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

David L Dickinson[1,2,3], Joseph Cazier[1] and Thomas Cech[4]

Abstract

We validated a Fitbit sleep tracking device against typical research-use actigraphy across four nights on 38 young adults. Fitbit devices overestimated sleep and were less sensitive to differences compared to the Actiwatch, but nevertheless captured 88 (poor sleepers) to 98 percent (good sleepers) of Actiwatch estimated sleep time changes. Bland–Altman analysis shows that the average difference between device measurements can be sizable. We therefore do not recommend the Fitbit device when accurate point estimates are important. However, when qualitative impacts are of interest (e.g. the effect of an intervention), then the Fitbit device should at least correctly identify the effect's sign.

Keywords

actigraphy, Fitbit, longitudinal studies, sleep, validation studies

Introduction

The usefulness and validity of research-grade actigraphy devices are well known (Sadeh, 2011). The rise in interest regarding consumer sleep tracking devices for research implies the need for testing such devices against accepted sleep monitoring technologies. This article 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, although recent validation attempts have produced mixed results (Evenson et al., 2015; Meltzer et al., 2014; Montgomery-Downs et al., 2012). Other consumer sleep trackers have also been the subject of validation tests. Validation studies of the Jawbone UP device, for example, have produced similar mixed results (de Zambotti et al., 2015; Evenson et al., 2015; Toon et al., 2015). 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. As we will show, our data are somewhat in line with previous conclusions. We provide evidence suggesting serious reservations about using the Fitbit device if accurate measurements are desired, but it may prove useful for qualitative purposes in certain settings.

Methods

This study adhered to the guidelines outlined in the Declaration of Helsinki as revised in 2008. We recruited 38 adult participants (23 females, 15 males; 26.05 ± 7.99 years) who each simultaneously wore a commonly utilized research-grade actigraph (Actiwatch Spectrum Plus: Philips Respironics) and a popular commercial sleep tracker (Fitbit Charge HR) 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 used the Pittsburgh Sleep Quality Index (PSQI; Buysse et al., 1989) to identify both good (PSQI $\leqslant$ 5; $n = 20$) and poor sleepers (PSQI > 5). Participants kept sleep diaries, and we report both raw and diary-adjusted Fitbit data on total sleep time (TST) and

[1]Appalachian State University, USA
[2]Institute for the Study of Labor (IZA), Germany
[3]Chapman University, USA
[4]National Council for Community and Education Partnerships, USA

Corresponding author:
David L Dickinson, Department of Economics and Center for Economic Research & Policy Analysis, Appalachian State University, 416 Howard Street, Boone, NC 28608, USA.
Email: dickinsondl@appstate.edu

efficiency. The procedures used for diary-aided scoring of the Fitbit data were similar to validated actigraphy procedures (Goldman et al., 2007). Because participants simultaneously wore both devices, this assured that the diary-aided scoring of both Fitbit and Actiwatch data utilized the exact same sleep diary record. Participants were compensated US$50 for participation, and procedures were approved by the Institutional Review Board in the Office of Research Protection at Appalachian State University (IRB approval #15-0325).

On the first day, participants 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, participants were instructed to return to the lab each day for approximately 20 minutes. During this time, they completed sleep diaries online and lab technicians synced Fitbit devices with lab computers and downloaded participants' Fitbit and Actiwatch data from the previous day.

Statistical methods used

We compare each participant's Fitbit nightly sleep measure to the analogous actigraphy-produced measure: time-in-bed (TIB), TST, sleep efficiency (as automatically device-scored), and TST/TIB (which we call quasi-efficiency). As noted above, Actiwatch data are scored using validated procedures, and we examine Fitbit 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. 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 polysomnographic (PSG) data acquisition, but it is not always clear whether consumer device data are adjusted as part of the scoring procedure.

For each outcome measure, M, we estimate the following linear model:

$$\text{Fitbit}(M) = \alpha + \beta \times \text{Actigraphy}(M) + \varepsilon \qquad (1)$$

where ε is a random effects error term accounting for the multiple observations ($n = 4$) per participant (i.e. error terms are clustered by participant). The null hypotheses that both $\alpha = 0$ and $\beta = 1$ imply Fitbit outcomes are statistically no different than Actiwatch outcomes on average. Rejection of $\alpha = 0$ reflects a general over/underestimation by Fitbit of the actigraphy-based measure. Rejection of $\beta = 1$ indicates hypo- or hyper-sensitivity of the Fitbit to changes in the outcome measure, compared to actigraphy. All estimations of model (equation (1)) were performed using the panel data random effects option in Stata 13 software.

We also performed Bland–Altman analysis on the differences in device measurements (Bland and Altman, 1986). Enhanced Bland–Altman plots were constructed using SAS software, and these plots include the linear prediction and 95 percent confidence interval on the difference between the outcome measures of the two devices (sleep time or sleep efficiency).

Finally, our unique longitudinal approach (most studies validate a device based on one night with PSG measures, for example) allows us to examine whether any systematic measurement differences between devices are a function of multiple measurements on the same participant.

All reported results are based on diary-adjusted (i.e. "scored") Fitbit and Actiwatch measures, as is typically done with actigraphy data. Diary-adjusted scoring of the Fitbit data significantly reduces the variance in sleep outcome measures from the Fitbit (see section "Results"). In fact, the correlation between the Actiwatch raw versus scored data is .9582, compared to .6327 between Fitbit raw versus scored data. Diary adjustments are used not to calibrate all the device data to match the diary, but rather the diary is used as a complement to the device data when sleep start/stop times are ambiguous in the device data record.

Results

Figure 1 and Table 1 summarize the key correlational results, while Figure 2 highlights the importance of the diary-aided scoring of the Fitbit data (i.e. manual adjustments of raw Fitbit data similar to typical scoring procedures used with actigraphy in sleep research studies). In Figure 1, the scatterplot Fitbit data measures (TST and efficiency) are compared to the analogous Actiwatch measure, with the linear regression estimate of equation (1) superimposed. Table 1 shows the full estimation results of TST, sleep efficiency (shown in Figure 1), TIB, and quasi-efficiency (not shown in Figure 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 Actiwatch measure (i.e. rejection of $\alpha = 0$ in favor of $\alpha > 0$). The results most closely approximate $\alpha = 0$ and $\beta = 1$ for the subsample of good sleepers, for whom we estimate that the Fitbit measure of TST is statistically indistinguishable from the Actiwatch TST measure. This correlational analysis does not, however, draw our attention to the differences between device measurements, which may be sizable and still produce a high correlation measure between devices.

Standard and enhanced Bland–Altman plots showing measurement differences between devices were constructed for TST, sleep efficiency, and quasi-efficiency measures. In Figures 3 to 5, we show results from analysis on the pooled sample as well as the subsamples of good and poor sleeper data for each of these measures. The enhanced plots (right-hand side panel in each figure) include a linear prediction
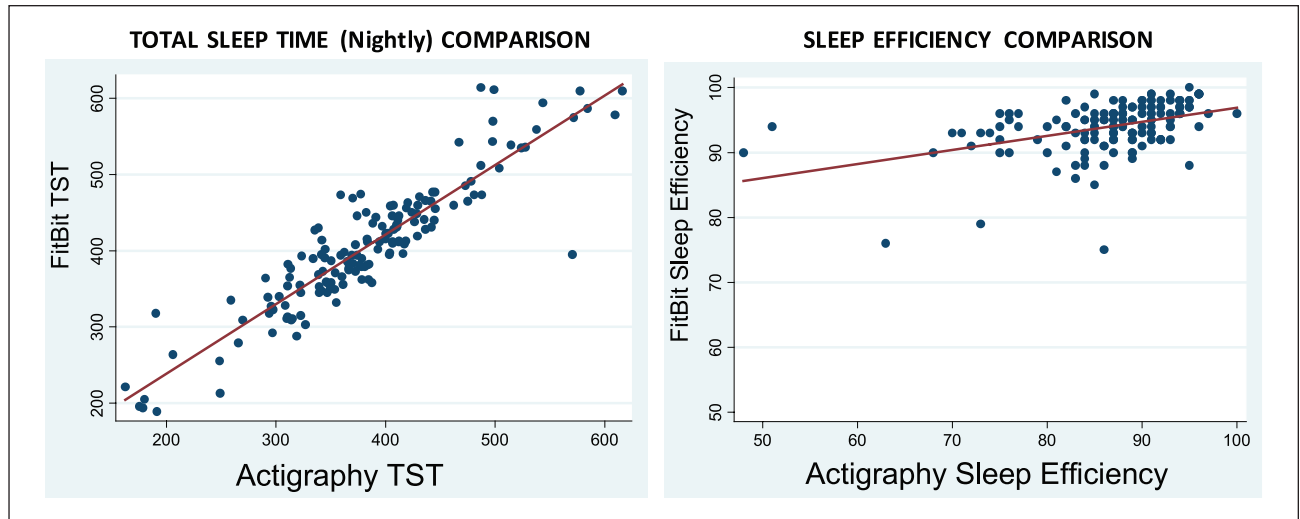
Figure 1. Fitbit versus actigraphy (ordinary least squares line fit shown).

Table 1. FitBit outcome measures regressed against actigraphy measures.

TIB	Dependent variable = Fitbit TIB		
	(1)	(2)	(3)
Variable	All subjects ($n = 152$)	Good sleepers (PSQI $\leqslant$ 5; $n = 72$)	Poor sleepers (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^2	.72	.74	.70
Test of $\beta = 1$	$X^2(1) = 9.32$***	$X^2(1) = 2.77$*	$X^2(1) = 7.48$***
TST	**Dependent variable = Fitbit TST**		
Variable	All subjects ($n = 152$)	Good sleepers (PSQI $\leqslant$ 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^2	.83	.84	.83
Test of $\beta = 1$	$X^2(1) = 2.69$*	$X^2(1) = .21$	$X^2(1) = 2.92$*
TST/TIB	**Dependent variable = Fitbit quasi-efficiency (TST/TIB)**		
Variable	All subjects ($n = 152$)	Good sleepers (PSQI $\leqslant$ 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^2	.26	.38	.19
Test of $\beta = 1$	$X^2(1) = 3.39$*	$X^2(1) = 1.87$	$X^2(1) = 1.75$
Efficiency	**Dependent variable = Fitbit efficiency (device defined)**		
Variable	All subjects ($n = 152$)	Good sleepers (PSQI $\leqslant$ 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^2	.19	.19	.21
Test of $\beta = 1$	$X^2(1) = 168.72$***	$X^2(1) = 1861.60$***	$X^2(1) = 39.61$***

TIB: time in bed; TST: total sleep time.
Random effects regression models with errors clustered by participant (four observations per participant). Robust standard errors shown in parenthesis. Statistical equivalence between actigraphy and Fitbit outcome variable implies $\alpha = 0$, $\beta = 1$.
*, **, ***Significance at the .10, .05, and .01 levels, respectively, for the two-tailed test.
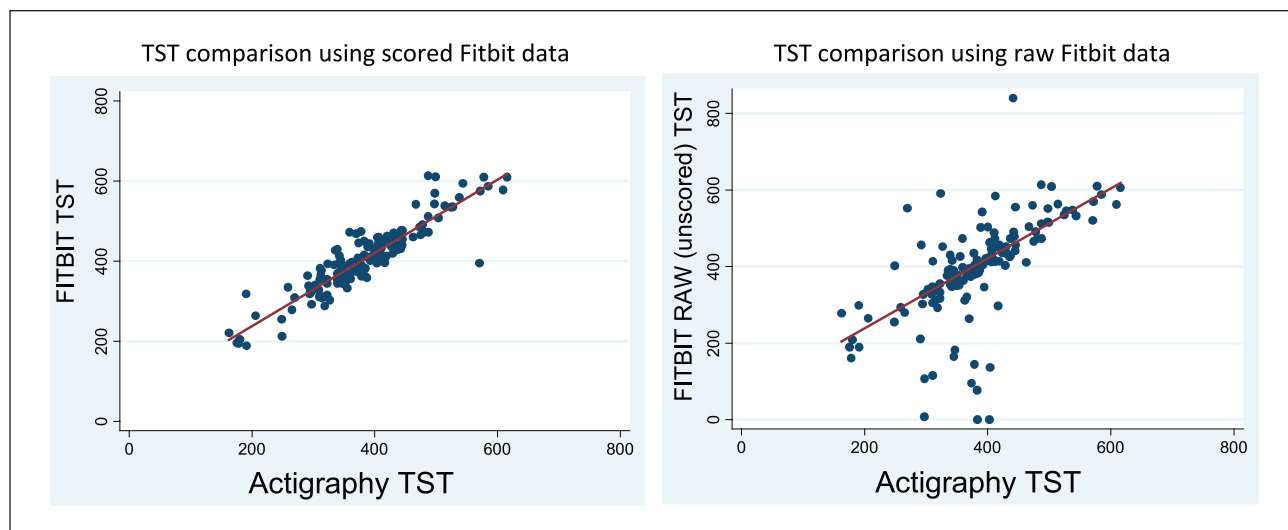
Figure 2. Fitbit scored versus raw nightly TST data compared to actigraphy (ordinary least squares line fit shown).
Left panel of Figure 2 reproduces the left panel of Figure 1 with axis rescaled for comparability with raw Fitbit data.

of the measurement difference and confidence intervals on that difference.

From Figures 3 to 5, we see a key concern regarding device point-estimate reliability of the Fitbit. In many instances, the difference in sleep parameter measurement is not only outside the random variation one might expect, but also the magnitude of the differences is substantial. Also, the Bland–Altman plots reveal that longer recorded values of Fitbit TST or sleep efficiency are associated with an even larger difference between the sleep parameters that the Fitbit and Actiwatch are measuring. Given the Actigraph Spectrum is a well-validated and commonly used research device for obtaining such sleep measures (i.e. it is our benchmark device between the two), this finding indicates that the Fitbit is not sufficiently accurate in the precision of its measurements compared to well-accepted device standards.

Finally, we also conducted longitudinal analysis on whether the day of testing (day 1, 2, 3, or 4) revealed any systematic tendencies regarding the difference between Fitbit and Actiwatch device measurements. The longitudinal data on each participant are shown in Figures 6 and 7, which in each case are separated by good and poor sleepers. One can see that the data show that the Fitbit tends to overestimate sleep efficiency and marginally overestimate TST across all days, but regression results in Tables 2 and 3 confirm no systematic trend across days. Thus, the Fitbit may yet provide useful information regarding the qualitative change in a participant's sleep trends, even though the specific values are likely biased.

Discussion

Given the prevalent use of actigraphy for monitoring participant sleep levels outside of a sleep laboratory environment, 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-participant 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 reasonable correlation on the key TST variable for good sleepers and somewhat lower but still high correlation on TST for poor sleepers. The regression fit between Actiwatch and Fitbit sleep efficiency (and it is unclear how that is defined with Fitbit) is inferior, which suggests that perhaps the use of the quasi-efficiency measure, TST/TIB, may be more reasonable. Nevertheless, the correlation between Fitbit and Actiwatch quasi-efficiency is substantially lower than the correlation between their TST measures.

Additional analysis with Bland–Altman plots show that the magnitude of the differences between device measurements can be substantial. In some instances, the difference in nightly sleep measured by the Fitbit is more than a full hour different from the analogous Actiwatch measure. Also, confidence intervals on the predicted difference between device measurements as a function of the Fitbit measure typically do not include the "zero difference" line, and the predicted difference in device measurements is not constant across the range of values in our data set. Finally,
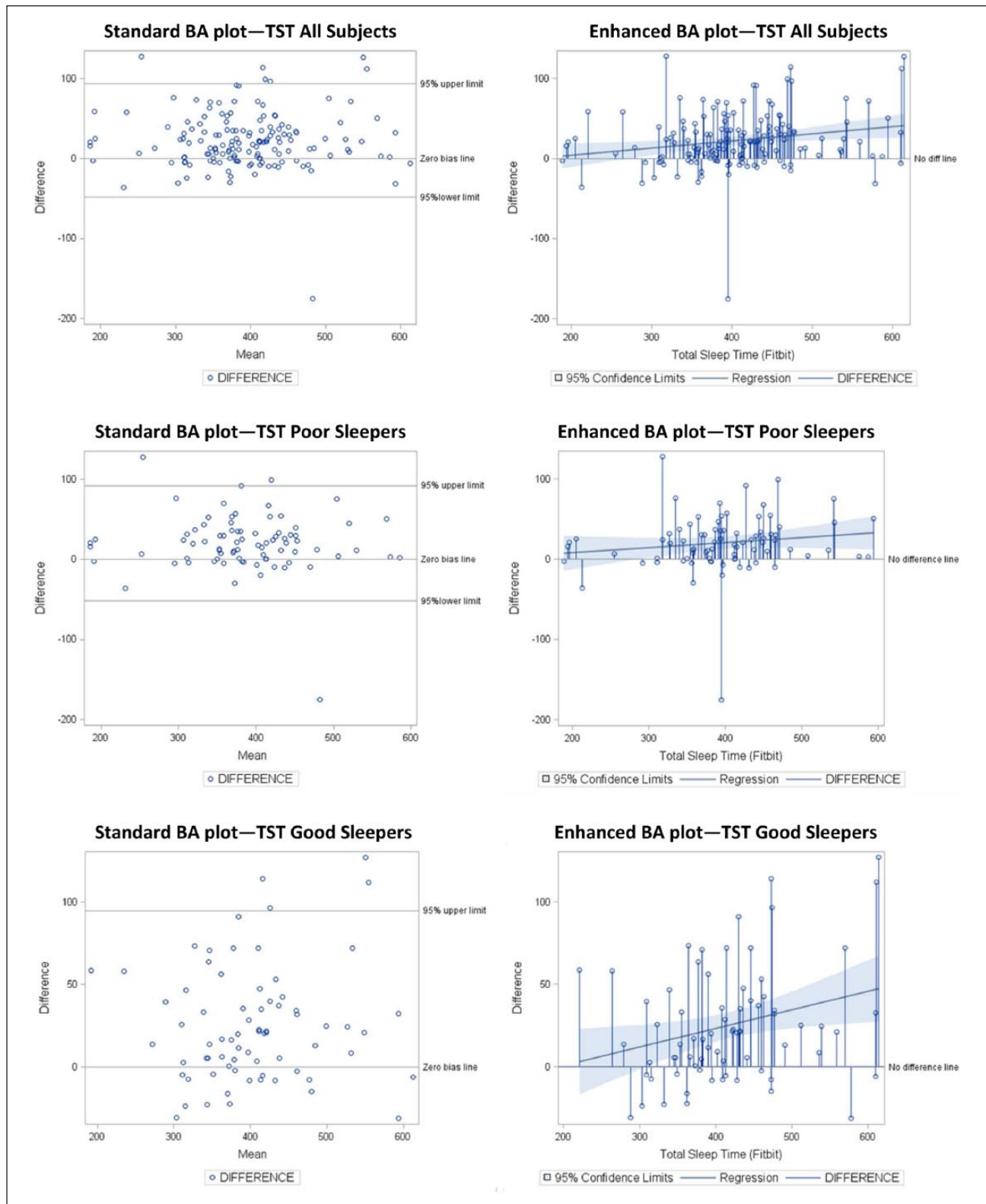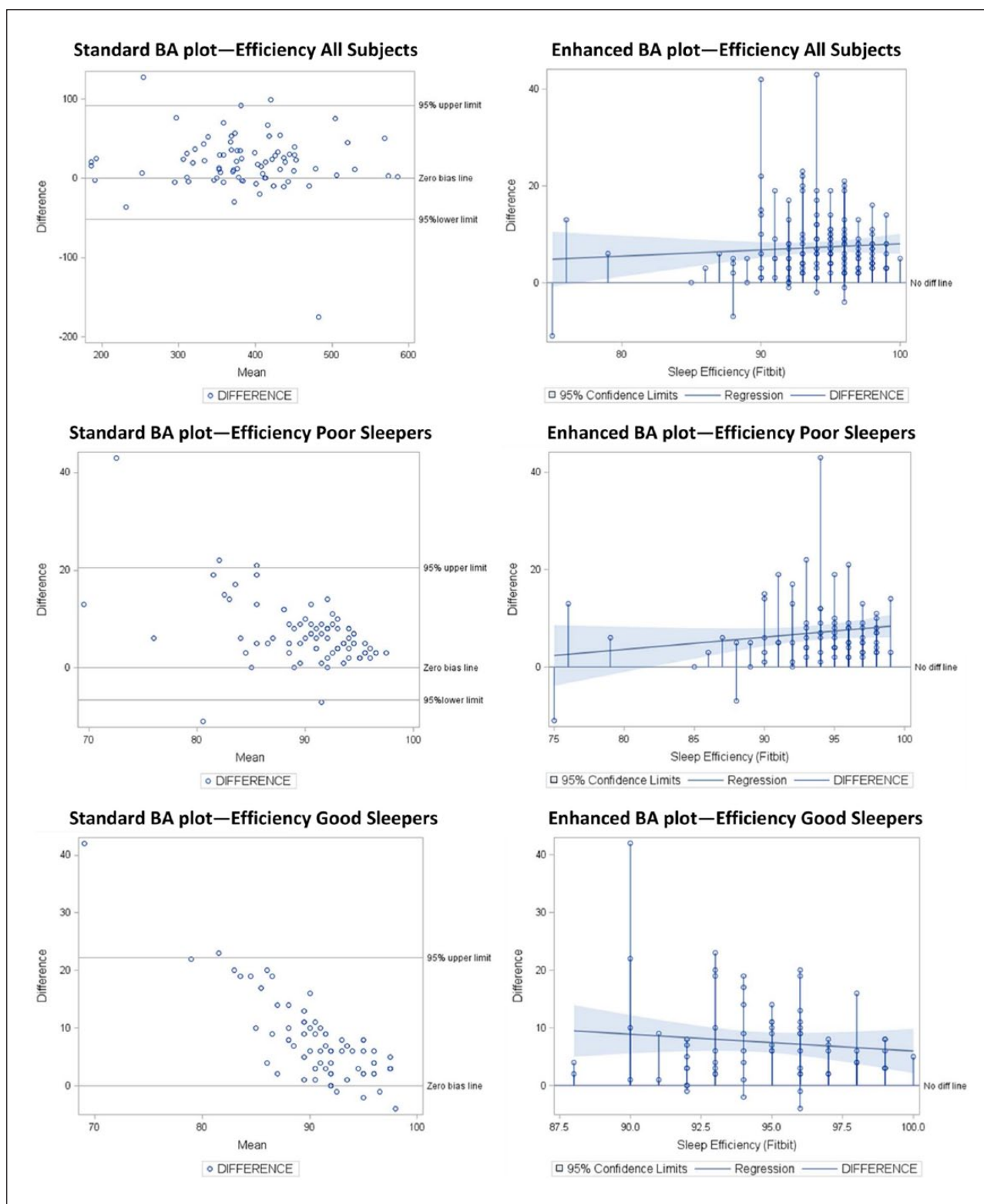
Figure 3. Bland–Altman plots—total sleep time (TST). 95 percent confidence interval shown.

we exploit the unique longitudinal nature of our data set by examining whether the difference between Fitbit and Actiwatch measures of TST and sleep efficiency differs systematically over the course of the four evenings of data
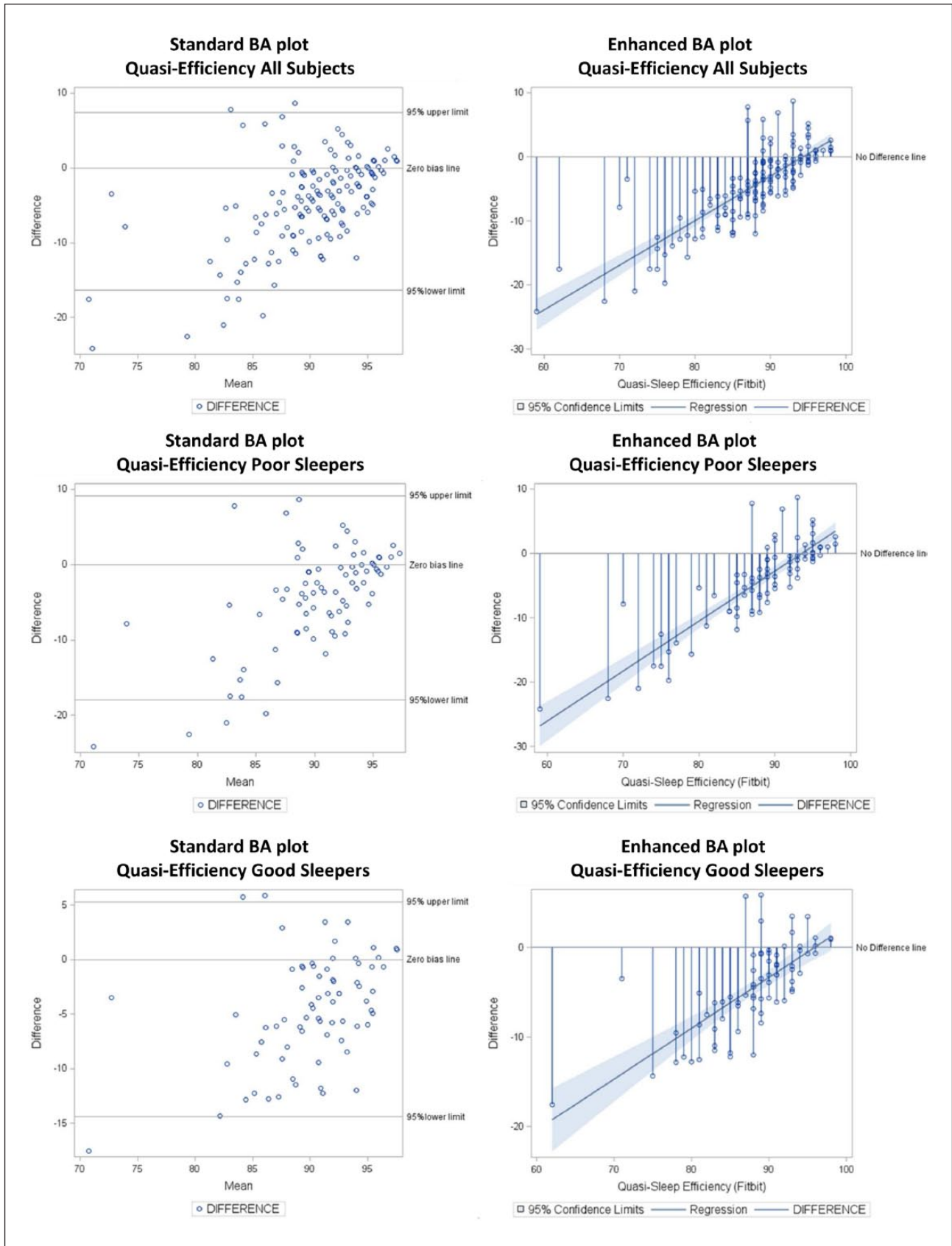
Figure 4. Bland–Altman plots—efficiency (device-scored). 95 percent confidence interval shown.

collection—such analysis is not possible with typical validation studies examining only a single night of device testing. We do not find evidence of differences in device measurement differences across consecutive evenings of testing. Overall, while the Fitbit may be useful for promoting a heightened awareness and concern over one's sleep,

Figure 5. Bland–Altman plots—quasi-efficiency (TST/TIB). 95 percent confidence interval shown.
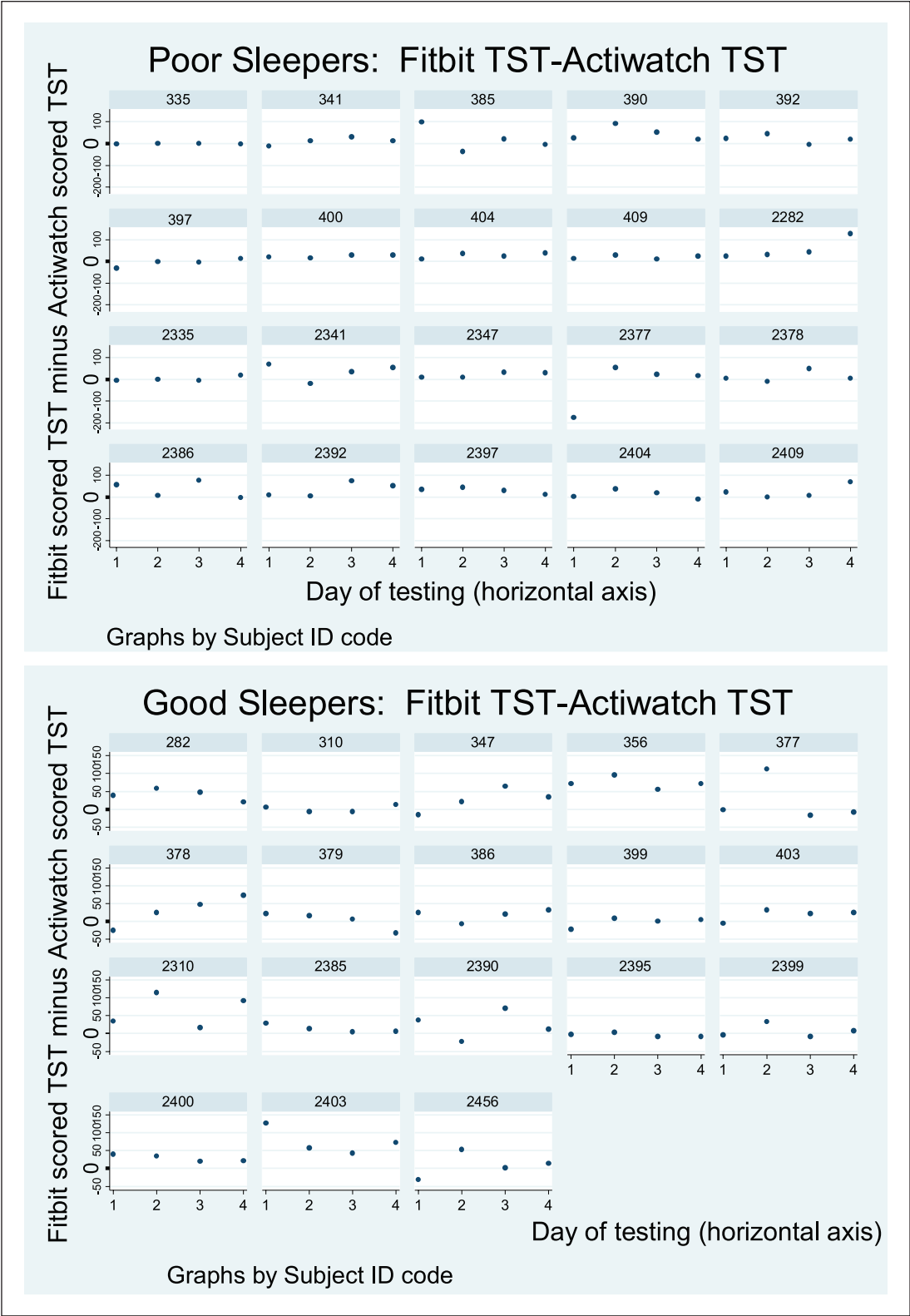
Figure 6. Fitbit–Actiwatch longitudinal TST device differences.

we do not recommend it as an alternative to traditional actigraphy when accurate point estimates of TST or sleep efficiency are desired. However, the significant positive average relationship between device measurements
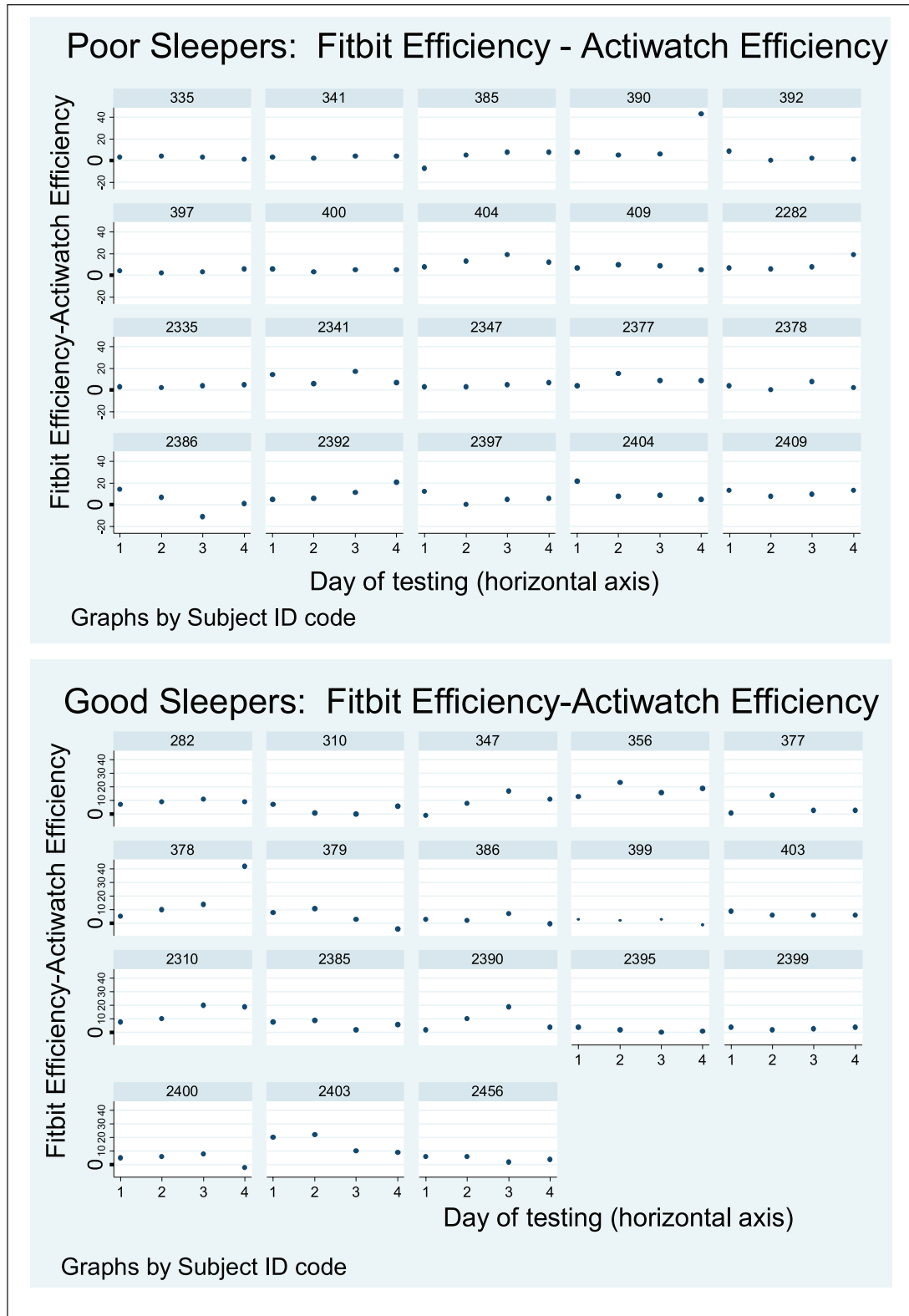
Figure 7. Fitbit–Actiwatch longitudinal device-measured sleep efficiency differences.

suggests a limited but useful role for the Fitbit for those instances where the average sign of the effect is all the researcher needs (e.g. assessing the directional impact of an intervention, assuming sufficient sample size). The qualitative value of the Fitbit data appears to be present for both good and poor sleepers.

Table 2. Longitudinal analysis of the difference between Fitbit TST and Actiwatch TST (dependent variable is Fitbit TST–Actiwatch TST).

Variable	All subjects ($n=152$)	Poor sleepers ($n=80$)	Good sleepers ($n=72$)
Constant	13.737 (7.526)*	9.900 (11.926)	18.000 (9.317)*
Day 2	12.500 (9.675)	7.675 (15.596)	17.861 (11.550)
Day 3	10.658 (7.577)	17.400 (12.083)	3.167 (9.017)
Day 4	11.868 (7.526)	16.125 (11.926)	7.139 (9.445)
Model test (X^2)	2.37	4.46	2.46

Random effects regression models with errors clustered by participant (four observations per participant). Robust standard errors shown in parenthesis. Impact of each identified day in the study timeline is in comparison with day 1 (the omitted reference group in the set of indicator variables).
*, **, ***Significance at the .10, .05, and .01 levels, respectively, for the two-tailed test.

Table 3. Longitudinal analysis of the difference between Fitbit sleep efficiency and Actiwatch sleep efficiency (dependent variable is Fitbit efficiency–Actiwatch efficiency).

Variable	All subjects ($n=152$)	Poor sleepers ($n=80$)	Good sleepers ($n=72$)
Constant	6.684 (.883)***	7.100 (1.366)***	6.222 (1.146)***
Day 2	.105 (1.014)	−1.850 (1.509)	2.278 (1.202)*
Day 3	.632 (1.341)	−.400 (1.919)	1.778 (1.926)
Day 4	1.632 (1.804)	1.900 (2.600)	1.333 (2.623)
Model test (X^2)	.94	4.48	3.63

Random effects regression models with errors clustered by participant (four observations per participant). Robust standard errors shown in parenthesis. Impact of each identified day in the study timeline is in comparison with day 1 (the omitted reference group in the set of indicator variables).
*, **, ***Significance at the .10, .05, and .01 levels, respectively, for the two-tailed test.

References

Bland MJ and Altman DG (1986) Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet* 327(8476): 307–310.

Buysse DJ, Reynolds CF III, Monk TH, et al. (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, et al. (2015) Evaluation of a consumer fitness-tracking device to assess sleep in adults. *Chronobiology International* 32(7): 1024–1028.

Evenson KR, Goto MM and Furberg RD (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 SE, Stone KL, Ancoli-Israel S, et al. (2007) Poor sleep is associated with poorer physical performance and greater functional limitations in older women. *Sleep* 30(10): 1317–1324.

Meltzer LJ, Hiruma LS, Avis K, et al. (2014) Comparison of a commercial accelerometer with polysomnography and actigraphy in children and adolescents. *Sleep* 38(8): 1323–1330.

Montgomery-Downs HE, Insana SP and Bond JA (2012) Movement toward a novel activity monitoring device. *Sleep and Breathing* 16(3): 913–917.

Russo K, Goparaju B and Bianchi MT (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 Review* 15(4): 259–267.

Toon E, Davey MJ, Hollis SL, et al. (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* 12(3): 343–350.