Assessing the Validity of the Garmin Venu SQ for Estimating VO_{2max}

Rachelle Foley

College of Arts and Sciences, Northwest University

Honors Thesis

Clint Bryan and Leland Nielsen, advisors

May 1, 2022

Abstract

This study assesses the accuracy and validity of the Garmin Venu SQ's VO_{2max} estimations. We asked 13 healthy college-aged students (ages 18-23) to complete three exercise trials while wearing the Garmin Venu SQ. Participants first ran two trials of the 1.5-Mile run to get a watch estimation before completing a VO_{2max} test on a lab treadmill to measure the actual VO_{2max} value. The watch estimation was then compared to the lab value to assess the level of agreement between the two measures. The mean estimated VO_{2max} from the Garmin watch was 53 ml/kg/min (ranged from 43 to 63 ml/kg/min). The mean VO_{2max} measured in the lab was 54.1 ml/kg/min (ranged from 37.3 to 72.4 ml/kg/min). The largest difference between the watch and lab results was 14.2 ml/kg/min and the smallest the difference was 2.7 ml/kg/min, with an overall mean difference of 1.1 ml/kg/min. A Bland-Altman plot shows good agreement between the watch estimations and actual VO_2 values, but also shows a proportional bias in high-fitness participants. Those with higher fitness tend to get lower watch predictions, and vice versa, and the spread among the data points is higher in those with higher fitness. Our study found that the estimated VO_{2max} produced by the Garmin Venu SQ has good overall agreement with actual VO_{2max} values measured in a laboratory setting.

Assessing the Validity of the Garmin Venu SQ for Estimating VO_{2max}

Maximal oxygen consumption (VO_{2max}) is recognized as the predominant indicator of cardiorespiratory fitness (CRF). A low cardiorespiratory fitness level is associated with an increase in risk for developing cardiovascular disease (CVD) in all populations (Ross, et al, 2016 & Myers, 2003). The American Heart Association examined the importance of CRF in determining one's risk of CVD and overall mortality and found that CRF can be an overall indicator of total body health (Ross, et al., 2016 & Myers, 2003). However, while it is an important risk factor to consider when assessing a patient's health, it is not routinely checked in clinical settings (Ross, et al., 2016). It is often difficult to perform maximal or submaximal tests of CRF and predictive equations aren't always helpful when one does not have a lot of physiological data. This is where fitness trackers can help individuals estimate their CRF and make sure that they are in a healthy range (Smith, 2021). However, they cannot do this if they cannot trust that their fitness trackers are correct in their estimations of VO_{2max}.

In simple terms, CRF is the ability of one's body to bring in oxygen and utilize it within cells (i.e., muscle fibers) to produce energy. In more scientific terms, CRF is determined by the ability of the respiratory system to perform ventilation and pulmonary diffusion, the ability of the cardiovascular system to effectively deliver oxygen to working musculature, the ability of the muscles to utilize that oxygen via oxidative phosphorylation in the mitochondria to match ATP production and ATP usage during exercise, and the ability of the nervous system to integrate these systems (Ross, et al., 2016). Therefore, if one has good CRF, then we could say that their respiratory system, circulatory system, muscular system, and nervous system are all in good working order, allowing for adequate energy-producing capacity needed to complete physical tasks such as activities of daily living and exercise.

When it comes to risk prediction, both measured and estimated values of CRF are strong (Ross, et al., 2016). These predictions are true regardless of age, sex, the presence of CVD, or other comorbidities. Independent of other risk factors, CRF has a profound impact on one's health. In fact, an increase in fitness of just one MET (metabolic equivalent of a task; a unit associated with CRF representing a VO_2 of 3.5 ml/kg/min) can improve one's chance of survival by 10-25% (Ross, et al., 2016). This means that a person does not necessarily need to have a high VO_{2max} to reap the benefits of increased CRF fitness, they just need to be in a healthy range for their specific characteristics.

One of the simplest ways for an individual to track his or her own CRF is through a fitness tracker. These trackers can come in many forms, but the most convenient are fitness watches. According to a survey from 2019, 1 in 5 U.S. adults report regularly wearing a fitness tracker (Vogels, 2020). Another worldwide survey found that fitness trackers are the number one fitness trend of 2022 (Thompson, 2021). In fact, wearable technology has been identified as the number one fitness trend in six of the last seven years since its debut in 2016. Fitness trackers, such as the Garmin Venu SQ (Garmin) in this study, can provide the wearer countless data points at any time, anywhere. One piece of data that activity trackers can give is an estimation of one's VO_{2max}. VO_{2max} specifically refers to the maximum amount of oxygen that an individual can bring in and then put to use in his or her body during a given task. As previously mentioned, this directly relates to CRF as it quantifies the functional ability of the respiratory system and the cardiovascular system.

To estimate one's VO_{2max}, the Garmin uses Firstbeat Analytics. Firstbeat Analytics is a performance analytics company that has spent years studying heartrate variability to get the best algorithms for heartrate data (Firstbeat Technologies, 2022). The FirstBeat algorithm for VO_{2max}

works by combining one's heart rate and speed data (pace) to see how hard the body is working at a given exercise intensity (Firstbeat, n.d.). It also eliminates outlying data to ensure that the data it uses is representative of the work being done (Firstbeat Technologies, 2014). For instance, if a runner must stop at a crosswalk, they have a speed of zero, but still have a very elevated heart rate. Firstbeat will remove that data point so that the speed and heartrate data are accurate. (Firstbeat, n.d.). Firstbeat has been measured against lab values for many different types of exercise and has been found to be 95% accurate with running activities (Firstbeat Technologies, 2014). In most measurements, the VO_{2max} value from Firstbeat was only off by 3.5ml/kg/min and was evenly distributed around the mean. Firstbeat is considered a reliable technology to estimate VO_{2max} and to use for improving fitness.

With CRF being such an important predictor of health and the individuality of CRF numbers, it is important that those seeking to improve their CRF have accurate numbers to go from. A fitness tracker can be a convenient and consistent tool in seeing one's current fitness and reaching new goals, but it must be accurate if it is going to help someone reduce their risk of CVD and overall mortality. Therefore, we want to ensure that the Garmin watch gives an estimation of CRF that is accurate enough to be a tool to use for health. If it is accurate then many people could begin to use it, and potentially other fitness watches, to keep track of their health and give them some peace of mind about their risks for CVD and other health issues.

Methods

Participants

Participants were recruited from the Northwest University campus via a flyer and a recruiting email (sent via a student listserv). Eligible participants were healthy males and females between the ages of 18-25 years that were willing and able to complete a maximal exercise test.

Procedures

All study procedures were approved by the Northwest University Institutional Review Board. All participants provided written informed consent, adhered to the approved COVID protocols, and completed all three testing sessions at Northwest University. During the initial visit, participants completed the informed consent form and a Physical Activity Readiness Questionnaire (PAR-Q). Each participant was then assigned an alphanumeric code and a Garmin Venu SQ watch for use during testing sessions. Participant age, sex, height, and weight were then recorded and used to complete the setup of the Garmin watch. Participants were then ready to complete the first of two 1.5-mile run tests.

1.5-Mile Runs

The 1.5-mile run tests were completed around the turf football field on the Northwest University campus. The watch was placed on the participant's wrist once outside in order to ensure that the GPS coordinates and the heart rate monitor synced with the watch. Once at the field, the participant completed a three- to five-minute warm-up, consisting primarily of light jogging to reduce the risk of injury. Once the warm-up was complete, the watch was set to "Run" mode and was started once the participant began the test. The participant was instructed to run the 1.5-mile run as fast as they could while keeping a steady pace. The 1.5-mile distance was equivalent to 7.5 laps around the border of the field (verified via meter wheel). Once finished, the watch was stopped and run data was recorded (time, distance, pace, and heart rate). The second 1.5-mile run test followed identical protocols and was completed within 1 week of the initial test. Once both 1.5-mile run tests were completed, the predicted VO_{2max} based on the watch data was recorded for each participant.

VO_{2max} test

The third and final testing session took place within one week of session two and had participants completing a VO_{2max} test on a treadmill to measure actual CRF. Upon arrival to the Exercise Science Laboratory, participants were fitted with a heart rate monitor chest strap and connected to the metabolic analyzer (Parvomedics TrueOne 2400, Salt Lake City, UT) for exercise testing. Subjects begin with a three-minute warm-up on the treadmill (Nordictrack x22i, ICON Health & Fitness, Logan, UT) at 3 miles per hour (mph) and 0% grade. For male participants, the treadmill protocol was as follows: two minutes at 5 mph/0% grade; two minutes at 6 mph/0% grade; two minutes at 7 mph/0% grade, then grade increased by 2% each minute thereafter (speed remains at 7 mph) until volitional exhaustion was reached. The highest recorded VO_2 value was identified as the participant's VO_{2max}. For female participants, the only deviation in protocol was a decrease in treadmill speed. Females ran 1 mph slower at each stage relative to the male protocol. For the test to be counted as a valid "maximal" test, the respiratory exchange ratio (RER) needed to be 1.10 or higher and the highest HR value needed to be within 10 beats of the participant's age-predicted maximal (220-age). Each participant met these criteria for the VO_{2max} testing sessions.

Statistical Analysis

The research goal was to evaluate the level of agreement between the predicted VO_{2max} from the Garmin activity tracker and the measured VO_{2max} in the laboratory (using a metabolic analyzer). There were two challenges presented when doing this. One, we must determine the appropriate method to measure, or assess, the level of agreement. Two, we must determine whether that level of agreement is acceptable (or define the level of agreement that we call "acceptable"). These challenges are addressed by using the Bland-Altman analysis.

It would likely be easiest to simply run a correlation analysis between the two VO_{2max} measures and determine statistical significance based on the strength of that correlation. The problem with a correlation in this instance is that because we are assessing agreement between two methods that are producing the same outcome variable (i.e., VO_{2max}), a significant correlation is to be expected. A correlation, after all, measures the strength of the relationship between one variable and another, not the differences. Therefore, correlation analysis is not recommended as a method for comparing two testing methods that are designed to measure the same variable. The appropriate analysis to describe agreement between two quantitative measurements is the Bland-Altman plot, which is an established method to quantify agreement by constructing limits of agreement (1). The resulting plot (see Figure 1) evaluates the mean difference between the two measures and the spread of the mean difference (standard deviation). Bland & Altman recommended that 95% of the data points should lie within ± 2 SD of the mean difference (1). Specifically, all case-wise differences between two methods showing "good agreement" are expected to fall within the limits of agreement set at ± 2 SD of the average difference. Therefore, a Bland-Altman plot of case-wise differences between the two methods is constructed and 95% confidence intervals are set within the plot.

Results

Thirteen participants (age: 20.2 ± 1.7 yrs.; ht.: 171.2 ± 10.5 cm; wt.: 65.5 ± 6.6 kg) completed all three sessions of the study (Table 1). The participant pool was comprised of seven males and six females. The mean estimated VO_{2max} from the Garmin watch was 53 ml/kg/min (ranged from 43 to 63 ml/kg/min). The mean VO_{2max} measured in the lab was 54.1 ml/kg/min (ranged from 37.3 to 72.4 ml/kg/min). The largest difference between the watch and lab results was 14.2 and the smallest the difference was 2.7 with an average difference of 1.1.

Table 1. Participant characteristics and results

ID	Sex (M-1, F-0)	Age	Height (cm)	Weight (kg)	Watch VO_{2max}	Measured VO_{2max}
1	1	21	178	70	57	60.3
2	0	20	176	70	47	44.2
3	0	19	168	59.5	48	43.6
4	1	22	185	72.7	63	72.4
5	1	18	164.5	58.6	58	64.3
6	0	23	152	62.3	46	52.1
7	1	18	187	80.7	55	52.3
8	0	20	167	59.5	48	43.4
9	1	20	182.5	66.8	60	53.5
10	0	20	160	61.4	43	37.3
11	1	23	167.6	59.5	57	71.2
12	1	21	182.9	70.5	57	68.5
13	0	18	162	59.5	54	46.9

The Bland-Altman analysis reveals several points of interest (Figure 1). First and foremost, all points within the plot lie within the limits of agreement and, therefore, suggest "good agreement" between the two methods of assessing VO_{2max}. However, regression analysis (as a means of post hoc testing) revealed that there is evidence of a proportional bias (p = .002). Based on the plot, proportional bias was evident in two ways. First, it appeared that for participants with a higher level of fitness (a VO_{2max} near or above 60 ml/kg/min), the measured VO_{2max} was higher than the predicted VO_{2max} via the Garmin and vice versa. Second, the spread of the difference scores was greater at higher VO_{2max} values when compared to the spread of the difference scores at lower VO_{2max} values. As a result, the predicted value (Garmin) tended to be slightly higher than the measured value in those participants with fitness levels lower than 60 ml/kg/min. Contrastingly, measured values tended to be higher than the predicted values in those participants with higher fitness levels and the magnitude of difference appeared to be more variable. Despite this proportional bias, the findings suggest that the agreement overall between the predicted and the measured VO_{2max} value is considered to be "good."

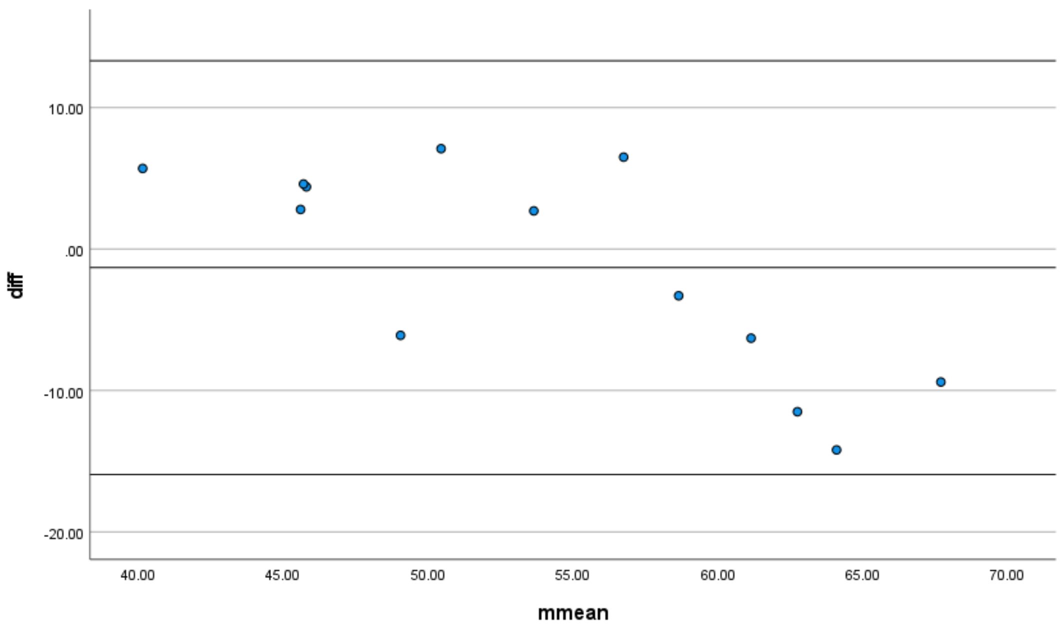

Fig. 1. Bland-Altman Method Comparison Plot (mean of the two VO_{2max} values on the x-axis; the difference between the two VO_{2max} values on the y-axis).

Discussion

The purpose of this study was to assess the validity of the VO_{2max} estimation on the Garmin Venu SQ fitness watch. We especially wanted to see if the watch could be used to give accurate CRF information to help an individual attain their health goals. We concluded that the watch does have good agreement with the actual CRF value measured in a laboratory and can therefore be used with confidence to inform one's fitness and health. However, we did find evidence of proportional bias for those with high VO_{2max} levels. Those with high CRF fitness tended to get lower estimations from the watch relative to the laboratory measurement, which could give them a false understanding of their fitness level. Other studies did not find this difference among fitness levels within their conclusions. The bias in our study is most likely due to the fact that we used young, fit, and healthy participants, many of whom are collegiate athletes. We, therefore, had a larger amount of high fitness levels, and more specifically a large spread of values from average to high fitness, among our participants that allowed for greater data surrounding high fitness participants. That said, it appears that those that would be using the watch for health purposes (i.e., those who have low or average fitness levels and want to improve) would be able to use the watch's estimations with confidence. Therefore, the watch appears to be most accurate for those who need it the most.

To get better data for all fitness levels, it would be beneficial to recruit more participants in the future. While we had sufficient data to make conclusions, having such a small number of participants does limit inference. It would be interesting to expand the age range and get more participants in the high and average fitness categories to see if the bias still stands.

This study did have some limitations. For instance, we found participants on a college campus with a narrower age range. We also recruited participants on a volunteer basis. This led to participants that were active and willing to do the exercise necessary to complete the study. Lastly, none of our participants had low fitness levels. They were all in the average to high range. Therefore, the results may not be widely applicable to other age and fitness populations. In terms of the procedures, there were a few limitations to our study. First, due to limitations on equipment availability, participants only wore the watch twice and for a short period of time. While we do not know, this may have made a difference on the watch's ability to estimate their VO_{2max} as it was unable to collect continuous data from each person over a longer time span. Second, the 1.5-mile run portion of the study was completely up to the participant's own effort and ability to pace themself. The inclusion of heart rate in addition to pace does help account for variations in effort. However, the integration of pace and heart rate is susceptible to heart rate variability due to any number of extraneous factors (i.e., sleep quality/quantity, immune function, time of day). Some may argue that the lab maximal test is also subjective since the person decides when they are exhausted, but there are end-test criteria that each participant satisfied, allowing the test results to be more objective and ensuring that they all ran a true max test to completion.

Our findings are mostly consistent with those of other studies. One study conducted by Freeberg and colleagues assessed the accuracy of the Fitbit Charge 2's (FB) VO_{2max} estimate (2019). The Fitbit differs from the Garmin in that it does not use the FirstBeat algorithm but its own. Freeberg et al. found that the FB's measurements were consistent and unbiased, but overall high (2019). This differed from our results because they did not find the proportional bias that we did. Their study had 30 participants compared to our 13 and had a wider age range, going from 18-35. This most likely led to more evenly distributed data and lower VO_{2max} numbers than ours.

Passler et al. found less optimistic results in their study on fitness tracker's ability to estimate VO_{2max} (2019). They looked at many watches, but the two most relevant were the Garmin Forerunner 920XT (GF) and the Garmin Vivosmart HR (GV). They found that the GF significantly underestimated the VO_{2max}, but fell within the 10% error necessary for a fitness tracker to be considered accurate. However, Passler et al. found other internal reasons to deem the watch unfit to be used for VO_{2max} estimates in sport or health. They also found that the GV deviates significantly from the gold standard for VO_{2max} estimation and therefore should not be used for sport or health either. Our findings differ as we found that the Garmin watch (and therefore the Firstbeat technology) were acceptably accurate. However, we did not look at as many variables as this study and did not approach our study with as strict of expectations.

Conclusion

Our study found that the Garmin Venu SQ has good overall agreement with the measured lab values for VO_{2max}. However, it does show some bias in that the watch underestimates those with a high VO_{2max} and has a wider spread in data at higher fitness levels. Although interesting, this bias was not concerning for the purpose of this study as we wanted to focus on the use of the watch for health. It appears that those who would be using the watch for this purpose (i.e., those that have low or average fitness levels and want to improve) would be able to use the watch's estimations with confidence. Therefore, the watch is the most accurate for those who need it the most.

References

Firstbeat (n.d.). Fitness level (VO_{2Max}). *Firstbeat Technologies.* https://www.firstbeat.com/en/science-and-physiology/fitness-level/

Firstbeat Technologies (2014). Automated fitness level (VO2Max) estimation with heartrate and speed data. *Firstbeat Technologies.* https://assets.firstbeat.com/firstbeat/uploads/2017/06/white_paper_VO2max_30.6.2017.pdf

Firstbeat Technologies (2022). Firstbeat. *Firstbeat Technologies.* https://www.firstbeat.com/en/

Freeburg, K. A., Baughman, B. R., Vickey, T., Sullivan, J.A., & Sawyer, B. J. (2019). Assessing the ability of the Fitbit Charge 2 to accurately predict VO_{2Max}. *mHealth 5*(39). http://dx.doi.org/10.21037/mhealth.2019.09.07

Giavarina, D. (2015). Understanding Bland Altman analysis. *Biochemia Medica, 25*(2), 141-151. https://doi.org/10.11613/bm.2015.015

Hutchinson, A. (2016). *Should Your Doctor Check Your VO_{2Max}?.* Runner's World. https://www.runnersworld.com/health-injuries/a20836245/should-your-doctor-check-your-vo2 max/#:~:text=That's%20because%20low%20VO2,depression%2C%20and%20Type%202%20diabetes

Klepin, K., Wing, D., Higgins, M., Nichols, J., & Godino, J. G. (2019). Validity of cardiorespiratory fitness measured with Fitbit compared to VO_{2Max}. *Medicine & Science in Sports and Exercise 51*(11), 2251-2256. https://doi.org/10.1249/MSS.0000000000002041

Larsen, R. T., Wagner, V., Korfitsen, C. B., Keller, C., Juhl, C. B., Langberg, H., Christensen, J. (2021). Effectiveness of physical activity monitors in adults: Systematic review and meta-analysis. *BMJ,* 1-14, *https:*//doi.org/10.1136/bmj-2021-068047

Myers, J. (2003). Exercise and cardiovascular health. *Circulation, 107*(1), e2–e5. https://doi.org/10.1161/01.CIR.0000048890.59383.8D

Passler, S., Bohrer, J., Blochinger, L., & Senner, V. (2019). Validity of wrist-worn activity trackers for estimating VO_{2Max} and energy expenditure. *International Journal of Environment Research and Public Health, 16*(3037). https://doi.org/10.3390/ijerph16173037

Ross, R., Blair, S. N., Arena, R., Church, T. S., Despres, J. Franklin, B. A., Haskell, W. L., Kaminsky, L. A., Levine, B. D., Lavie, C. J., Myers, J., Niebauer, J., Sallis, R., Sawada, S. S., Sui, X., & Wisleff, U. (2016). Importance of assessing fitness in clinical practice: A case for fitness as a clinical vital sign: A scientific statement from the American Heart Association. *Circulation, 134*(24), 653-699. https://doi.org/10.1161/CIR.0000000000000461

Smith, N. (2021). *New Study Finds Cardiorespiratory Fitness (VO_{2Max}) Measured from Activity Data on your iPhone to be Quite Accurate.* MyHealthyApple. https://www.myhealthyapple.com/new-study-finds-cardiorespiratory-fitness-vo2-max-measured-from-activity-data-on-your-iphone-to-be-quite-accurate/

Thompson, W. R. (2021). Worldwide survey of fitness trends for 2022. *American College of Sports Medicine, 26*(1), 11-20. https://doi.org/10.1249/FIT.0000000000000732

Vogels, E. M. (2020). *About One-in-Five Americans Use a Smart Watch or Fitness Tracker.* Pew Research Center. *https://www.pewresearch.org/fact-tank/2020/01/09/about-one-in-five-americans-use-a-smart-watch-or-fitness-tracker/*