Journal for the Measurement of Physical Behaviour, 2018, 1, 97-99
https://doi.org/10.1123/jmpb.2018-0055

Human Kinetics
EDITORIAL

Validation of a Popular Consumer Activity Tracker

Patty Freedson
University of Massachusetts, Amherst

In this issue we feature the paper, "Methods for Activity Monitor Validation Studies: An Example With the Fitbit Charge," by Kathryn DeShaw and colleagues (2018). Kathryn is a doctoral student working under the direction of Dr. Greg Welk at Iowa State University. This paper examines free-living validity of the Fitbit Charge by comparing its performance to the activPal and Actigraph GT3X research grade accelerometers. The correlations between the research grade accelerometers and the Fitbit Charge are fairly high ($r = 0.80$ for daily moderate to vigorous physical activity [MVPA] minutes and $r = 0.76$ for daily steps). However, examination of the mean absolute percent errors were very high (71.5% for MVPA minutes and 30% for steps). Although the reference measures in this study do not directly assess MVPA and steps, they chose research-grade devices that reportedly estimate MVPA and steps reasonably well. These findings certainly have implications for surveillance and intervention studies that are using the Fitbit Charge to estimate MVPA and steps. The authors advocate for standardized methods for consumer device validation and their paper is a good first step in developing best practices for device validation. Kathryn agreed to be interviewed by me about this paper and her responses can be found below. A short commentary is then presented by Dr. Dinesh John (d.john@neu.edu), an associate professor in the Health Sciences Department at Northeastern University and an Associate Editor for this journal.

Interview With Kathryn J. DeShaw, MS, PhD Student at Iowa State University

What are the key 'take-home' messages from this study?

The study provided a comprehensive evaluation of the Fitbit Charge under free-living conditions using a robust reference method. There are no perfect criterion measures for extended, field-based studies, but the direct comparison of the Fitbit to a novel pattern recognition methodology called SIP (Sojourns Including Posture) provided a way to evaluate agreement across a full week of free-living activity in participants. The need for precision depends on the nature of the study, but the present results document high amounts of individual error when comparing the Fitbit to a more established reference device under free-living conditions.

In your paper you discuss the importance of standardizing methods for device validation. Can you provide a set of recommendations that you consider as key elements to standardizing validation methods to estimate various indices of physical activity and sedentary behavior?

Yes, a main message we sought to convey was the value of standardized reporting methods in these types of studies. There are many monitors available in the market and standardized methods are critical to enable the relative advantages and disadvantages of monitors to be better understood. One of our main recommendations is to report mean absolute percentage error (MAPE) as a standardized metric of individual error (for individual activities as well as for extended periods of time). The reporting of MAPE allows comparisons across studies since it is independent of the device, intensity, duration, and choice of activities compared; however, it is also important to report mean percent error (MPE) to quantify the direction of error. We also advocate for the use of 'equivalence testing' as a formal statistical test of agreement (instead of tests designed to evaluate differences) and the use of Bland-Altman plots to examine distribution of error and potential bias. These indicators can be used in both laboratory studies with simulated free-living activities or in true field evaluation studies like we used in this study.

Given the large mean absolute percent error in estimating MVPA from the Fitbit Charge, do you think this device can be employed in clinical trials to assess changes in MVPA?

Caution is certainly warranted when using these devices in clinical studies—especially if the goal is to evaluate individual changes in MVPA. As documented in the paper, the MAPE values reflective of individual error are quite large and the relative amount of systematic versus random error is difficult to determine. The monitors are likely acceptable for studies aimed at capturing overall group-level differences or for studies seeking to quantify relative amounts of activity in a sample. The devices are probably best suited for promoting adoption of self-monitoring and tracking of physical activity because this is what they were designed to do.

There are many ongoing clinical trials that use the Fitbit Charge or some Fitbit device as an outcome (or exposure) measure. At least for future trials employing the Fitbit Charge, how can the results from your study be used to inform researchers about methodological issues such as sample sizes for treatment and control groups? If you would like to discuss other methodology factors, please do so.

It is important for researchers to know the limitations of these consumer-based devices for research applications. While these devices are now widely used in clinical trials, researchers should be encouraged to collect additional measures and to employ stronger methods to enable triangulation of outcomes and results. Monitor location, reliability/durability, wear time, and syncing of data must be considered. Additional work is clearly needed to understand sources of individual variability and sensitivity to change for these types of applications. Lastly, as the results from this study provide indicators of individual error, they can be used by researchers in power calculations to help ensure that studies are appropriately powered to detect changes.

Freedson (psf@kin.umass.edu) is Professor Emerita, Dept. of Kinesiology, University of Massachusetts, Amherst.

We invite you to add other remarks as needed.

Physical activity and sedentary behavior are complex and independent behaviors that remain difficult to assess. The field has made great strides, but a key need is to continue working towards monitoring methods that would allow estimates from different monitors to be more directly comparable. The use of consumer blood pressure monitoring devices provides a good example because they enable consumers to independently and conveniently check blood pressure using the same metrics employed in health care settings. These devices vary in size and features, as well as accuracy, but the use of standardized metrics (i.e., mmHg) enables values to be directly compared. Similar steps are needed to promote standardization in consumer activity monitors. However, continued innovation is also needed to improve precision in research-grade monitors and methods to enable efficient use in larger trials and studies. The use of standardized reporting in validation studies will help to promote comparisons across devices.

Commentary by Dinesh John, PhD, Associate Professor, Health Sciences Department, Northeastern University

Validation studies are crucial for identifying accurate methods to measure physical behaviors. Such studies inform the scientific community on selecting appropriate measurement tools that are used to draw valid inferences on associations among physical behaviors and health outcomes. The validation study by DeShaw et al. (2018) highlighted significant shortcomings of a specific device offered by a leading brand of consumer wearables, in measuring two commonly used metrics of physical activity, MVPA and steps. This type of work is relevant because variability among an assortment of proprietary consumer devices accompanied by the application of such devices in both small scale and population-level scientific studies (Althoff et al., 2017; "Fitbit Selected for National Institutes of Health (NIH) Precision Medicine Research Program with The Scripps Research Institute (TSRI)," 2017; Wang et al., 2015) has the potential to increasingly limit cross-study comparability of behavioral outcomes. Such validation studies further reinforce the need to "standardize our definitions and measures" (Freedson, 2018) in studies using consumer wearables to measure physical behaviors.

Manufacturers of consumer wearables typically optimize device performance based on various factors including hardware (e.g., dynamic range of the sensor, sampling rate) and firmware (i.e., onboard data pre-processing such as signal filtering) specifications, battery life, data storage and compression, and the volume of onboard data processing (i.e., proprietary methods that translate motion signals into behavior attributes). Variability in few or several of these factors will compromise cross-study comparisons when using consumer wearables. In their paper, DeShaw et al. importantly advocate for uniformity in specific, but controllable aspects of consumer-wearable validation (i.e., validation data type and statistical tests of comparison). Cross-study uniformity in such aspects may enable a fairer comparison of the same outcome derived using two different consumer wearables.

Studies are already using robust criterion methodologies, including direct observation and video analyses of naturally occurring behavior attributes, to calibrate and validate research and consumer devices (Lyden, Petruski, Staudenmayer, & Freedson, 2014; Toth et al., 2018). While completely standardizing free-living protocols among studies is not feasible, carefully designed laboratory protocols that subject proprietary consumer devices to variations in typical behaviors (e.g., ambulation) and related conditions (e.g., walk/run, speed) and activities that hinder signal detection, and, thereby, behavior capture (e.g., wrist-worn step counting when holding an umbrella), may be necessary to better understand both the performance of the sensor within a consumer wearable and the validity of the methodology used to derive physical behavior attributes from sensor data. Such a menu of activities may be comprised of essential activity types that are included when testing any device and activities that are novel to a sensor type (e.g., accelerometer *vs*. gyroscope), or other influencing parameters (e.g., wear-location). A consensus among the research community on standardized and rigorous lab-based testing protocol that accompanies free-living validation against direct observation or video analyses may help in improving our understanding of "when" and "why" a device fails.

While the preference of analytical strategies may vary/evolve, a practice that is common in engineering/computer science, which allows researchers to inspect a method or device used in a scientific publication, is the access to raw data used in the analyses (via data sharing platforms). Easy access to data from published work may allow future researchers to reanalyze previous data that can then be contrasted impartially with current techniques.

It is highly likely that advancing sensor-based innovations will continue to influence the measurement of physical behavior. While the volume of information that can be extracted from such innovations holds immense promise in improving our understanding of how different physical behaviors impact health, the rate of device adoption tends to greatly exceed the accrual of scientific evidence on device validity and reliability. A consensus on standard testing protocols and comparison strategies (including data transparency and sharing) may significantly elevate the quality of evidence used by end-user research scientists when choosing a suitable activity monitor for a study, and thereby better inform device adoption.

Readers of this commentary or other papers published in the Journal for the Measurement of Physical Behaviour (JMPB) are invited to offer their reactions and comments to specific papers or commentaries. You can send your 'letters to the editor' to Patty Freedson at psf@kin.umass.edu. We will offer authors a chance to respond to comments and reactions. The letters and responses to letters will be published in a later issue of JMPB.

References

Althoff, T., Sosic, R., Hicks, J.L., King, A.C., Delp, S.L., & Leskovec, J. (2017). Large-scale physical activity data reveal worldwide activity inequality. *Nature, 547*(7663), 336–339. PubMed ID: 28693034 doi:10.1038/nature23018

DeShaw, K.J., Ellingson, L., Bai, Y., Lansing, J., Perez, M., & Welk, G. (2018). Methods for activity monitor validation studies: An example with the Fitbit Charge. *Journal for the Measurement of Physical Behaviour*, *1*(3).

Fitbit Selected for National Institutes of Health (NIH) Precision Medicine Research Program with The Scripps Research Institute (TSRI). (2017). [Press release]. Retrieved from https://investor.fitbit.com/press/press-releases/press-release-details/2017/Fitbit-Selected-for-National-Institutes-of-Health-NIH-Precision-Medicine-Research-Program-with-The-Scripps-Research-Institute-TSRI/default.aspx