
9. Pinedo-Jauregi, A.; Garcia-Tabar, I.; **Carrier, B.**; Navalta, J.W.; Cámara, J. Reliability and Validity of the Stryd Power Meter during Different Walking Conditions. Gait Posture 2021. doi: 10.1016/j.gaitpost.2021.11.041
10. Davis, D.; **Carrier, B.**; Cruz, K.; Barrios, B.; Navalta, J. A Protocol and Novel Tool to Systematically Review the Effects of Mindful Walking on Mental and Cardiovascular Health. PLOS ONE 2021. doi: 10.1371/journal.pone.0258424
11. Salatto, R.W.; McGinnis, G.R.; Davis, D.W.; **Carrier, B.**; Manning, J.W.; DeBeliso, M.; Navalta, J.W. Effects of Acute Beta-Alanine Ingestion and Immersion-Plus-Exercise on Connectedness to Nature and Perceived Pain. Int. J. Environ. Res. Public Health 2021, 18(15), 8134. doi: 10.3390/ijerph18158134

12. **Carrier, B.**; Creer, A.; Williams, L.R.; Holmes, T.M.; Jolley, B.D.; Dahl, S.; Weber, E.; Standifird, T. Validation of Garmin Fenix 3 HR Fitness Tracker Biomechanics and Metabolics (VO2max). J. Meas. Phys. Behav. 2020, 3(4), 331-337. doi: 10.1123/jmpb.2019-0066
13. **Carrier, B.**; Barrios, B.; Jolley, B.D.; Navalta, J.W. Validity and Reliability of Physiological Data in Applied Settings Measured by Wearable Technology: A Rapid Systematic Review. Technologies 2020, 8(4), 70. doi: 10.3390/technologies8040070
14. Barrios, B.; **Carrier, B.**; Jolley, B.D.; Davis, D.W.; Sertic, J.; Navalta, J.W. Establishing a Methodology for Conducting a Rapid Review on Wearable Technology Reliability and Validity in Applied Settings. Topics Exerc. Sci. Kinesiol. 2020, 1(2), Article 8. Available at: https://digitalscholarship.unlv.edu/scholarship_kin/vol1/iss2/8
15. Salatto, R.W.; Davis, D.W.; **Carrier, B.**; Barrios, B.; Sertic, J.; Cater, P.; Navalta, J.W. Efficient Method of Delivery for Powdered Supplement or Placebo for an Outdoor Exercise Investigation. Scholarship Kinesiology 2020, 1(2), Article 5. Available at: https://digitalscholarship.unlv.edu/scholarship_kin/vol1/iss2/5

Research Funding

<u>Awarded</u>

- University of Nevada, Las Vegas
    - Department of Kinesiology and Nutrition Sciences, Graduate Student Travel Grant, Oct. 2023, $452
    - Graduate & Professional Student Association Travel Grant, May 2022, $1,250
    - Department of Kinesiology and Nutrition Sciences, Graduate Student Research Award, Jan. 2022, $1,000

- Kinesiology Department - Mentor Directed Research Award, Dec. 2019, $1,500
- Graduate & Professional Student Association Research Grant, Nov. 2019, $1,040
- UNLV Access Grant, Aug. 2019, $1,000

- Utah Valley University, Orem UT
  - Scholarly Activities Committee Grant, College of Science, Apr. 2019, $4,684
  - Scholarly Activities Committee Grant, College of Science, Jan. 2019, $2,000
  - Scholarly Activities Committee Grant, College of Science, Nov. 2018, $4,000

Applied For (not awarded)

- University of Nevada, Las Vegas
  - Department of Kinesiology and Nutrition Sciences, Graduate Student Research Award, Dec. 2020, $1,500
- National Strength and Conditioning Association Foundation Grant
  - Graduate Student Research Grant, Feb. 2020, $15,000
- National Science Foundation (NSF)
  - NSF Graduate Research Fellowship Program (GRFP), Oct. 2019, $138,000
- Utah Valley University, Orem UT
  - Undergraduate Research Scholarly and Creative Activities Grant, Feb. 2017, $3,000
  - Undergraduate Research Scholarly and Creative Activities Grant, Oct. 2016, $3,000

Courses Taught

University of Nevada, Las Vegas – Aug. 2019 to Present

- Courses Taught:
  - KIN 391 - Exercise physiology (lecture)

- KIN 391 - Exercise physiology (lab)
- PEX 124 - Indoor soccer
- PEX 154 - Indoor cycling

College of Southern Nevada - Jan. 2022 to May 2022

- Courses Taught:
  - BIO 223 - Anatomy and Physiology (lab)

Utah Valley University – Aug. 2018 to Jul. 2019

- Courses Taught:
  - BIOL 1015 – Biology lab for non-science majors
  - BIOL 1615 – Biology lab for life science majors

Course / Teaching Student Evaluation Scores

	3-Year Average	5-Year Average
Overall Mean	4.28	4.33
Lecture Mean Score	4.28	4.34
Lab Mean Score	4.27	4.32

*Measurements all out of 5 points total.

Relevant Experience

University of Nevada, Las Vegas – Multiple Dates and Roles

- **Graduate Student (PhD)** - Aug. 2021 to Present

    - PhD student, studying interdisciplinary health sciences with an emphasis in exercise physiology.
- **Graduate Student (Master's)** - Aug. 2019 to May 2021
    - Master's student, studying Kinesiology with an emphasis in exercise physiology.
- **Student Researcher** - Aug. 2019 to Present
    - Researcher in Exercise Physiology lab, studying wearable technology, sport performance, sport nutrition, environmental physiology, and others under Dr. James Navalta.
    - Public Health & Epidemiology Researcher, studying health impacts of vector-borne diseases under Dr. Chad Cross.
- **Instructor / Graduate Assistant** - Aug. 2019 to Present
    - Instructor teaching exercise physiology labs and lecture (for graduate assistantship), as well as indoor soccer and spin courses (as adjunct / part-time instructor).
- **Sport Science Intern** - Aug - Dec. 2022
    - Worked with the sports science team to manage athlete technology, including Catapult, Hawkins force plates, Firstbeat, and others. Responsible for managing data and performing analyses. Worked with several teams, most closely with the women's soccer team.

College of Southern Nevada - Jan 2022 to May 2022

Adjunct instructor teaching Anatomy and Physiology lab (BIO 223).

Utah Valley University - Multiple Dates and Roles

- **Undergraduate Student** - Aug. 2013 - May 2017
- **Adjunct Faculty Member** - Aug. 2018 to Jul. 2019

    - Adjunct faculty member in the Biology department. Taught introductory biology labs for life science and non-science majors.
- **Research Assistant (Human Performance Lab)** - Jan. 2018 to Jul. 2019
    - Research assistant in Biomechanics and Human Performance Lab under Dr. Andrew Creer and Dr. Tyler Standifird. Tasks included collecting kinematic, EMG, metabolic, physiologic, and other types of data for multiple research studies.
- **Research Assistant (Brooks Lab)** - Nov. 2018 to Jul. 2019
    - Research assistant in Brooks Microbiology Lab. Researched microbial community of indoor rock-climbing walls. Tasks included PCR, analyzing sequenced genetic data, DNA extraction, electrophoresis, PCR cleaning, applying source tracking techniques to samples and culturing bacteria.
- **Research Assistant (Ogden Lab)**- Jun. 2015 to Jun. 2017
    - Research assistant in Ogden Bioinformatics Lab. Researched evolution of Mayflies. Tasks included PCR, analyzing sequenced genetic data, DNA extraction, electrophoresis, PCR cleaning, and constructing phylogenetic trees.
- **Teacher's Assistant** - Jan. 2016 to May 2016
    - Teacher's Assistant for introductory biology class. Tasks included multiple weekly review sessions, one-on-one tutoring, grading, and weekly coordination with professors.

Certifications

- ASA-Graduate Statistician (GSTAT)
- UNLV Grad Academy, Graduate Research Certification
- American Red Cross Adult and Pediatric First Aid/CPR/AED Certification
- Indoor Cycling Instructor Certification

- Collaborative Institutional Training Initiative (CITI) Certification

Skills

Analytical Skills

- Biostatistics and Epidemiology
- Business Intelligence Software
- Data Analysis and Data Cleaning
- Google Docs Suite
- IBM SPSS Statistics Software
- Jamovi Statistics Software
- Microsoft Office Suite
- R (programming language)
- SQL (querying language)
- Tableau and Tableau Prep

Other Scientific Skills

- Biodex Isokinetic Testing
- Blood Glucose Testing
- Bod Pod Body Composition Testing
- Catapult Sports Wearable Technology
- COSMED K5 Metabolic System
- DNA Extraction and Purification
- Functional Threshold Power (FTP) Testing
- Gel Electrophoresis
- Graded Treadmill and Cycle Ergometer Exercise Testing (VO2max)
- Hawkin Dynamics Force Plate Testing
- Hematocrit and Hemoglobin Testing
- LabQuest
- Lactate Threshold Testing
- ParvoMedics Metabolic System
- PCR Gene Amplification
- Phlebotomy
- RNA Extraction and Purification
- Skinfold Body Composition Testing
- Submaximal & Field-Based Aerobic Capacity Testing
- Wingate Testing

Honors & Awards

Professional Organizations

- Graduate Student Research Competition Finalist, Southwest Chapter of the American College of Sports Medicine Annual Meeting, Oct., 2023.

University of Nevada, Las Vegas

- Western Association of Graduate Schools Nominee
    - Department Nominee (Kinesiology and Nutrition Sciences) for Western Association of Graduate Schools (WAGS) Outstanding Master's Thesis Award for the "Distinguished Master's Thesis Award STEM" Category, Jul. 2021
    - College Nominee (School of Integrated Health Sciences) for Western Association of Graduate Schools (WAGS) Outstanding Master's Thesis Award for the "Distinguished Master's Thesis Award STEM" Category, Aug. 2021
    - University Nominee (UNLV) for Western Association of Graduate Schools (WAGS) Outstanding Master's Thesis Award for the "Distinguished Master's Thesis Award STEM" Category, Nov. 2021
- 2nd Place Winner for Poster Presentation Group, GPSA Research Forum, Apr. 2021

Utah Valley University

- Graduated with Distinction (Cum Laude), Aug. 2017
- Dean's List 2014 to 2017

Service & Volunteer Work

- Peer reviewer for scientific journals:
    - ACSM's Health & Fitness Journal - Jan. 2024 to Present

    - Topics in Exercise Science and Kinesiology - Oct. 2023 to Present
    - Sports Medicine - Jul. 2023 to Present
    - Frontiers in Digital Health - Mar. 2022 to Present
    - International Journal of Exercise Science - Dec. 2019 to Present
- Student Leadership Committee Member, UNLV Sport Innovation Institute - Aug. 2023 - Present
- UNLV Diversity, Equity, Inclusion, & Justice Advisory Board Member, School of Integrated Health Sciences Student Representative - Aug. 2022 - Present
- Las Vegas Golden Knights performance and physiological testing technician - Aug. 2022, Jul. 2023
- Las Vegas Lights FC performance and physiological testing technician - Jan. 2020

Scientific Presentations

Oral Presentations

1. **Carrier, B.**; Bunn, J.; Reece, J.D.; Aguilar, C.D.; Eschbach, C.; Navalta, J.W. "The Risk of Bias in Validity and Reliability Studies Testing Physiological Variables using Consumer-Grade Wearable Technology: A Systematic Review and WEAR-BOT Analysis," Int. J. Exerc. Sci.: Conf. Proc. 2023, 14(3), Article 6. Available at: https://digitalcommons.wku.edu/ijesab/vol14/iss3/6
    - Presented as a finalist in the Graduate Student Research Competition.
2. **Carrier, B.**; Helm, M.M.; Davis, D.W.; Cruz, K.; Barrios, B.; Navalta, J.W., FACSM. "Validation of VO2max and Lactate Threshold Estimates in Wearable Technology in High-Level Runners," American College of Sports Medicine, May 2022, San Diego, CA, USA.
    - Presented as a thematic poster presentation.

3. **Carrier, B.**; Navalta, J.W. "Understanding Heart Rate Monitor Technology, Validity, and Appropriate Use-Cases in Wearable Technology During Exercise," University of Nevada, Las Vegas Annual Graduate and Professional Student Research Forum, Apr. 2022, Las Vegas, NV, USA
    - Presented as a pre-recorded video presentation.
4. Symposium: Wearable Activity Monitors. Introduction of student presenters, Navalta, J.W.; The evolution of wearable devices, Salatto, R.W.; The current state of technology devices in applied settings, Barrios, B.; The needed considerations in current testing models, Jolley, B.D.; The future of wearable exercise testing, **Carrier, B.** Virtual Annual Meeting of the Southwest American College of Sports Medicine, 2020
    - Presented as a pre-recorded video presentation.
5. **Carrier, B.**; Salatto, R.W.; Manning, J.W.; Barrios, B.; Sertic, J.V.L.; Davis, D.W.; Cater, P.C.; McGinnis, G.; DeBeliso, M.; Navalta, J.W. "Does Acute Beta-Alanine Supplementation Improve Performance, Rating of Perceived Exertion and Heart Rate During Hiking?" American College of Sports Medicine, May 2020, San Francisco, CA, USA
    - Presented as a thematic poster presentation.

<u>Poster Presentations</u>

1. **Carrier, B.**; Salatto, R.W.; Davis, D.W.; Sertic, J.V.L.; Barrios, B.; Cater, P.; Navalta, J.W., "Assessing the Validity of Several Heart Rate Monitors in Wearable Technology While Mountain Biking," Southwest Am. Coll. Sports Med., Oct. 2021, Costa Mesa, CA, USA
2. **Carrier, B.**; Cruz, K.; Farmer, H.; Navalta, J., "Validation of the Lactate Threshold Estimate from the Garmin fenix 6 Fitness Tracker," Am. Coll. Sports Med., Jun. 2021, Washington D.C., USA
   Presented as a digital poster.

3. **Carrier, B.**; Cruz, K.; Farmer, H.; Navalta, J., "Validation of the Lactate Threshold Estimate from the Garmin fenix 6 Fitness Tracker," Univ. Nevada, Las Vegas Ann. Grad. Prof. Stud. Res. Forum, Apr. 2021, Las Vegas, NV, USA
4. **Carrier, B.**; Trainor, T.; Jolley, B.W.; Navalta, J.W.; Creer, A., "Validation of the Humon Hex Lactate Threshold Estimate," Southwest Am. Coll. Sports Med., Oct. 2019, Costa Mesa, CA, USA
5. **Carrier, B.**; Holmes, T.; Williams, L.; Dahl, S.; Weber, L.; Creer, A.; Standifird, T., "Validation of Garmin Fitness Tracker Biomechanics," Am. Coll. Sports Med., May 2019, Orlando, FL, USA
6. **Carrier, B.**; Richards, S.; Hancock, C.; Brooks, L., "Who Brought the Microbes? Investigating the source of fecal veneer on rock climbing holds," Intermountain Am. Soc. Microbiol., Apr. 2019, Provo, UT, USA
7. **Carrier, B.**; Holmes, T.; Williams, L.; Dahl, S.; Weber, L.; Creer, A.; Standifird, T., "Validation of Garmin Fitness Tracker Biomechanics," Southwest Am. Coll. Sports Med., Oct. 2018, Costa Mesa, CA, USA
8. **Carrier, B.**; Ferguson, D.; Ogden T.H., "Molecular Phylogeny of Baetidae (Ephemeroptera)," Evolution 2017, June 2017, Portland, OR, USA
9. **Carrier, B.**; Ferguson, D.; Ogden T.H., "Molecular Phylogeny of Baetidae (Ephemeroptera)," Utah Conf. Undergrad. Res., Feb. 2017, Orem, UT, USA
10. **Carrier, B.**; Ogden T.H., "Phylogenetic Relationships of Mayfly Family Baetidae (Ephemeroptera)," Utah Conf. Undergrad. Res., Feb. 2016, Salt Lake City, UT, USA

Abstracts

Submitted Abstracts

1. Cross, C.L.; **Carrier, B.**; Alcala, M.; "Soil-transmitted helminths in the United States: Using Big Data to characterize patients and analyze disease trends". Submitted to *American Society of Parasitologists National Meeting*, Mar. 2024

Published Peer Reviewed Abstracts

1. **Carrier, B.**; Bunn, J.; Reece, J.D.; Aguilar, C.D.; Eschbach, C.; Navalta, J.W.; "The Risk of Bias in Validity and Reliability Studies Testing Physiological Variables using Consumer-Grade Wearable Technology: A Systematic Review and WEAR-BOT Analysis", *Int. J. Exerc. Sci.: Conf. Proc. 2023, 14(3), Article 6*. Available at: https://digitalcommons.wku.edu/ijesab/vol14/iss3/6
2. Carballo, T.; Blank, M.; **Carrier, B.**; Cruz, S.; Bovell, J.; Davis, D.; Sweder, T.; Malek, E.; Zarei, S.; Navalta, J.; "Does hand use affect metabolic measures during pickleball". *Int. J. Exercise Sci. Conf. Proc. 2023, 14(1), 123*. Available at: https://digitalcommons.wku.edu/ijesab/vol14/iss3/123
3. Blank, M.; Davis, D.; **Carrier, B.**; Carballo, T.; Bovell, J.; Sweder, T.; Cruz, S.; Yu, Z.; Zarei, S.; Navalta, J.; "Evaluation of caloric expenditure metrics of Garmin Instinct wearable technology devices during pickleball:. *Int. J. Exercise Sci. Conf. Proc. 2023, 14(3), 118*. Available at: https://digitalcommons.wku.edu/ijesab/vol14/iss3/118
4. Zarei, S.; Cruz, S.; Carballo, T.; **Carrier, B.**; Davis, D.; Bovell, J.; Sweder, T.; Blank, M.; Malek, E.; Navalta, J.; "Validity and reliability of the Garmin Instinct in measuring heart rate during pickleball". *Int. J. Exercise Sci. Conf. Proc. 2023, 14(3), 134*. Available at: https://digitalcommons.wku.edu/ijesab/vol14/iss3/134
5. Cruz, S.; Carballo, T.; Davis, D.; **Carrier, B.**; Bovell, J.; Blank, M.; Sweder, T.; Yu, Z.; Zarei, S.; Navalta, J.; "Does handedness impact pulmonary measures during pickleball?" *Int. J. Exercise Sci. Conf. Proc. 2023, 14(3), 28*.

6. Bovell, J.; Davis, D.; **Carrier, B.**; Zarei, S.; Carballo, T.; Blank, M.; Sweder, T.; Cruz, S.; Malek, E.; Navalta, J.; "Validity and reliability of the Polar OH1 biceps-band heart rate monitor during pickleball". *Int. J. Exercise Sci. Conf. Proc. 2023, 14(3), 131*. Available at: https://digitalcommons.wku.edu/ijesab/vol14/iss3/131
7. Ziegler, K. K.; McKenzie, A.; Ziegler, W.; Maxwell, S.; **Carrier, B.**; Aguilar, C.; Routsis, A.; Thornton, T.; Bovell, J.; Zarei, S.; Green, D.; Lavin, K. L. A.; Hawkes, A.; Cowley, J.; Funk, M.; Navalta, J. W. FACSM; Lawrence, M. M. Perceived Fatigue and Physical Activity Enjoyment Following Indoor and Outdoor Moderately Heavy Superset Resistance Training. *Int. J. Exercise Sci. Conf. Proc. 2023, 14, 144*. Available at: https://digitalcommons.wku.edu/ijesab/vol14/iss3/144.
8. Ziegler, W. F.; **Carrier, B.**; Aguilar, C. D.; Pearce, D.; Graffius, J. M.; Ellingford, B.; Fullmer, W.; Cowley, J.; Funk, M.; Bodell, N.; Navalta, J. W. FACSM; Lawrence, M. M.; "Repetition Count Concurrent Validity of Various Garmin Wrist Watches During Light Circuit Resistance Training". *Int. J. Exercise Sci. Conf. Proc. 2023, 14, 143*. Available at: https://digitalcommons.wku.edu/ijesab/vol14/iss3/143
9. Maxwell, S. M.; **Carrier, B.**; Aguilar, C.; Ziegler, K.; McKenzie, A.; Ziegler, W.; Routsis, A.; Thornton, T.; Zarei, S.; Green, D.; Bovell, J.; Lavin, K. A.; Hawkes, A.; Cowley, J. C.; Funk, M.; Navalta, J. W. FACSM; Lawrence, M. M.; "Rating of Perceived Exertion, Average Heart Rate, and Energy Expenditure Following Indoor and Outdoor Moderately Heavy Superset Resistance Training". *Int. J. Exercise Sci. Conf. Proc. 2023, 14, 139*. Available at: https://digitalcommons.wku.edu/ijesab/vol14/iss3/139
10. Thorton, T.; Ziegler, W.; Maxwell, S.; McKenzie, A.; Routsis, A.; Ziegler, K.; **Carrier, B.**; Aguilar, C.D.; Green, D.; Bovell, J.; Lavin, K.L.A.; Zarei, S.; Cowley, J.; Hawkes, A.; Funk, M.; Navalta, J.W.; Lawrence, M.M. "Heart Rate and Energy Expenditure Concurrent Validity of Identical Garmin

Wrist Watches During Moderately Heavy Resistance Training" Southwest Am. Coll. Sports Med., Oct. 2023, Costa Mesa, CA, USA

11. Graffius, J.; Pearce, D.; Ellingford, B.; **Carrier, B.**; Aguilar, C.; Fullmer, W.; Ziegler, W.; Gil, D.; Torres, M.; Davis, D.; Peck, M.; Vargas, N.; Weyers, B.; Carlos, K.; Bodell, N.; Manning, J.; Funk, M.; DeBeliso, M.; Navalta, J.; Lawrence, M. "Garmin wrist watches heart rate and energy expenditure validity during light circuit resistance training" Med. Sci. Sports Exercise 2023, 55(9S), 1185. https://www.doi.org/10.1249/01.mss.0000983372.14302.26
12. Pearce, D.; Graffius, J.; Ellingford, B.; **Carrier, B.**; Aguilar, C.; Gil, D.; Torres, M.; Davis, D.; Ziegler, W.; Fullmer, W.; Peck, M.; Vargas, N.; Weyers, B.; Carlos, K.; Bodell, N.; Manning, J.; Cowley, J.; Funk, M.; Navalta, J.; Lawrence, M. "Heart rate and energy expenditure validity and reliability in Garmin Instinct watches during resistance training," Med. Sci. Sports Exercise 2023, 55(9S), 1186. https://www.doi.org/10.1249/01.mss.0000983376.69328.37
13. Torres, M.; Pearce, D.; Graffius, J.; Ellingford, B.; Gil, D.; **Carrier, B.**; Aguilar, C.; Ziegler, W.; Fullmer, W.; Weyers, B.; Peck, M.; Vargas, N.; Carlos, K.; Davis, D.; Funk, M.; Lawrence, M.; Manning, J.; Navalta, J.; Bodell, N. "Outdoor resistance training decreases Rate of Perceived Exertion during light-intensity resistance training," Med. Sci. Sports Exercise 2023, 55(9S), 1820. https://www.doi.org/10.1249/01.mss.0000985576.76514.c9
14. Pearce, D.; Graffius, J.M.; Ellingford, B.; **Carrier, B.**; Aguilar, C.D.; Gil, D.; Torence, M.; Davis, D.W.; Ziegler, W.; Fullmer, W.; Peck, M.; Vargas, N.R.; Weyers, B.; Carlos, K.; Bodell, N.; Manning, J.W.; Navalta, J.W.; DeBeliso, M.; Funk, M.; Lawrence, M.M. "Concurrent Validity and Reliability of Average Heart Rate and Energy Expenditure of Identical Garmin Instinct Watches During Low Intensity Resistance Training," Southwest Am. Coll. Sports Med., Oct. 2022, Costa Mesa, CA, USA
15. Graffius, J.M.; Pearce, D.; Ellingford, B.; **Carrier, B.**; Aguilar, C.D.; Fullmer, W.; Ziegler, W.; Gil, D.; Torence, M.; Davis, D.W.; Peck, M.; Vargas, N.R.; Weyers, B.; Carlos, K.; Bodell, N.; Manning, J.W.;

DeBeliso, M.; Navalta, J.W.; Funk, M.; Lawrence, M.M. "Average Heart Rate and Energy Expenditure Validity of Garmin Vivoactive 3 and Fenix 6 Wrist Watches During Light Circuit Resistance Training," Southwest Am. Coll. Sports Med., Oct. 2022, Costa Mesa, CA, USA

16. Carlos, K.; Davis, D.W.; **Carrier, B.**; Perdomo Rodriguez, J.; Vargas, N.R.; Malek, E.M.; Weyers, B.; Navalta, J.W. "The Validity of Bicep Located Heart Rate Monitors During Running," Southwest Am. Coll. Sports Med., Oct. 2022, Costa Mesa, CA, USA
17. Perdomo Rodriguez, J.; Davis, D.W.; Vargas, N.R.; Malek, E.M.; **Carrier, B.**; Carlos, K.; Weyers, B.; Navalta, J.W. "Comparing Exercise Intensity as a Percentage of the Age-Estimated Heart Rate Max Among Walking, Jogging, and Skipping," Southwest Am. Coll. Sports Med., Oct. 2022, Costa Mesa, CA, USA
18. Weyers, B.; Davis, D.W.; Vargas, N.R.; Perdomo Rodriguez, J.; Malek, E.M.; Carlos, K.; **Carrier, B.**; Navalta, J.W. "Determining Validity and Reliability of Caloric Expenditure Recorded by Wearable Technology While Walking and Running," Southwest Am. Coll. Sports Med., Oct. 2022, Costa Mesa, CA, USA
19. Vargas, N.R.; **Carrier, B.**; Davis, D.W.; Rodriguez, J.P.; Malek, E.M.; Weyers, B.; Carlos, K.; Navalta, J.W. "The Validity and Reliability of the Garmin Instinct in Measuring Heart Rate, Energy Expenditure, and Steps During Skipping," Southwest Am. Coll. Sports Med., Oct. 2022, Costa Mesa, CA, USA
20. **Carrier, B.**; Helm, M.M.; Davis, D.W.; Cruz, K.; Barrios, B.; Navalta, J.W. "Validation of VO2max and Lactate Threshold Estimates in Wearable Technology in High-Level Runners," Am. Coll. Sports Med., May 2022, San Diego, CA, USA
21. Cruz, K.; Davis, D.W.; **Carrier, B.**; Navalta, J.W. "Validity of the K5 Wearable Metabolic System during the YMCA Bench Press Test - A Pilot Study," Am. Coll. Sports Med., May 2022, San Diego, CA, USA

22. Fullmer, W.; **Carrier, B.**; Gil, D.; Cruz, K.; Aguilar, C.; Davis, D.; Malek, E.; Bodell, N.; Montes, J.; Manning, J.; DeBeliso, M.; Navalta, J.; Lawrence, M. "Validity of Average Heart Rate and Energy Expenditure in Polar Armband Devices While Self-Paced Biking," Utah Conf. Undergrad. Res. (UCUR), Feb. 2022, St. George, UT, USA
23. Helm, M.M.; **Carrier, B.**; Davis, D.W.; Cruz, K.; Barrios, B.; Navalta, J.W. "Validation of the Garmin Fenix 6S Maximal Oxygen Consumption (VO2max) Estimate," Int. J. Exercise Sci.: Conf. Proc. 2021, 14(1), Article 29. Available at: https://digitalcommons.wku.edu/ijesab/vol14/iss1/29
24. Helm, M.M.; **Carrier, B.**; Davis, D.W.; Cruz, K.; Barrios, B.; Navalta, J.W. "Validation of Sweat Rate, Fluid Loss, and Sodium Loss in Wearable Technology," Int. J. Exercise Sci.: Conf. Proc. 2021, 14(1), Article 8. Available at: https://digitalcommons.wku.edu/ijesab/vol14/iss1/8
25. **Carrier, B.**; Salatto, R.W.; Davis, D.W.; Sertic, J.V.L.; Barrios, B.; Cater, P.; Navalta, J.W. "Assessing the Validity of Several Heart Rate Monitors in Wearable Technology While Mountain Biking," Int. J. Exercise Sci.: Conf. Proc. 2021, 14(1), Article 18. Available at: https://digitalcommons.wku.edu/ijesab/vol14/iss1/18
26. Helm, M.M.; **Carrier, B.**; Davis, D.W.; Cruz, K.; Barrios, B.; Navalta, J.W. "Validation of the Garmin Fenix 6S Maximal Oxygen Consumption (VO2max) Estimate," Int. J. Exercise Sci.: Conf. Proc. 2021, 14(1), Article 29. Available at: https://digitalcommons.wku.edu/ijesab/vol14/iss1/29
27. Davis, D.W.; **Carrier, B.**; Cruz, K.; Barrios, B.; Navalta, J.W. FACSM "The Effects of Meditative and Mindful Walking on Mental and Cardiovascular Health," Int. J. Exercise Sci.: Conf. Proc. 2021, 14(1), Article 8. Available at: https://digitalcommons.wku.edu/ijesab/vol14/iss1/8
28. Bodell, N.; **Carrier, B.**; Gil, D.; Fullmer, W.; Cruz, K.; Aguilar, C.D.; Davis, D.W.; Malek, E.M.; Montes, J.; Manning, J.W.; Navalta, J.W.; Lawrence, M.M.; DeBeliso, M. "Validity of Average Heart Rate and Energy Expenditure in Polar OH1 and Verity Sense While Self-Paced Walking," Int. J.

Exercise Sci.: Conf. Proc. 2021, 14(1), Article 69. Available at: https://digitalcommons.wku.edu/ijesab/vol14/iss1/69

29. Gil, D.; **Carrier, B.**; Fullmer, W.; Cruz, K.; Aguilar, C.D.; Davis, D.W.; Malek, E.M.; Bodell, N.; Montes, J.; Manning, J.W.; Navalta, J.W.; Lawrence, M.M.; DeBeliso, M. "Validity of Average Heart Rate and Energy Expenditure in Polar OH1 and Verity Sense While Self-Paced Running," Int. J. Exercise Sci.: Conf. Proc. 2021, 14(1), Article 27. Available at: https://digitalcommons.wku.edu/ijesab/vol14/iss1/27
30. Fullmer, W.B.; **Carrier, B.**; Malek, E.; Gil, D.; Cruz, K.; Aguilar, C.; Davis, D.; Bodell, N.; Montes, J.; Manning, J.; DeBeliso, M.; Navalta, J.; Lawrence, M.M. "Validity of Average Heart Rate and Energy Expenditure in Polar Armband Devices While Self-Paced Biking," Int. J. Exercise Sci.: Conf. Proc. 2021, 14(1), Article 26. Available at: https://digitalcommons.wku.edu/ijesab/vol14/iss1/26
31. Cruz, K.; Navalta, J.W.; Davis, D.W.; **Carrier, B.** "Validity of the K5 Wearable Metabolic System during the YMCA Bench Press Test - A Pilot Study," Int. J. Exercise Sci.: Conf. Proc. 2021, 14(1), Article 22. Available at: https://digitalcommons.wku.edu/ijesab/vol14/iss1/22
32. **Carrier, B.**; Cruz, K.; Farmer, H.; Navalta, J. "Validation of the Lactate Threshold Estimate from the Garmin fenix 6 Fitness Tracker," Am. Coll. Sports Med., Jun. 2021, Washington D.C., USA
33. Cruz, K.; **Carrier, B.**; Farmer, H.; Navalta, J. "The Validity of VO2 Max: Treadmill GXT and Wearable Technology," Am. Coll. Sports Med., Jun. 2021, Washington D.C., USA
34. **Carrier, B.**; Cruz, K.; Farmer, H.; Navalta, J. "Validation of the Lactate Threshold Estimate from the Garmin fenix 6 Fitness Tracker," Univ. Nevada, Las Vegas Ann. Grad. Prof. Stud. Res. Forum, Apr. 2021, Las Vegas, NV, USA
35. Cruz, K.; **Carrier, B.**; Farmer, H.; Navalta, J. "The Validity of VO2 Max: Treadmill GXT and Wearable Technology," Univ. Nevada, Las Vegas Ann. Grad. Prof. Stud. Res. Forum, Apr. 2021, Las Vegas, NV, USA

36. Cruz, K.; Salatto, R.W.; Davis, D.W.; **Carrier, B.**; Barrios, B.; Cater, P.; Farmer, H.; Navalta, J.W. "Evaluation of Rating of Perceived Exertion During Mountain Biking," Southwest Am. Coll. Sports Med., Oct. 2020, Costa Mesa, CA, USA
37. Farmer, H.; Salatto, R.W.; Davis, D.W.; **Carrier, B.**; Barrios, B.; Cater, P.; Cruz, K.; Navalta, J., FACSM. "Felt Arousal Scale is Not Reliable for Use in Repeated Mountain Biking Trial Application," Southwest Am. Coll. Sports Med., Oct. 2020, Costa Mesa, CA, USA
38. **Carrier, B.**; Salatto, R.W.; Manning, J.W.; Barrios, B.; Sertic, J.V.L.; Davis, D.W.; Cater, P.C.; McGinnis, G.; DeBeliso, M.; Navalta, J.W. "Does Acute Beta-Alanine Supplementation Improve Performance, Rating of Perceived Exertion and Heart Rate During Hiking?", Am. Coll. Sports Med., May 2020, San Francisco, CA, USA
39. Barrios, B.; **Carrier, B.**; Cater, P.C.; Sertic, J.V.L.; Salatto, R.W.; Navalta, J.W. "Validation of Heart Rate Monitoring of Fenix 5 During Mountain Biking," Am. Coll. Sports Med., May 2020, San Francisco, CA, USA
40. Sertic, J.V.L.; **Carrier, B.**; Cater, P.C.; Barrios, B.; Salatto, R.W.; Navalta, J.W. "Validation of Two Wearable Chest Straps for Heart Rate Monitoring During Mountain Biking," Am. Coll. Sports Med., May 2020, San Francisco, CA, USA
41. Salatto, R.W.; Navalta, J.W.; Montes, J.; Bodell, N.; **Carrier, B.**; Sertic, J.V.L; Barrios, B.; Cater, P.C.; Davis, D.W.; Manning, J.W.; DeBeliso, M. "Evaluating the Validity of Heart Rate Measured by the Suunto Spartan Sport Watch During Trail Running," Am. Coll. Sports Med., May 2020, San Francisco, CA, USA
42. Navalta, J.W.; McGinnis, G.R.; Manning, J.W.; Salatto, R.W.; **Carrier, B.**; Davis, D.W.; Sertic, J.V.L.; Cater, P.C.; Barrios, B.; Malek, E.M.; Reynolds, C.K.; DeBeliso, M. "Acute Beta-Alanine Supplementation and Pain Perception Before and After Hiking," Am. Coll. Sports Med., May 2020, San Francisco, CA, USA

43. Trainor, T.; **Carrier, B.**; Jolley, B.W.; Creer, A. "Validation of the Humon Hex Lactate Threshold Estimate," Am. Coll. Sports Med., May 2020, San Francisco, CA, USA
44. Standifird, T.; Williams, L.; **Carrier, B.**; Creer, A. "Differences Between Predicted And Measured V02 During Level And Uphill Walking," Am. Coll. Sports Med., May 2020, San Francisco, CA, USA
45. **Carrier, B.**; Trainor, T.; Jolley, B.W.; Navalta, J.W.; Creer, A. "Validation of the Humon Hex Lactate Threshold Estimate," Southwest Am. Coll. Sports Med., Oct. 2019, Costa Mesa, CA, USA
46. Barrios, B.; Sertic, J.V.L.; Cater, P.C.; Davis, D.W.; **Carrier, B.**; Salatto, R.W.; Montes, J.; Bodell, N.; Manning, J.W.; DeBeliso, M.; Navalta, J.W. "Evaluating the Validity of Heart Rate Measured by the Jabra Elite During Trail Running," Southwest Am. Coll. Sports Med., Oct. 2019, Costa Mesa, CA, USA
47. Cater, P.C.; Sertic, J.V.L.; Davis, D.W.; Barrios, B.; **Carrier, B.**; Salatto, R.W.; Montes, J.; Bodell, N.; Manning, J.W.; DeBeliso, M.; Navalta, J.W. "Evaluating the Validity of Heart Rate Measured by the Rhythm During Trail Running," Southwest Am. Coll. Sports Med., Oct. 2019, Costa Mesa, CA, USA
48. Davis, D.W.; Barrios, B.; **Carrier, B.**; Salatto, R.W.; Sertic, J.V.L.; Cater, P.C.; Montes, J.; Bodell, N.; Manning, J.W.; DeBeliso, M.; Navalta, J.W. "Evaluating the Validity of Heart Rate Measured by the Garmin Fenix 5 During Trail Running," Southwest Am. Coll. Sports Med., Oct. 2019, Costa Mesa, CA, USA
49. Salatto, R.W.; Navalta, J.W.; Montes, J.; Bodell, N.; **Carrier, B.**; Sertic, J.V.L; Barrios, B.; Cater, P.C.; Davis, D.W.; Manning, J.W.; DeBeliso, M. "Evaluating the Validity of Heart Rate Measured by the Suunto Spartan Sport Watch During Trail Running," Southwest Am. Coll. Sports Med., Oct. 2019, Costa Mesa, CA, USA
50. Sertic, J.V.L.; Cater, P.C.; Davis, D.W.; Barrios, B.; **Carrier, B.**; Salatto, R.W.; Montes, J.; Bodell, N.; Manning, J.W.; DeBeliso, M.; Navalta, J.W. "Validating the Heart Rate Feature of the Motiv Ring on Outside Graded Terrain," Southwest Am. Coll. Sports Med., Oct. 2019, Costa Mesa, CA, USA

51. Navalta, J.W.; Salatto, R.W.; Montes, J.; Bodell, N.; **Carrier, B.**; Sertic, J.V.L.; Barrios, B.; Cater, P.; Davis, D.; Manning, J.W.; DeBeliso, M. "Wearable Device Price is Correlated with the Limits of Agreement Range as a Measure of Heart Rate Validity during Trail Running," Southwest Am. Coll. Sports Med., Oct. 2019, Costa Mesa, CA, USA
52. **Carrier, B.**; Holmes, T.; Williams, L.; Dahl, S.; Weber, L.; Creer, A.; Standifird, T. "Validation of Garmin Fitness Tracker Biomechanics," Am. Coll. Sports Med., May 2019, Orlando, FL, USA
53. Jolley, B.W.; **Carrier, B.**; Standifird, T.; Creer, A. "Validation of Garmin Fitness Tracker Metabolic Data (VO2max)," Am. Coll. Sports Med., May 2019, Orlando, FL, USA
54. **Carrier, B.**; Richards, S.; Hancock, C.; Brooks, L. "Who Brought the Microbes? Investigating the source of fecal veneer on rock climbing holds," Intermountain Am. Soc. Microbiol., Apr. 2019, Provo, UT, USA
55. **Carrier, B.**; Holmes, T.; Williams, L.; Dahl, S.; Weber, L.; Creer, A.; Standifird, T. "Validation of Garmin Fitness Tracker Biomechanics," Southwest Am. Coll. Sports Med., Oct. 2018, Costa Mesa, CA, USA
56. **Carrier, B.**; Ferguson, D.; Ogden T.H. "Molecular Phylogeny of Baetidae (Ephemeroptera)," Evolution 2017, June 2017, Portland, OR, USA
57. **Carrier, B.**; Ferguson, D.; Ogden T.H. "Molecular Phylogeny of Baetidae (Ephemeroptera)," Utah Conf. Undergrad. Res., Feb. 2017, Orem, UT, USA
58. **Carrier, B.**; Ogden T.H. "Phylogenetic Relationships of Mayfly Family Baetidae (Ephemeroptera)," Utah Conf. Undergrad. Res., Feb. 2016, Salt Lake City, UT, USA