Appalachian
STATE UNIVERSITY

Department of Economics Working Paper

Number 16-04 | February 2016

A Practical Validation Study of a Commercial Accelerometer Using Good and Poor Sleepers

David L. Dickinson
Appalachian State University

Joseph Cazier
Appalachian State University

Thomas Cech
Appalachian State University

Department of Economics
Appalachian State University
Boone, NC 28608
Phone: (828) 262-2148
Fax: (828) 262-6105
www.business.appstate.edu/economics

“A practical validation study of a commercial accelerometer using good and poor sleepers.”

David L. Dickinson

Joseph Cazier

Thomas Cech

Summary

We validated the performance of Fitbit sleep tracking devices against research-grade actigraphy across four days/nights on 38 young adult good and poor sleepers. Fitbit devices underestimated changes in nightly sleep compared to standard actigraphy. Nevertheless, we estimated the Fitbit captures 88% (poor sleepers) to 98% (good sleepers) of actigraphy estimated sleep time changes, which may still be useful for qualitative sleep analysis over time.

Keywords: Sleep tracking, Fitbit, validation, actigraphy

INTRODUCTION

The usefulness and validity of research-grade actigraphy devices are well known (see Sadeh, 2011, and references therein). The rise in interest regarding consumer sleep tracking devices for research implies the need for testing such devices against accepted sleep monitoring technologies. This paper reports results from a validation study of the Fitbit sleep tracking device against standard actigraphy. Fitbit is a leading maker of devices that claim to track sleep, though recent validation attempts have produced mixed results (Meltzer *et al*., 2015; Montgomery-Downs *et al*., 2012; Evenson *et al*., 2015).[1] A summary of the claims and validity of numerous consumer sleep monitors is found in Russo *et al*. (2015), with a focus on the question of their possible usefulness even absent clinical-level data validity. Our study intends to contribute to this debate. We find that the Fitbit, while generally producing biased point estimates of sleep time and efficiency, accurately tracks directional changes of nightly sleep over time, even in poor sleepers. This implies contexts where the Fitbit may be useful, such as in testing the impact of a sleep intervention on a given subject.

MATERIALS AND METHODS

We recruited 38 adult subjects (23 females, 15 males; 26.05 ± 7.99 years old) who each simultaneously wore a commonly utilized research-grade actigraph (Actiwatch Spectrum Plus) and a Fitbit Charge HR device for 4 weekdays/nights. Both the Actiwatch and Fitbit were set to sample data at 30-second epochs, and the Fitbit was set to "normal" mode. We use the Pittsburg Sleep Quality Index (PSQI) to identify both good ($PSQI \leq 5$; n=20) and poor sleepers ($PSQI > 5$). Subjects kept sleep diaries, and we report both raw and diary-adjusted Fitbit data on total sleep time and efficiency.[2] Subjects were compensated $50 for participation and procedures were approved by the Institutional Review Board at the authors' university

On the first day, subjects visited our lab and provided written informed consent, completed the PSQI, received device instructions, and were assigned both an Actiwatch and a Fitbit device. Before departing, subjects were instructed to return to the lab each day for approximately twenty-minutes. During this time, they completed sleep diaries online and lab technicians synced Fitbit devices with lab computers and downloaded subjects' Fitbit and Actiwatch data from the previous day.

Statistical Analysis

We compare each subject's nightly sleep measure to the analogous actigraphy-produced measure: time-in-bed (TIB), total sleep time (TST), sleep efficiency (as automatically device-scored), and TST/TIB (which we call quasi-efficiency). Actigraphy data are scored using validated procedures, and we examine Fitbit

[1] Other consumer sleep trackers have been the subject of validation tests. For example, similar mixed results have been found in validation studies of the Jawbone UP device (Toon *et al*., 2015; de Zombotti *et al*., 2015; Evenson *et al*., 2015).

[2] The procedures used for diary-aided scoring of the Fitbit data were similar to validated actigraphy procedures (e.g., Goldman *et al*., 2007). Because subjects simultaneously wore both devices, this assured that the diary-aided scoring of both Fitbit and actigraphy data utlized the exact same sleep diary record.

measures of TST using both raw and diary-adjusted data. To our knowledge, existing validation studies of consumer monitoring devices do not always adjust device data with input from sleep diaries, even though this is common in many research studies.[3]

For each outcome measure, *M*, we estimate the following:

(1) $Fitbit(M) = \alpha + \beta * Actigraphy(M) + \varepsilon$

Where ε is a random effects error term accounting for the multiple observations (n=4) per subject. The null hypotheses that both α=0 and β=1 implies Fitbit outcomes are statistically no different than Actiwatch outcomes. Rejection of α=0 reflects a general over/underestimation by Fitbit of the actigraphy-based measure. Rejection of β=1 indicates hypo- or hyper-sensitivity of the Fitbit to changes in the outcome measure, compared to actigraphy.

RESULTS

Figures 1, 2, and Table 1 summarize the key results. Figure 1 shows the scatterplot Fitbit data measures (TST and Efficiency) compared to the analogous Actigraphy measure, with the linear regression estimate of equation (1) superimposed. Table 1 shows the full estimation results of TST, sleep efficiency (shown in Fig. 1), Time-in-Bed (TIB), and quasi-efficiency (not shown in Fig. 1) as well as estimates for the separate subsamples of good and poor sleepers. In most instances, Table 1 indicates that the Fitbit generally overestimates TIB, TST, and Efficiency relative to the actigraphy measure (i.e., rejection of α=0 in favor of α >0). The results most closely approximate α=0 and β=1 for the subsample of good sleepers, for whom we estimate that the Fitbit measure of (diary-adjusted) TST is statistically indistinguishable from the actigraphy TST measure.

FIGS 1&2 HERE

Figure 1 and Table 1 results are based on diary-adjusted (i.e., "scored") Fitbit measures, as is typically done with actigraphy data. Figure 2 reproduces equation (1) estimates for TST using raw data as directly provided to the consumer by the Fitbit device. The scatterplot shows the reduction in Fitbit measure variance resulting from diary-aided scoring. Though not reported in Table 1, estimates of equation (1) using Fitbit raw data fail to reject α=0 and β=1. However, we attribute this to the higher variance in the raw Fitbit data (see Fig. 2). To further highlight the importance of diary-aided scored of Fitbit data, we note that the correlation between the Actiwatch raw versus scored data is .9582, compared to .6327 between Fitbit raw versus scored data.

TABLE 1 HERE

[3] Some devices require user activation of "sleep mode", which may serve as a diary-type measure. The Fitbit Charge HR does not require such user activation. Also, some validation studies involve concurrent PSG data acquisition, but it is not always clear whether consumer device data is adjusted as part of scoring.

DISCUSSION

Given the prevalent use of actigraphy for extra-lab monitoring of subject sleep levels, we aimed to assess the practical usefulness of the Fitbit device as an alternative to actigraphy in certain contexts. The National Sleep Foundation places significant emphasis on sleep level targets and guidelines, and they routinely identify sleep deficits by comparing nightly sleep guidelines to self-report measures. One use of low-cost sleep monitoring devices may be to help assess within-subject sleep trends in settings where clinical accuracy is not necessary. In other words, consumer sleep tracking devices may still be *qualitatively* useful for personal goal tracking or even some applied research purposes (e.g., did intervention X significantly increase John Doe's nightly sleep?).

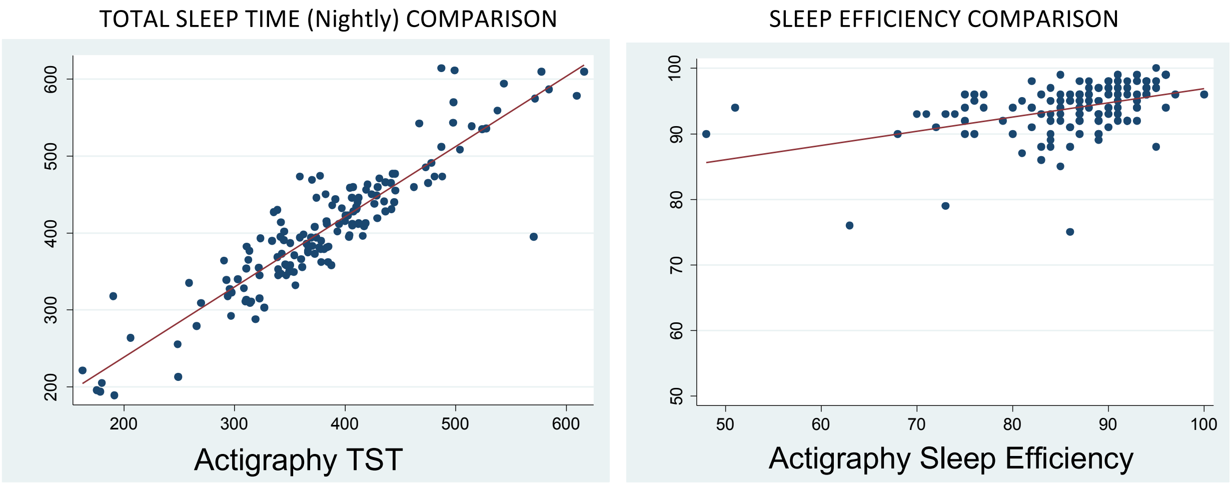
Our statistical analysis finds that diary-adjusted Fitbit data show fairly good quantitative predictions on the key TST variable for good sleepers, and at least qualitatively useful data on poor sleepers. The statistical fit between actigraphy and Fitbit sleep efficiency (and it is unclear how that is defined with Fitbit) is poor, which suggests the use of the quasi-efficiency measure, TST/TIB. This is not surprising given the manual construction of TST/TIB benefits from diary-aided scoring. In sum, while caution should be exercised to properly score Fitbit data, we find it can be a useful and informative measure of subject sleep measures in certain contexts.

ACKNOWLEDGEMENTS: We gratefully acknowledge funding from the XXXXX University YYYYY Program. Valuable research assistance with the data collection and preparation was also provided by WWWWW and BBBBBBB.

CONFLICTS OF INTERESTS: The authors report no conflict of interests. Actigraphy devices were purchased for an earlier project funded through NSF grant BCS-1229067 to ZZZZZZ. Fitbit devices purchased directly through Amazon.com.

AUTHOR CONTRIBUTIONS: DLD and JC designed the study; DLD and JC wrote the protocol; JC and TC collected the data; DLD conducted the analysis; TC scored the data; DLD interpreted the data; all authors contributed to writing. All authors have approved the final manuscript.

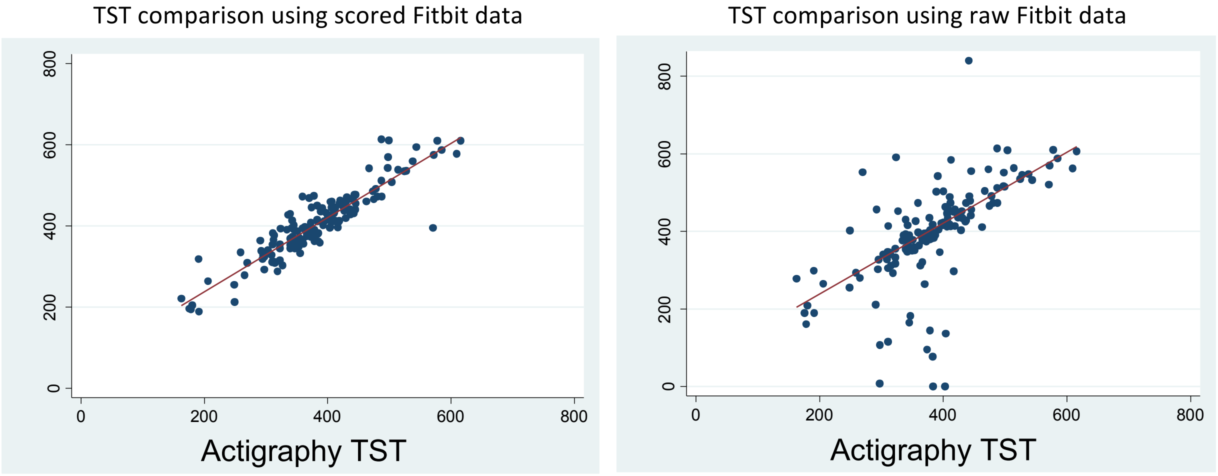
FIGURE 1: FitBit v. Actigraphy (OLS line fit shown)



TOTAL SLEEP TIME (Nightly) COMPARISON

SLEEP EFFICIENCY COMPARISON

FIGURE 2: FitBit scored v. raw nightly TST data comparison to actigraphy (OLS line fit shown)



TST comparison using scored Fitbit data

TST comparison using raw Fitbit data

Note: Left panel of Fig. 2 reproduces the left panel of Fig. 1 with axis rescaled for comparability with raw Fitbit data

TABLE 1: FitBit outcome measures regressed against Actigraphy measures.

TIB	**Dependent Variable=FitBit Time in Bed**		
Variable	**(1)** All Subjects (n=152)	**(2)** Good Sleepers (PSQI ≤ 5; n=72)	**(3)** Poor Sleeperss (PSQI > 5; n=80)
Constant α	104.981 (22.035)***	84.577 (27.908)***	122.113 (31.93)***
Actigraphy TIB β	.854 (.048)***	.912 (.053)***	.803 (.072)***
R-squared	.72	.74	.70
Test of β=1	X^2(1) = 9.32***	X^2(1) = 2.77*	X^2(1) = 7.48***
TST	**Dependent Variable=FitBit Total Sleep Time**		
Variable	All Subjects (n=152)	Good Sleepers (PSQI ≤ 5; n=72)	Poor Sleepers (PSQI > 5; n=80)
Constant α	54.203 (18.649)***	35.121 (22.013)	65.529 (25.247)***
Actigraphy TST β	.917 (.050)***	.974 (.056)***	.879 (.071)***
R-squared	.83	.84	.83
Test of β=1	X^2(1) = 2.69*	X^2(1) =.21	X^2(1) = 2.92*
TST/TIB	**Dependent Variable=FitBit Quasi-Efficiency (TST/TIB)**		
Variable	All Subjects (n=152)	Good Sleepers (PSQI ≤ 5; n=72)	Poor Sleepers (PSQI > 5; n=80)
Constant α	19.342 (13.175)	10.212 (10.845)	24.656 (22.440)
Actigraphy TST/TIB β	.742 (.140)***	.840 (.117)***	.685 (.238)***
R-squared	.26	.38	.19
Test of β=1	X^2(1) = 3.39*	X^2(1) = 1.87	X^2(1) = 1.75
Efficiency	**Dependent Variable=FitBit Efficiency (device defined)**		
Variable	All Subjects (n=152)	Good Sleepers (PSQI ≤ 5; n=72)	Poor Sleepers (PSQI > 5; n=80)
Constant α	76.096 (5.432)***	85.413 (1.767)***	69.368 (10.140)***
Actigraphy Efficiency β	.207 (.061)***	.105 (.021)***	.279 (.115)**
R-squared	.19	.19	.21
Test of β=1	X^2(1) = 168.72***	X^2(1) = 1861.60***	X^2(1) = 39.61***

Notes: Random effects regression models with errors clustered by subject (4 observations per subject). Robust standard errors shown in parenthesis. *, **, *** indicate significance at the .10, .05, and .01 levels, respectively, for the 2-tailed test. Statistical equivalence between Actigraphy and Fitbit outcome variable implies α=0, β=1.

REFERENCES

Buysse, D.J., Reynolds, C.F. III., Monk, T.H., Berman, S.R., & Kupfer, D.J. (1989). The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research. *Psychiatry Research*, *28(2)*, 193-213.

de Zambotti, M., Claudatos, S., Inkelis, S., Colrain, I. M., & Baker, F. C. (2015). Evaluation of a consumer fitness-tracking device to assess sleep in adults. *Chronobiology international*, *32*(7), 1024-1028.

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), 1-22.

Goldman, S. E., Stone, K. L., Ancoli-Israel, S., Blackwell, T., Ewing, S. K., Boudreau, R., ... & Newman, A. B. (2007). Poor sleep is associated with poorer physical performance and greater functional limitations in older women. *Sleep*, *30*(10), 1317-1326.

Meltzer, L. J., Hiruma, L. S., Avis, K., Montgomery-Downs, H., & Valentin, J. (2014). Comparison of a Commercial Accelerometer with Polysomnography and Actigraphy in Children and Adolescents. *Sleep*, *38*(8), 1323-1330.

Montgomery-Downs, H. E., Insana, S. P., & Bond, J. A. (2012). Movement toward a novel activity monitoring device. *Sleep and Breathing*, *16*(3), 913-917.

Russo, K., Goparaju, B., & Bianchi, M. T. (2015). Consumer sleep monitors: is there a baby in the bathwater? *Nature and science of sleep*, *7*, 147-157.

Sadeh, A. (2011). The role and validity of actigraphy in sleep medicine: an update. *Sleep medicine reviews*, *15*(4), 259-267.

Toon, E., Davey, M. J., Hollis, S. L., Nixon, G. M., Horne, R. S., & Biggs, S. N. (2015). Comparison of Commercial Wrist-Based and Smartphone Accelerometers, Actigraphy, and PSG in a Clinical Cohort of Children and Adolescents. *Journal of clinical sleep medicine: JCSM: official publication of the American Academy of Sleep Medicine*. (e-pub ahead of print)