HHS Public Access

Author manuscript

*Behav Sleep Med*. Author manuscript; available in PMC 2021 September 01.

Published in final edited form as:

*Behav Sleep Med*. 2020 ; 18(5): 637–652. doi:10.1080/15402002.2019.1651316.

# Field-based measurement of sleep: Agreement between six commercial activity monitors and a validated accelerometer

**Andrew G. Kubala, M.S.**[1], **Bethany Barone Gibbs, Ph.D.**[1], **Daniel J. Buysse, M.D.**[2], **Sanjay R. Patel, M.D. M.S.**[3], **Martica H. Hall, Ph.D.**[2], **Christopher E. Kline, Ph.D.**[1]

[1]Department of Health and Physical Activity, University of Pittsburgh, Pittsburgh, PA

[2]Department of Psychiatry, University of Pittsburgh, Pittsburgh, PA

[3]Department of Medicine, University of Pittsburgh, Pittsburgh, PA.

## Abstract

**Objective—**To examine agreement between multiple commercial activity monitors (CAMs) and a validated actigraph to measure sleep.

**Methods—**Thirty adults without sleep disorders wore an Actiwatch Spectrum (AW) and alternated wearing 6 CAMs for one 24-h period each (Fitbit Alta, Jawbone Up3, Misfit Shine 2, Polar A360, Samsung Gear Fit2, Xiaomi Mi Band 2). Total sleep time (TST) and wake after sleep onset (WASO) were compared between edited AW and unedited CAM outputs. Comparisons between AW and CAM data were made via paired t-tests, mean absolute percent error (MAPE) calculations, and intra-class correlations (ICC). Intra-model reliability was performed in 10 participants who wore a pair of each AW and CAM model.

**Results—**Fitbit, Jawbone, Misfit, and Xiaomi overestimated TST relative to AW (53.7–80.4 min, P≤.001). WASO was underestimated by Fitbit, Misfit, Samsung and Xiaomi devices (15.0–27.9 min; P≤.004) and overestimated by Polar (27.7 min, P≤.001). MAPEs ranged from 5.1% (Samsung) to 25.4% (Misfit) for TST and from 36.6% (Fitbit) to 165.1% (Polar) for WASO. TST ICCs ranged from .00 (Polar) to .92 (Samsung), while WASO ICCs ranged from .38 (Misfit) to .69 (Samsung). Differences were similar between poor sleepers (Pittsburgh Sleep Quality Index global score >5; n=10) and good sleepers. Intra-model reliability analyses revealed minimal between-pair differences and high ICCs.

**Conclusions—**Agreement between CAMs and AW varied by device, with greater agreement observed for TST than WASO. While reliable, variability in agreement across CAMs with traditional actigraphy may complicate the interpretation of CAM data obtained for clinical or research purposes.

### Keywords

sleep; actigraphy; accelerometry; reliability

**Address for Correspondence:** Christopher E. Kline, Ph.D., Physical Activity and Weight Management Research Center, Department of Health and Physical Activity, University of Pittsburgh, 32 Oak Hill Court, Pittsburgh, PA 15261; chriskline@pitt.edu.

## INTRODUCTION

Sleep is an essential human behavior, with multiple sleep dimensions linked to a variety of health outcomes (Buysse, 2014). The growing evidence supporting this relationship has increased recognition of the importance of sleep by public health practitioners and the general public (Ohayon et al., 2017). Not coincidentally, many commercial activity monitors (CAMs), wearable devices that measure physical activity and other health behaviors, now report the ability to track sleep.

The accuracy of CAMs for assessing physical activity has been comprehensively examined (Bai et al., 2016; Evenson, Goto, & Furberg, 2015; Ferguson, Rowlands, Olds, & Maher, 2015; Kooiman et al., 2015; O'Driscoll et al., in press). In contrast, the sleep-tracking ability of CAMs has been infrequently validated against polysomnography (PSG), the gold standard method of sleep measurement (Cook, Prairie, & Plante, 2018; de Zambotti, Baker, & Colrain, 2015; de Zambotti, Claudatos, Inkelis, Colrain, & Baker, 2015; de Zambotti, Goldstone, Claudatos, Colrain, & Baker, 2018; Kang et al., 2017; Pesonen & Kuula, 2018). While PSG is relatively intrusive, expensive, and non-representative of habitual sleep, actigraphy is a commonly used field-based alternative to measure sleep for both clinical and research purposes (Ancoli-Israel et al., 2015). Actigraphy is well-validated compared to PSG, typically demonstrating excellent sensitivity to detect sleep and a low-to-moderate ability to detect wakefulness (Marino et al., 2013).

While CAMs utilize the same basic technology as actigraphy, their agreement with either PSG or actigraphs is not well-established (Cellini, McDevitt, Mednick, & Buman, 2016; Meltzer, Hiruma, Avis, Montgomery-Downs, & Valentin, 2015; Montgomery-Downs, Insana, & Bond, 2012; Scott et al., 2019; Toon et al., 2016). Additionally, certain CAMs have multisensory components that measure heart rate, movement, and even skin conductance, whereas others rely solely on movement to estimate sleep/wake status. However, the ways in which CAM-specific inputs are utilized for sleep estimation are unknown due to their proprietary algorithms. Nevertheless, research studies are now using CAMs to measure sleep, due in part to the lower cost of these devices compared to traditional actigraphy (Inderkum & Tarokh, 2018; Weaver et al., 2019; Xu et al., 2018). Accordingly, the purpose of the current study was to examine the agreement between sleep-tracking data provided by 6 CAMs (Fitbit Alta, Jawbone Up3, Misfit Shine 2, Polar A360, Samsung Gear Fit2, Xiaomi Mi Band 2) and a traditional actigraph (Philips Actiwatch Spectrum).

## METHOD

### Participants

Participants were recruited from the University of Pittsburgh community by flyers and word of mouth. Recruitment efforts sampled healthy men and women between the ages of 18–60 years. Prospective participants were excluded for any of the following reasons: body mass index (BMI) ≥35 kg/$m^2$, major physical or mental health condition (e.g., depression), pregnant/breastfeeding, excessively short or long self-reported sleep duration (<5 or >10 hours), or self-reported diagnosis or significant symptoms of a sleep disorder (assessed by a

locally developed questionnaire). Thirty participants (15 male, 15 female) provided written informed consent and were monetarily compensated after participation in the primary study. To examine the reliability between two devices of the same model, a subset of 10 participants (5 males, 5 females) consented to participate in ancillary intra-model reliability assessments (described below). All research procedures were approved by the University of Pittsburgh Institutional Review Board.

## Commercial Activity Monitors (CAMs)

Seven CAMs that included sleep monitoring were evaluated in this study. Model names, manufacturers, and software/application versions of the CAMs are summarized in Table 1. Six CAMs were included in the analyses: Fitbit Alta (Fitbit), Jawbone Up3 (Jawbone), Misfit Shine 2 (Misfit), Polar A360 (Polar), Samsung Gear Fit2 (Samsung), and Xiaomi Mi Band 2 (Xiaomi). All devices had sensors for heart rate detection (light sensors: Polar, Samsung, Xiaomi; skin conductance: Jawbone) except for the movement-based Fitbit and Misfit monitors. The multi-sensory Garmin Vivosmart HR (Garmin) was also examined, but a high number of device failures precluded its inclusion in these analyses.

All CAMs were Bluetooth-compatible and synced to study-specific iPads (Apple, Inc.; Cupertino, CA). Data were downloaded and accessible through the application associated with each CAM. Fitbit could either detect sleep with a ‘normal’ or ‘sensitive’ setting; the normal setting was utilized for this study. Each CAM featured automatic sleep detection; although the Jawbone, Misfit, and Xiaomi have manual sleep tracking modes as an option, no inputs regarding sleep/wake status are required to be provided by the participant or researcher. Specific sleep metrics differed in name across CAMs (Table 1), but all devices provided indices of total sleep time (TST) and wake after sleep onset (WASO). Because TST and WASO are highly relevant sleep health outcomes (Buysse, 2014), these indices were compared to the AW.

## Diary and other self-reported sleep data

Participants completed a modified version of the Pittsburgh Sleep Diary (Monk et al., 1994) each day. Each morning, participants indicated the time they went to bed with the intention of going to sleep the prior night and the time that they stopped attempting sleep that morning. These inputs were used to edit the AW rest intervals (see below). At the initial study visit, participants also completed the Insomnia Severity Index (ISI), Pittsburgh Sleep Quality Index (PSQI), and Epworth Sleepiness Scale (ESS) (Bastien, Vallieres, & Morin, 2001; Buysse, Reynolds III, Monk, Berman, & Kupfer, 1989; Johns, 1991) to characterize their insomnia symptoms, sleep quality, and daytime sleepiness, respectively.

## Philips Actiwatch Spectrum (AW)

The AW Spectrum Classic (Philips Respironics; Murrysville, PA) is a widely used device for monitoring sleep in free-living conditions (Kushida et al., 2001; Patel et al., 2015; Quante et al., 2018). The AW can incorporate accelerometry (activity counts), ambient light, and user inputs to estimate rest and sleep intervals. Thus, it was considered the criterion measure for comparisons within this study. AW data were collected in 1-min epochs using Actiware software (version 6.0.9). Participants were instructed to press an event marker when they got

into bed each night with the intention of going to sleep and when they stopped attempting sleep in the morning. Rest intervals were manually established by a trained technician who followed a standardized approach that incorporated the following inputs, ranked in order of importance: event marker, light intensity, sleep diary, and activity counts (Patel et al., 2015). Once rest intervals were established, sleep/wake status for each epoch was determined by the Actiware algorithm using the default settings for sleep onset and offset (10 min and 10 min, respectively) and a wake threshold of 40. TST, defined as the total amount of time scored as sleep from sleep onset to sleep offset, and WASO, defined as the total amount of time scored as awake from sleep onset to sleep offset, were retained for comparisons against CAM outputs. Primary analyses of this study compared unedited CAM data (as might typically be used by a researcher) against edited AW data (best practice method). Analyses comparing unedited CAM data with unedited AW data (i.e., using the automated rest interval setting that incorporated event marker inputs) provided similar results and are presented as supplemental data (Supplemental Tables 3 and 4, Supplemental Figures 1–3).

## Procedures

Two sets of devices (7 CAMs, 1 AW) were utilized in this study, with one set designated for males (Set 1) and another for females (Set 2). Multiple sets were utilized to facilitate timely completion of the study, allow for smaller band sizes to be utilized by the female participants, and ensure optimal CAM placement during wear.

Prior to a participant's arrival, each CAM was updated with the participant's height, weight, age, sex, and handedness. Upon arrival, participants were given a carrying case with labeled CAMs, an AW, and a sleep diary. Participants were instructed to wear two devices each day on the same (nondominant) wrist across seven days. The AW (criterion) was worn continuously for all seven days, while CAMs were changed every 24 hours. CAMs were worn in a specific order based upon their battery life: Samsung, Garmin, Fitbit, Polar, Jawbone, Xiaomi, Misfit. To reduce sample variability in sleep measurement due to device placement on the wrist, participants were randomized to wear the AW closer to or further from the wrist. Each participant provided one night of wear per CAM device and 7 days of wear for the AW; thus, there were 30 nights for comparison of each CAM with an AW. In instances of device failure or user error (e.g., inadequate charge, participant non-wear), participants were asked to re-wear the missing devices for an additional night.

## Ancillary Intra-Model Reliability Study

Two sets of each device (AW and each CAM model) were used in the primary study. To test the reliability between the pairs of each model, 10 participants consented to 8 additional days of device wear. Procedures for device setup were identical to those previously mentioned and participants were instructed to consecutively wear each pair of same-model devices for one 24-hour period in the following order: Samsung, Garmin, Fitbit, Polar, Jawbone, Xiaomi, Misfit, AW. Both devices were worn on the same (nondominant) wrist; participants were randomized to wear one of the devices closer to or further from the wrist. There were 10 nights of comparison between the device pairs.

### Statistical Analysis

Statistical analyses were conducted using IBM SPSS Statistics (v. 23; IBM, Chicago, IL) and significance was set at $P<0.05$. Paired sample t-tests compared the mean differences between CAM and AW data. Agreement between the CAM and AW data was evaluated using intra-class correlation coefficients (ICCs), Bland-Altman plots, and mean absolute percent error (MAPE) calculations. We used ICCs (2,1) with absolute agreement and reported confidence intervals for each estimate; ICC values were classified as poor (<0.50), moderate (0.50–0.75), good (0.75–0.90), or excellent (>0.90) based on established guidelines (Koo & Li, 2016). Bland-Altman plots were used to visualize the CAM-AW differences and evaluate if there was differential bias across the range of values. Limits of agreement (LOA) were computed (mean difference±1.96 SD) to indicate the range in which the differences between the two measures would occur with 95% probability (Bland & Altman, 1986). MAPE values were calculated to indicate the relative measurement error of each CAM compared to the AW for TST and WASO. MAPE was calculated as the absolute difference between the CAM and AW measure divided by the AW measure multiplied by 100 (e.g., [(CAM TST – AW TST)] / AW TST x 100) (Cellini et al., 2016). These analyses were conducted for the overall sample but also following stratification by sleep quality. Those with a PSQI global >5 were classified as poor sleepers, while those with a score ≤ 5 were considered good sleepers.

For the ancillary reliability component of the study, paired t-tests and ICCs were calculated according to device set and device placements (closer to or further from wrist) for each of the 7 devices (6 CAMs, 1 AW). Finally, analyses were stratified by sex to evaluate whether measures of agreement were similar between men and women.

## RESULTS

### Sample Characteristics

Participant characteristics are summarized in Table 2. The sample was primarily white (87%) with a mean age of 24.8±4.1 years. Participants reported mean PSQI scores of 4.9±3.2, ESS scores of 6.4±3.3, and ISI scores of 4.2±3.5. Ten participants (33%) were categorized as having poor sleep quality (PSQI >5; mean PSQI score of 8.6±2.7). Most of the sample (80%) had earned an undergraduate degree or higher. Males and females did not differ on any demographic or sleep characteristic. Additionally, the demographic and sleep characteristics of the reliability subsample (n=10) did not differ from the main study sample.

### TST: Mean Bias

Results of the paired t-tests between unedited CAM and edited AW TST are found in Table 3. Fitbit, Jawbone, Misfit, and Xiaomi each overestimated TST in comparison to AW (>50 min, each $P<.01$). In contrast, Polar underestimated TST in comparison to AW (−81.8 min, $P<.001$). Samsung TST did not differ from AW TST (7.1 min, $P=.27$). In good sleepers, all CAMs overestimated TST compared to the AW (>10 min, each $P<.01$) with the exception of the Polar which underestimated TST in comparison to the AW (−90.4, $P=.004$). In poor sleepers, Xiaomi overestimated TST compared to the AW (75.3, $P=.002$), while all other CAMs TST did not differ from AW TST (>±20 min, each $P>.10$).

### TST: Agreement

ICCs between the unedited CAM and edited AW for TST are reported in Table 3. ICCs for Fitbit and Samsung were classified as good and excellent, respectively. Jawbone, Misfit, and Xiaomi ICCs were classified as moderate, while Polar ICC was classified as poor. MAPE for TST across the CAMs is displayed in Figure 1. Overall, Samsung MAPE was low (5.1%), but MAPE for Fitbit, Jawbone, Polar, Misfit, and Xiaomi ranged from 14.6% (Fitbit) to 25.4% (Misfit). Figure 2 displays Bland-Altman plots of the pattern of differences between unedited CAM and edited AW TST data. Upon visual inspection, Jawbone underestimated TST at low values and overestimated TST at higher values; no other patterns were observed across CAMs. In contrast to the narrow LOA observed for Samsung, wider LOA were seen for Fitbit, Jawbone, Misfit, Polar, and Xiaomi. ICCs and MAPE values were similar across good and poor sleepers.

### WASO: Mean Bias

Results of the paired t-tests between unedited CAM and edited AW WASO are found in Table 4. Fitbit, Misfit, Samsung, and Xiaomi each underestimated WASO in comparison to AW (>14.0 min, each $P \leq .01$). Polar overestimated WASO (27.7 min, $P < .001$), while Jawbone WASO did not differ from AW (6.8 min, $P = .40$). For the Jawbone, Misfit, Polar, and Xiaomi, there were no differences in results between good and poor sleepers. In good sleepers, the Fitbit underestimated WASO compared to the AW (−16.2 min, $P = .02$), while in poor sleepers WASO was not different (−12.5 min, $P = .14$).

### WASO: Agreement

ICCs between the unedited CAM and edited AW WASO data are reported in Table 4. ICCs for Samsung, Fitbit, Jawbone, and Xiaomi were classified as moderate, while Polar and Misfit ICCs were classified as poor. MAPE for CAM WASO ranged from 36.6% (Fitbit) to 165.1% (Polar) (Figure 1). In Figure 3, Bland-Altman plots display the pattern of differences in WASO between the CAMs and AW. Upon visual inspection, the spread in differences increased as WASO increased for Fitbit and Jawbone, suggesting a pattern of increased error with higher WASO. Samsung had the narrowest LOA of all the CAMs. ICCs and MAPE values were similar across good and poor sleepers.

### Intra-Model Reliability of CAM and AW

Results of the intra-model reliability analyses according to device placement and device pairs for TST and WASO are found in Table 5 and Table 6, respectively. Paired t-tests indicated that TST did not differ based upon device placement (each $P \geq .20$; Table 5). In addition, no difference in TST was observed between device pairs for AW, Jawbone, Misfit, Polar, Samsung, or Xiaomi (each $P \geq .15$); TST differed across Fitbit devices ($P = .03$; Table 5). ICCs across devices for TST were classified as excellent with the exception of Jawbone (classified as good).

Paired t-tests indicated that WASO did not differ according to device placement for any of the devices (each $P \geq .17$; Table 6). In addition, WASO did not differ between device sets for AW, Fitbit, Jawbone, Misfit, Polar, or Xiaomi (each $P \geq .08$); WASO differed across Samsung

devices (P=.001). ICCs across devices for WASO were each classified as excellent with the exception of the Jawbone (classified as good).

### Device Failures

In total, CAMs failed to collect sleep data, either because of user error or software/device malfunction, on 38 individual nights. User errors accounted for 18% of data loss (Garmin: 1 night, Jawbone: 1 night, Misfit: 1 night, Polar: 1 night, Samsung: 3 nights). Software or device malfunction accounted for 82% of data loss (Fitbit: 4 nights, Garmin: 17 nights, Jawbone: 2 nights, Polar: 1 night, Samsung: 6 nights, Xiaomi: 1 nights). In every instance, aside from Garmin devices, data were obtained during subsequent re-wear. As a result, each CAM-AW comparison included 30 nights of data.

### Ancillary Analyses

When stratified by sex, the results were similar to the results reported above and are presented as supplemental data (Supplemental Tables 1 and 2). No difference was found between unedited or edited AW TST or WASO data (P=.67 and P=.65, respectively). Using unedited AW data as the comparison against CAM data did not result in substantially different results from those presented above (Supplemental Figures 1–3, Supplemental Tables 3 and 4).

## DISCUSSION

Despite the popularity of CAMs and their increasing use for clinical and research purposes, few data are available to determine whether these devices provide similar sleep estimates to traditional actigraphy. We evaluated agreement between six CAMs and a common actigraph using a study design that directly compared each CAM against the criterion actigraph in a sample of 30 adults. In addition, we evaluated whether pairs of each CAM model provided similar estimates in a subsample of participants.

Broadly, we observed greater agreement between the various CAMs and AW for TST than for WASO. Most CAMs overestimated TST in relation to AW, with Samsung being the only device with a nonsignificant difference in TST, low MAPE, and excellent ICC. Other devices either overestimated (Fitbit, Jawbone, Misfit, Xiaomi) or underestimated TST (Polar) in relation to AW, though Fitbit had a relatively acceptable MAPE and ICC. For WASO, none of the CAMs exhibited excellent agreement with AW. Jawbone WASO did not differ from AW WASO, as others over- (Polar) or underestimated (Fitbit, Misfit, Samsung, Xiaomi) WASO relative to AW. While MAPE for WASO exceeded 35% for each CAM, the absolute mean difference in WASO between Fitbit, Jawbone, and Samsung with AW was small (<16 min) and ICCs suggested modest agreement with AW. In Bland-Altman plots for Fitbit and Jawbone WASO data, a slight funnel shape appears to indicate a greater discrepancy between the respective CAM and AW with greater amounts of WASO. This funnel shape has been observed in validation research involving the Jawbone Up compared to PSG (de Zambotti, Claudatos, et al., 2015).

When compared to the AW, there were no discernible differences in the patterns of agreement between the multi-sensory CAMs (Jawbone, Polar, Samsung, Xiaomi) and

movement-only CAMs (Fitbit, Misfit). Although integrating multiple physiological signals for sleep/wake estimation improves CAM prediction against PSG (Beattie et al., 2017), it is unsurprising that multi-sensory CAMs did not have better agreement with actigraphy since traditional actigraphy estimates sleep/wake status solely on movement.

To the authors' knowledge, this paper is the first to evaluate the sleep–tracking capability of the Samsung Gear Fit2, Polar A360, and Xiaomi Mi Band 2 in comparison with traditional actigraphy and few studies have evaluated the agreement between Fitbit, Jawbone, and Misfit devices with traditional actigraphs (individually discussed below). Our results largely corroborate previous literature comparing Fitbit models with actigraphy. Montgomery-Downs and colleagues found that when compared to AW-64 (a predecessor to the AW Spectrum used in the current study), the Fitbit Classic overestimated TST by 24.1±46.6 minutes and overestimated sleep efficiency, of which WASO is a contributor (Montgomery-Downs et al., 2012). Meltzer and colleagues compared the Fitbit Ultra to AW Spectrum and found that the Fitbit (using the normal setting) overestimated TST and underestimated WASO by >30 min each (Meltzer et al., 2015). Another study found that TST of the Fitbit Charge HR was ~20 min greater than Actiwatch 2 (Lee et al., 2017). In contrast to these results, Cook and colleagues found that the Fitbit Flex produced similar TST estimates compared to Actiwatch 2, but underestimated WASO in a sample of adults with depression (Cook, Prairie, & Plante, 2017). However, the similar TST estimates in the latter study could be attributed to their manual adjustment of Fitbit bed and wake times to match PSG times, which was not performed in the current study. Thus, our findings of Fitbit's overestimation of TST and underestimation of WASO are generally consistent with prior research (Castner et al., 2019; Dickinson, Cazier, & Cech, 2016; Ferguson et al., 2015; Scott et al., 2019; Visovsky, Kip, Rice, Hardwick, & Hall, 2013).

Jawbone devices have been studied in comparison to PSG (de Zambotti, Baker, et al., 2015; de Zambotti, Claudatos, et al., 2015; Mantua, Gravel, & Spencer, 2016), but few comparisons have been made to actigraphs commonly used in sleep research. Previous studies found that the Jawbone Up and Up3 did not differ from Actiwatch 2 in measuring TST and WASO (Cook et al., 2018; Toon et al., 2016). These findings partially conflict with those from our study: we found that the Up3 overestimated TST but provided a similar WASO estimate. Lastly, Ferguson and colleagues explored the concordance between the Fitbit One, Jawbone Up, and Misfit Shine with a BodyMedia SenseWear armband. They found good agreement of sleep time between the CAMs and the SenseWear armband, with ICCs of .71–.90 (Ferguson et al., 2015). These conflicting results could be attributed to the differences in CAM and traditional actigraphs used between prior studies and ours.

Because the performance of CAMs in poor sleepers has been infrequently evaluated, we also examined whether the level of CAM-AW agreement differed between good and poor sleepers as identified by the PSQI. We did not find consistent evidence that would indicate poorer agreement among devices for poor sleepers, as ICCs were of similar magnitude for many devices. These results mirror a study from de Zambotti and colleagues that found the Jawbone UP estimated sleep similarly in women with insomnia or normal sleep when compared to PSG (de Zambotti, Claudatos, et al., 2015). In contrast, a study from Kang and colleagues found that when comparing the Fitbit Flex and an Actiwatch 2 with PSG, ICCs

were classified as excellent for individuals with and without insomnia, but the insomnia group was more likely to have differences of >30 min of TST in the Fitbit Flex compared to the good sleepers (Kang et al., 2017). In a study more similar to the current report, Dickinson and colleagues found lower TST correlations between the Fitbit Charge HR and Actiwatch Spectrum Plus in poor sleepers compared to good sleepers (Dickinson et al., 2016). Our findings may be due to the small sample of poor sleepers (n=10) and the relatively mild sleep disturbance of the poor sleepers, as none indicated significant sleep disorder symptoms.

A secondary purpose of the current study was to evaluate whether identical CAM devices of the same model provided similar TST and WASO estimates. In general, concordance between devices of the same model was excellent for CAMs as well as AW, with all devices other than Jawbone exhibiting an ICC >.90 for both TST and WASO. Additionally, the only differences in TST or WASO observed across each pair of identical models were observed for Fitbit (TST) and Samsung (WASO). Finally, we found that device placement (i.e., closer to or further from the wrist) did not impact TST or WASO estimates for any device pairs. Prior studies that have explored CAM and actigraphy device reliability have found similar results. Montgomery-Downs and colleagues compared two identical Fitbit Classic devices on three participants for a single night and found agreement rates >96% for each participant (Montgomery-Downs et al., 2012). Meltzer et al. found no differences in TST or WASO between two Fitbit Ultra devices worn on the same wrist by 7 children (Meltzer et al., 2015). Similar findings of high intra-device reliability have been observed for other validated actigraphs such as the Fatigue Science Readiband (Driller, McQuillan, & O'Donnell, 2016). Individual CAM devices of the same model appear to have high reliability for a single night of wear and suggest that these devices provide consistent measures of TST and WASO.

Similar to other reports, we also evaluated the number of device failures (Baroni, Bruzzese, Di Bartolo, & Shatkin, 2016; Mantua et al., 2016). The high number of Garmin device failures precluded their inclusion in these analyses. However, after excluding user errors such as inadequate charging or improper wear, few device-based data failures were observed. Only Samsung and Fitbit had >10% device-based data loss. These results are similar to Mantua and colleagues, as they found that the Misfit Shine and the Fitbit Flex (predecessors to the Misfit Shine 2 and Fitbit Alta), had 10% or greater loss due to device-based errors (Mantua et al., 2016). Our results found fewer device failures in the Fitbit Alta compared to other Fitbit models used in previous reports from Baroni and colleagues, showing 70% device failure in the Fitbit Flex (Baroni et al., 2016), and Lillehei and colleagues who found 86% loss in the Fitbit One (Lillehei, Halcón, Savik, & Reis, 2015).

Overall, these data suggest that most CAMs have only modest agreement with a traditional actigraph, particularly for WASO. Despite these findings, our results do not discredit the use of CAMs for clinical or research use. For instance, Dickinson and colleagues suggested that, although CAMs are only able to provide crude estimates of TST and WASO in relation to traditional actigraphy, they can provide insight into the directionality and magnitude of changes in sleep (Dickinson et al., 2016). In a recent review on the use of wearable technology for assessing sleep, de Zambotti and colleagues suggest that regulated CAMs could be useful in clinical settings for the assessment of sleep-wake patterns (de Zambotti et

al., 2019). Additionally, our study utilized an AW as the criterion measure despite its limitations, particularly regarding wake detection, relative to PSG (Marino et al., 2013). Many CAMs use movement to estimate sleep, but more recent models incorporate multiple signals in conjunction with motion (e.g., heart rate, skin temperature). As noted by Goldstone and colleagues, these newer CAMs may predict sleep/wake status more accurately than actigraphs currently utilized in research (Goldstone, Baker, & de Zambotti, 2018). Thus, our findings suggest that TST and WASO estimates are not interchangeable between the AW and most of the CAMs we studied, but do not indicate that CAM data are less accurate than AW data.

Our study has some limitations. First, the sample was primarily healthy young adults with minimal sleep disturbance. Although we stratified our analyses based upon good vs. poor sleeper classification, only 10 participants were categorized as poor sleepers. Additionally, this study only compared one night of device wear per participant. This prohibits any insight into comparisons of night-to-night variability within individuals or weekly averages as might typically be measured for devices. This research focused on the agreement between the readily available sleep outputs from the CAMs and AW that are accessible to consumers and clinicians who may utilize these devices. Thus, this study is limited by a lack of epoch-by-epoch comparison between the CAMs and AW. Details regarding the AW hardware, software, and algorithm are available to researchers, but the proprietary nature of the CAMs' sleep estimation methods introduces added uncertainty and accuracy of measurement. This study compared the automatic sleep detection of the CAMs to the edited AW data to reflect how consumers may utilize the devices compared to how researchers commonly use the AW. With our chosen research methodology, we cannot infer how comparability could change if the manual sleep tracking mode from the Jawbone, Xiaomi, and Misfit were utilized. Finally, we lack comparison between the CAMs and PSG which precludes any determination of whether the AW or the CAMs are measuring TST and WASO with greater accuracy. Because of these limitations, future studies should explore the comparability and accuracy of CAMs in larger samples with poor sleep or sleep disorders (e.g., insomnia), and should include comparison against both traditional actigraphy and PSG, ideally over multiple nights of observation.

In conclusion, we found moderate agreement for TST and WASO measured by six different CAMs as compared to an AW. Of the six CAMs, Samsung, followed by Fitbit and Jawbone, had the highest agreement with the AW across various indices when considering both TST and WASO. Moreover, the CAMs had greater agreement with actigraphy when estimating TST and less agreement when estimating WASO. There were no patterns indicating that CAM performance differed by poor versus good sleeper classification. While CAMs may represent an appealing alternative to clinicians and researchers looking to assess sleep due to their popularity and affordability, they should understand that these devices provide sleep estimates that are not interchangeable with traditional actigraphy.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## ACKNOWLEDGMENTS

All devices evaluated in this study were purchased using laboratory funds set aside for CEK; device manufacturers had no role in the study design, reporting of results, or interpretation of results. Investigator support for CEK was provided by National Institutes of Health grant K23HL118318.

CONFLICTS OF INTEREST:

BBG has received grant support independent of this project from the American Heart Association, National Institutes of Health (NIH), and the Tomayko Foundation. DJB has received grant support independent from this project from the NIH, served as a paid consultant for the American Academy of Physician Assistants, Bayer, BeHealth, CME Institute, Ebb Therapeutics, and Emmi Solutions, and received licensing fees for the Pittsburgh Sleep Quality Index, copyrighted to the University of Pittsburgh. MHH and CEK have received grant support independent of this project from the NIH. SRP has received grant support independent of this project from the NIH, American Sleep Medicine Foundation, ResMed Foundation, Bayer Pharmaceuticals, and Philips Respironics.

## REFERENCES

Ancoli-Israel S, Martin JL, Blackwell T, Buenaver L, Liu L, Meltzer LJ, … Taylor DJ (2015). The SBSM guide to actigraphy monitoring: clinical and research applications. Behavioral Sleep Medicine, 13 Suppl 1, S4–s38. [PubMed: 26273913]

Bai Y, Welk GJ, Nam YH, Lee JA, Lee J-M, Kim Y, … Dixon PM (2016). Comparison of consumer and research monitors under semistructured settings. Medicine & Science in Sports & Exercise, 48(1), 151–158. [PubMed: 26154336]

Baroni A, Bruzzese JM, Di Bartolo CA, & Shatkin JP (2016). Fitbit Flex: an unreliable device for longitudinal sleep measures in a non-clinical population. Sleep and Breathing, 20(2), 853–854. [PubMed: 26449552]

Bastien CH, Vallieres A, & Morin CM (2001). Validation of the Insomnia Severity Index as an outcome measure for insomnia research. Sleep Medicine, 2(4), 297–307. [PubMed: 11438246]

Beattie Z, Oyang Y, Statan A, Ghoreyshi A, Pantelopoulos A, Russell A, & Heneghan C. J. P. m. (2017). Estimation of sleep stages in a healthy adult population from optical plethysmography and accelerometer signals. Physiological Measurement, 38(11), 1968. [PubMed: 29087960]

Bland JM, & Altman D (1986). Statistical methods for assessing agreement between two methods of clinical measurement. Lancet, 327(8476), 307–310.

Buysse DJ (2014). Sleep health: can we define it? Does it matter? Sleep, 37(1), 9–17. [PubMed: 24470692]

Buysse DJ, Reynolds CF III, Monk TH, Berman SR, & Kupfer DJ (1989). The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research. Psychiatry Research, 28(2), 193–213. [PubMed: 2748771]

Castner J, Mammen MJ, Jungquist CR, Licata O, Pender JJ, Wilding GE, & Sethi S (2019). Validation of fitness tracker for sleep measures in women with asthma. Journal of Asthma, 56(7), 719–730. doi:10.1080/02770903.2018.1490753 [PubMed: 29972657]

Cellini N, McDevitt EA, Mednick SC, & Buman MP (2016). Free-living cross-comparison of two wearable monitors for sleep and physical activity in healthy young adults. Physiology & Behavior, 157, 79–86. [PubMed: 26821185]

Cook JD, Prairie ML, & Plante DT (2017). Utility of the Fitbit Flex to evaluate sleep in major depressive disorder: a comparison against polysomnography and wrist-worn actigraphy. Journal of Affective Disorders, 217, 299–305. [PubMed: 28448949]

Cook JD, Prairie ML, & Plante DT (2018). Ability of the multisensory Jawbone UP3 to quantify and classify sleep in patients with suspected central disorders of hypersomnolence: a comparison against polysomnography and actigraphy. Journal of Clinical Sleep Medicine, 14(5), 841–848. [PubMed: 29734975]

de Zambotti M, Baker FC, & Colrain IM (2015). Validation of sleep-tracking technology compared with polysomnography in adolescents. Sleep, 38(9), 1461–1468. [PubMed: 26158896]

de Zambotti M, Cellini N, Goldstone A, Colrain IM, & Baker FC (2019). Wearable Sleep Technology in Clinical and Research Settings. Medicine & Science in Sports & Exercise, 51(7), 1538–1557. [PubMed: 30789439]

de Zambotti M, Claudatos S, Inkelis S, Colrain IM, & Baker FC (2015). Evaluation of a consumer fitness-tracking device to assess sleep in adults. Chronobiology International, 32(7), 1024–1028. [PubMed: 26158542]

de Zambotti M, Goldstone A, Claudatos S, Colrain IM, & Baker FC (2018). A validation study of Fitbit Charge 2 compared with polysomnography in adults. Chronobiology International, 35(4), 465–476. [PubMed: 29235907]

Dickinson DL, Cazier J, & Cech T (2016). A practical validation study of a commercial accelerometer using good and poor sleepers. Health Psychology Open, 3(2), 2055102916679012.

Driller M, McQuillan J, & O'Donnell S (2016). Inter-device reliability of an automatic-scoring actigraph for measuring sleep in healthy adults. Sleep Science, 9(3), 198–201. [PubMed: 28123660]

Evenson KR, Goto MM, & Furberg RD (2015). Systematic review of the validity and reliability of consumer-wearable activity trackers. Internatonal Journal of Behavioral Nutrition and Physical Activity, 12, 159.

Ferguson T, Rowlands AV, Olds T, & Maher C (2015). The validity of consumer-level, activity monitors in healthy adults worn in free-living conditions: a cross-sectional study. Internatonal Journal of Behavioral Nutrition and Physical Activity, 12(1), 42.

Goldstone A, Baker FC, & de Zambotti M (2018). Actigraphy in the digital health revolution: still asleep? Sleep, 41(9), zsy120.

Inderkum AP, & Tarokh L (2018). High heritability of adolescent sleep–wake behavior on free, but not school days: a long-term twin study. Sleep, 41(3), zsy004.

Johns MW (1991). A new method for measuring daytime sleepiness: the Epworth Sleepiness Scale. Sleep, 14(6), 540–545. [PubMed: 1798888]

Kang SG, Kang JM, Ko KP, Park SC, Mariani S, & Weng J (2017). Validity of a commercial wearable sleep tracker in adult insomnia disorder patients and good sleepers. Journal of Psychosomatic Research, 97, 38–44. [PubMed: 28606497]

Koo TK, & Li MY (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of Chiropractic Medicine, 15(2), 155–163. [PubMed: 27330520]

Kooiman TJ, Dontje ML, Sprenger SR, Krijnen WP, van der Schans CP, & de Groot M (2015). Reliability and validity of ten consumer activity trackers. BMC Sports Science Medicine and Rehabilitation, 7(1), 24.

Kushida CA, Chang A, Gadkary C, Guilleminault C, Carrillo O, & Dement WC (2001). Comparison of actigraphic, polysomnographic, and subjective assessment of sleep parameters in sleep-disordered patients. Sleep Medicine, 2(5), 389–396. [PubMed: 14592388]

Lee H-A, Lee H-J, Moon J-H, Lee T, Kim M-G, In H, … Kim, L. (2017). Comparison of wearable activity tracker with actigraphy for sleep evaluation and circadian rest-activity rhythm measurement in healthy young adults. Psychiatry Investigation, 14(2), 179–185. [PubMed: 28326116]

Lillehei AS, Halcón LL, Savik K, & Reis RJ (2015). Effect of inhaled lavender and sleep hygiene on self-reported sleep issues: a randomized controlled trial. The Journal of Alternative Complementary Medicine, 21(7), 430–438. [PubMed: 26133206]

Mantua J, Gravel N, & Spencer RM (2016). Reliability of sleep measures from four personal health monitoring devices compared to research-based actigraphy and polysomnography. Sensors, 16(5), E646.). [PubMed: 27164110]

Marino M, Li Y, Rueschman MN, Winkelman JW, Ellenbogen J, Solet JM, … Buxton OM (2013). Measuring sleep: accuracy, sensitivity, and specificity of wrist actigraphy compared to polysomnography. Sleep, 36(11), 1747–1755. [PubMed: 24179309]

Meltzer LJ, Hiruma LS, Avis K, Montgomery-Downs H, & Valentin J (2015). Comparison of a commercial accelerometer with polysomnography and actigraphy in children and adolescents. Sleep, 38(8), 1323–1330. [PubMed: 26118555]

Monk TH, Reynolds CF 3rd, Kupfer DJ, Buysse DJ, Coble PA, Hayes AJ, … Ritenour AM (1994). The Pittsburgh Sleep Diary. Journal of Sleep Research, 3, 111–120.

Montgomery-Downs HE, Insana SP, & Bond JA (2012). Movement toward a novel activity monitoring device. Sleep and Breathing, 16(3), 913–917. [PubMed: 21971963]

O'Driscoll R, Turicchi J, Beaulieu K, Scott S, Matu J, Deighton K, … Stubbs J (in press). How well do activity monitors estimate energy expenditure? A systematic review and meta-analysis of the validity of current technologies. British Journal of Sports Medicine. doi:10.1136/bjsports-2018-099643

Ohayon M, Wickwire EM, Hirshkowitz M, Albert SM, Avidan A, Daly FJ, … Gozal D (2017). National Sleep Foundation's sleep quality recommendations: first report. Sleep Health, 3(1), 6–19. [PubMed: 28346153]

Patel SR, Weng J, Rueschman M, Dudley KA, Loredo JS, Mossavar-Rahmani Y, … Seiger AN (2015). Reproducibility of a standardized actigraphy scoring algorithm for sleep in a US Hispanic/Latino population. Sleep, 38(9), 1497–1503. [PubMed: 25845697]

Pesonen AK, & Kuula L (2018). The validity of a new consumer-targeted wrist device in sleep measurement: an overnight comparison against polysomnography in children and adolescents. Journal of Clinical Sleep Medicine, 14(4), 585–591. [PubMed: 29609722]

Quante M, Kaplan ER, Cailler M, Rueschman M, Wang R, Weng J, … Redline S (2018). Actigraphy-based sleep estimation in adolescents and adults: a comparison with polysomnography using two scoring algorithms. Nature and Science of Sleep, 10, 13–20.

Scott J, Grierson A, Gehue L, Kallestad H, MacMillan I, Hickie IJS (2019). Can consumer grade activity devices replace research grade actiwatches in youth mental health settings? Sleep and Biological Rhythms, 17(2), 223–232.

Toon E, Davey MJ, Hollis SL, Nixon GM, Horne RS, & Biggs SN (2016). 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. [PubMed: 26446248]

Visovsky C, Kip K, Rice J, Hardwick M, & Hall P (2013). Choosing instruments for research: an evaluation of two activity monitors in healthy women. Journal of Novel Physiotherapies, 3, 171.

Weaver RG, Beets MW, Perry M, Hunt E, Brazendale K, Decker L, … Maydeu-Olivares A (2019). Changes in children's sleep and physical activity during a 1-week versus a 3-week break from school: a natural experiment. Sleep, 42(1), zsy205. doi:10.1093/sleep/zsy205.

Xu X, Conomos M, Manor O, Rohwer J, Magis A, & Lovejoy J (2018). Habitual sleep duration and sleep duration variation are independently associated with body mass index. International Journal of Obesity, 42(4), 794–800. [PubMed: 28895585]

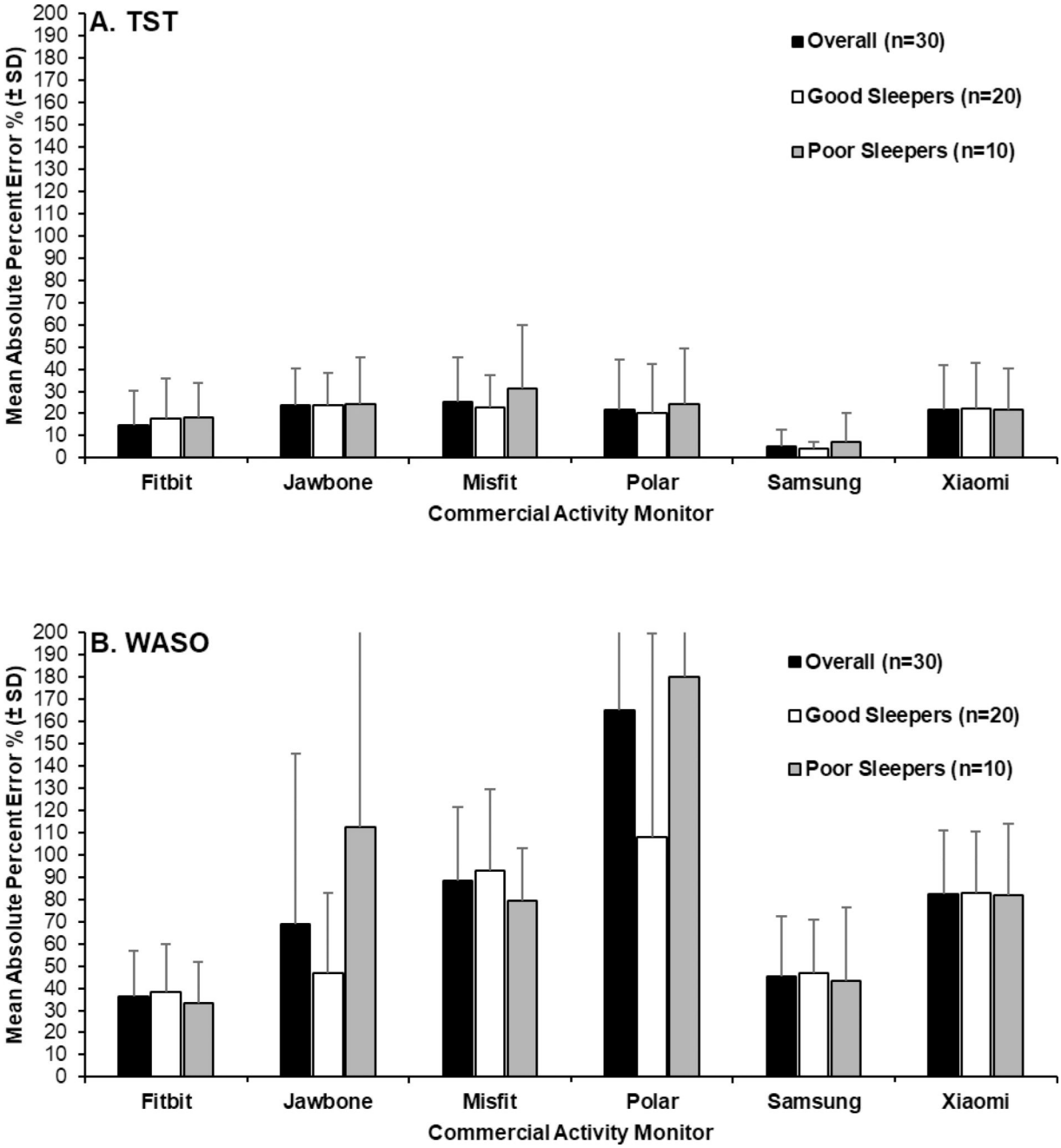


**Figure 1. Mean Absolute Percent Error for comparisons between unedited commercial activity monitor data relative to edited Actiwatch data.**

Panel A: total sleep time; panel B: wake after sleep onset. Data are shown as mean±standard deviation. Abbreviations: SD=standard deviation; TST=total sleep time; WASO=wake after sleep onset.

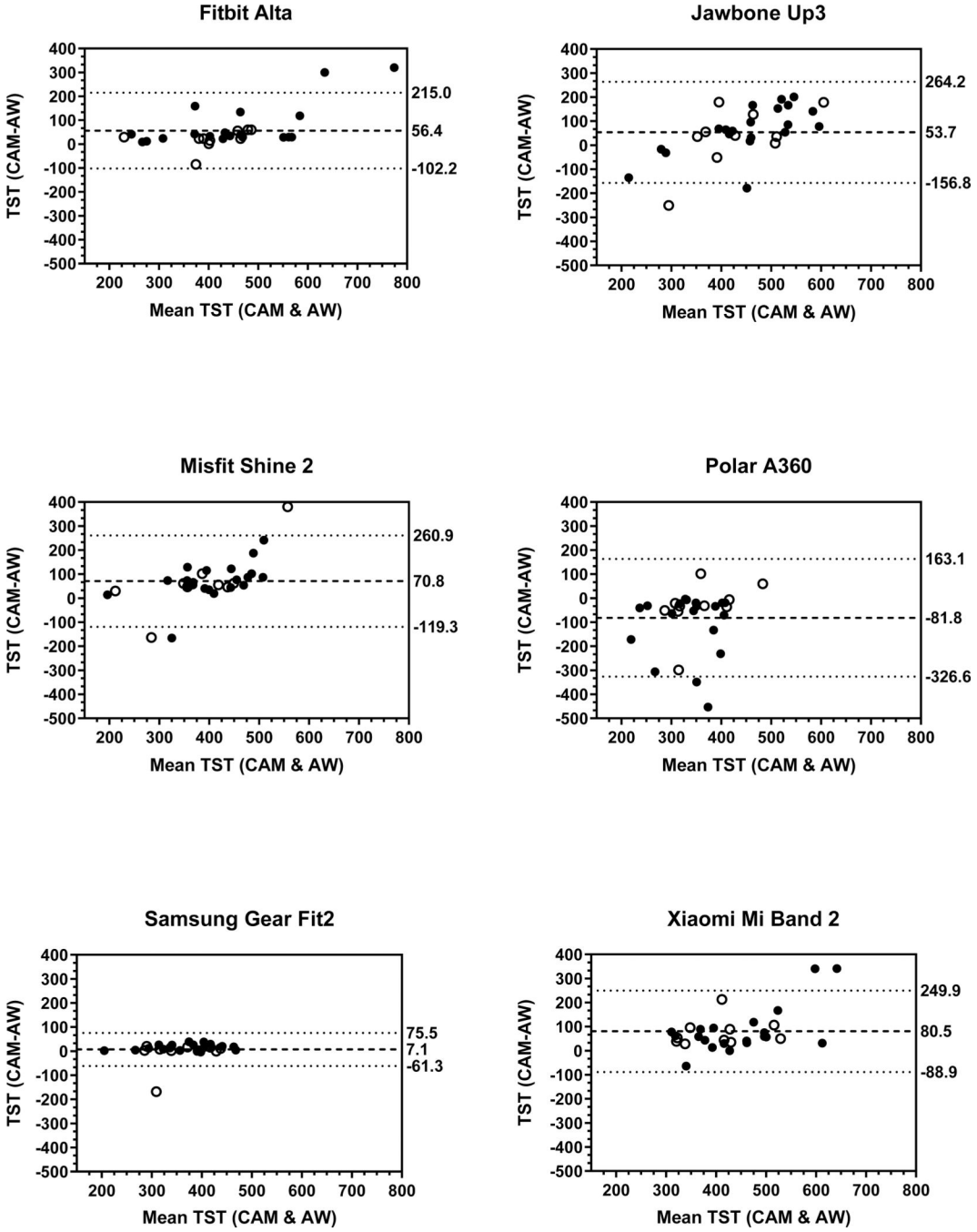


**Figure 2. Bland-Altman plots for unedited commercial activity monitor vs. edited Actiwatch total sleep time.**

Data are shown in minutes. X-axis indicates the mean total sleep time of the commercial activity monitor and Actiwatch (CAM & AW). Y-axis indicates the difference in total sleep time between the commercial activity monitor and Actiwatch (CAM-AW). Closed circles=good sleepers (i.e., Pittsburgh Sleep Quality Index [PSQI] global score ≤ 5); open circles=poor sleepers (i.e., PSQI global score > 5). Abbreviations: AW=Actiwatch; CAM=commercial activity monitor; TST=total sleep time.

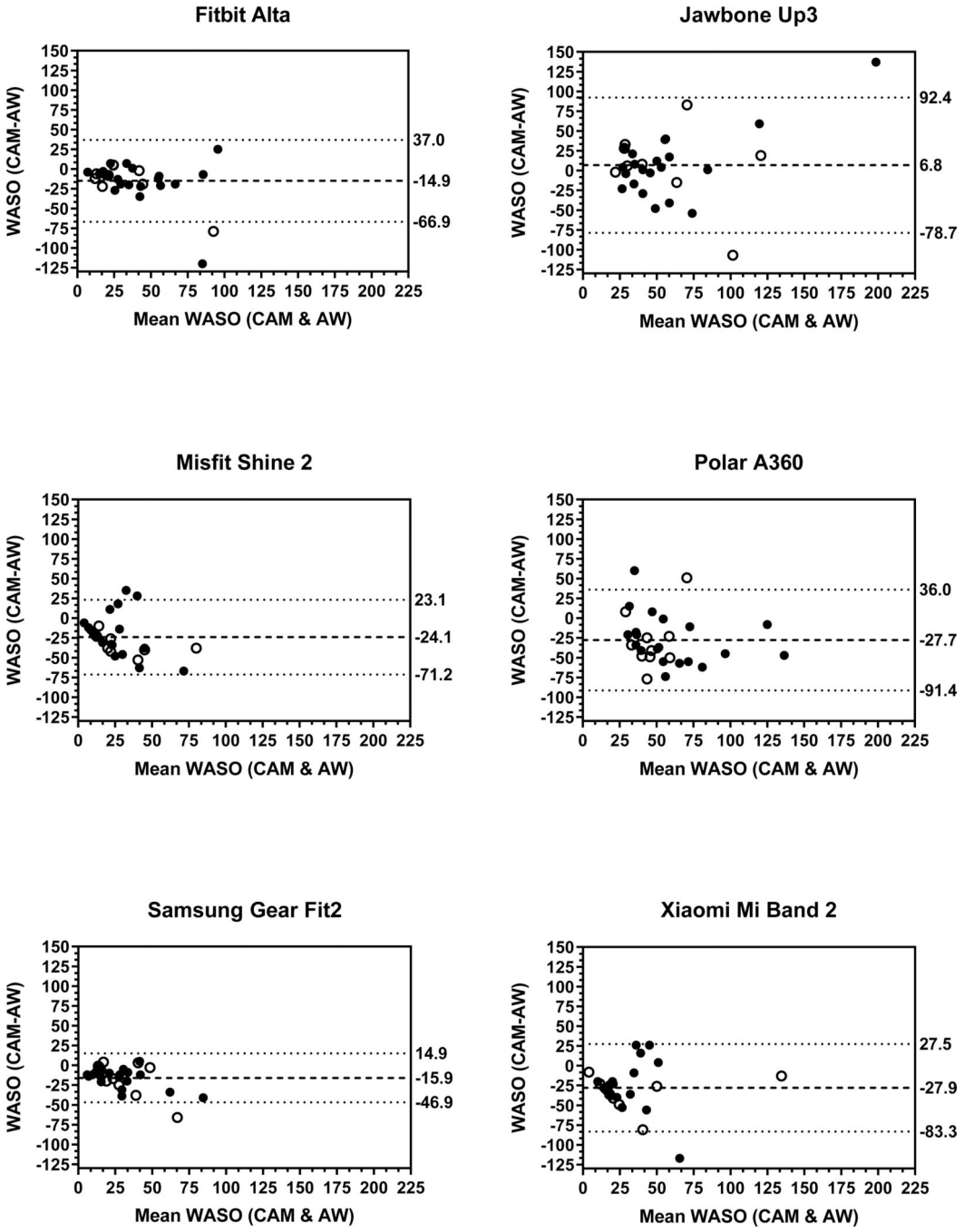


**Figure 3. Bland-Altman plots for unedited commercial activity monitor vs. edited Actiwatch wake after sleep onset.**

Data are shown in minutes. X-axis indicates the mean wake after sleep onset of the commercial activity monitor and Actiwatch (CAM & AW). Y-axis indicates the difference in wake after sleep onset between the commercial activity monitor and actiwatch (CAM-AW). Closed circles=good sleepers (i.e., Pittsburgh Sleep Quality Index [PSQI] global score ≤ 5); open circles=poor sleepers (i.e., PSQI global score > 5). Abbreviations: AW=Actiwatch; CAM=commercial activity monitor; WASO=wake after sleep onset.

**Table 1.**

Device information.

| Manufacturer | Model | App/Software | TST Variable | WASO Variable | Additional Variables |
|---|---|---|---|---|---|
| *CAM Devices* | | | | | |
| Fitbit | Alta | 2.38.1 | Time Asleep | Awake/Restless | Mins to fall asleep, Times Awake, Times Restless, Bed/Wake Time, Mins to fall asleep |
| Jawbone | Up3 | 4.29.0.100 | Total Sleep | Awake for | Deep Sleep, Light Sleep, Fell Asleep, REM Sleep, Times Woke Up, Bed/Wake Time |
| Misfit | Shine 2 | 2.15.2 | Total Sleep | Awake | Restful, Light, Bed/Wake Time |
| Polar | A360 | 3.5.4 | Night Sleep | Restless Sleep | Restful Sleep |
| Samsung | Gear Fit2 | 1.6.17030904 | Actual Sleep Time | Restless | Total Time Slept, Efficiency, Motionless, Light, Restless, Bed/Wake Time |
| Xiaomi | Mi Band 2 | 3.0.4 | Total Sleep | Awake Time | Deep Sleep, Light Sleep, Bed/Wake Time, Sleep Score |
| *Research Device* | | | | | |
| Philips Respironics | Actiwatch Spectrum | 6.0.9 | Total Sleep Time | Wake After Sleep Onset | |

TST: total sleep time; WASO: wake after sleep onset.

**Table 2.**

Participant characteristics.

| Characteristic | Total (*N*=30) | Males (*n*=15) | Females (*n*=15) | Intra-Model Reliability Subsample (*n*=10) |
|---|---|---|---|---|
| Age (years), mean (SD) | 24.8 (4.1) | 25.0 (3.9) | 24.6 (4.4) | 24.2 (4.6) |
| White race, *n* (%) | 26 (87) | 13 (87) | 13 (87) | 10 (100) |
| College degree or more, *n* (%) | 24 (80) | 12 (80) | 12 (80) | 10 (100) |
| BMI ($kg/m^2$), mean (SD) | 25.7 (3.8) | 26.6 (2.9) | 24.9 (4.4) | 24.4 (4.4) |
| PSQI, mean (SD) | 4.9 (3.2) | 5.6 (3.3) | 4.1 (3.1) | 3.3 (2.4) |
| PSQI > 5, *n* (%) | 10 (33) | 7 (23) | 3 (10) | 2 (20) |
| ISI, mean (SD) | 4.2 (3.5) | 4.6 (3.5) | 3.9 (3.6) | 3.1 (4.0) |
| ESS, mean (SD) | 6.4 (3.3) | 7.3 (3.6) | 5.5 (2.7) | 4.8 (3.0) |

BMI: Body Mass Index; ESS: Epworth Sleepiness Scale; ISI: Insomnia Severity Index; PSQI: Pittsburgh Sleep Quality Index; SD: Standard Deviation.

**Table 3.**

Overall and sleep quality-stratified comparison of total sleep time between the unedited commercial activity monitor and edited Actiwatch.

| CAM | CAM TST | AW TST | Difference | P-value | ICC (95% CI) |
|---|---|---|---|---|---|
| Fitbit Alta | | | | | |
| *Overall* | 466.4 (146.7) | 410.1 (97.4) | 56.4 (80.9) | .001 | .84 (.51–.93) |
| *PSQI ≤ 5* | 491.4 (165.7) | 417.0 (109.3) | 74.4 (90.1) | .002 | .83 (.32–.94) |
| *PSQI > 5* | 416.7 (85.0) | 396.3 (70.8) | 20.4 (42.1) | .16 | .91 (.66–.98) |
| Jawbone Up3 | | | | | |
| *Overall* | 473.1 (137.6) | 419.4 (75.2) | 53.7 (107.4) | .01 | .65 (.26–.83) |
| *PSQI ≤ 5* | 484.9 (139.3) | 422.7 (77.8) | 62.2 (99.9) | .01 | .69 (.21–.88) |
| *PSQI > 5* | 449.7 (138.3) | 413.9 (73.3) | 35.8 (124.8) | .39 | .52 (−.82–.89) |
| Misfit Shine 2 | | | | | |
| *Overall* | 434.2 (119.5) | 363.4 (65.3) | 70.8 (97.0) | < .001 | .56 (.03–.80) |
| *PSQI ≤ 5* | 442.3 (104.9) | 373.6 (64.3) | 68.7 (77.6) | .001 | .64 (−.05–.87) |
| *PSQI > 5* | 417.9 (149.5) | 343.1 (65.6) | 74.8 (132.6) | .11 | .46 (−.62–.85) |
| Polar A360 | | | | | |
| *Overall* | 305.1 (93.9) | 386.9 (78.2) | −81.8 (124.9) | .001 | .00 (−.70–.40) |
| *PSQI ≤ 5* | 300.1 (80.5) | 390.5 (86.3) | −90.4 (122.3) | .004 | .00 (−.83–.45) |
| *PSQI > 5* | 315.2 (121.0) | 379.7 (62.6) | −64.5 (134.9) | .17 | .03 (−1.0–.74) |
| Samsung Gear Fit2 | | | | | |
| *Overall* | 366.8 (68.1) | 359.7 (61.2) | 7.1 (34.9) | .27 | .92 (.84–.96) |
| *PSQI ≤ 5* | 377.4 (64.8) | 363.5 (62.9) | 13.9 (12.3) | <.001 | .97 (.74–.00) |
| *PSQI > 5* | 345.6 (72.9) | 352.2 (60.0) | −6.6 (57.4) | .73 | .78 (.11–.95) |
| Xiaomi Mi Band 2 | | | | | |
| *Overall* | 474.8 (119.5) | 394.3 (76.2) | 80.5 (86.4) | < .001 | .65 (−.01–.86) |
| *PSQI ≤ 5* | 490.5 (132.7) | 407.3 (75.2) | 83.2 (99.6) | .001 | .62 (.03–.86) |
| *PSQI > 5* | 443.5 (84.7) | 368.2 (74.9) | 75.3 (55.6) | .002 | .70 (.26–.93) |

Analyses for the overall sample are based on 30 participants for each CAM; n=20 for PSQI ≤ 5 and n=10 for PSQI > 5. AW: Actiwatch; CAM: commercial activity monitor; CI: confidence interval; ICC: intra-class correlation; PSQI: Pittsburgh Sleep Quality Index; TST: total sleep time in minutes. CAM TST, AW TST, and Difference data are reported as mean (standard deviation).

**Table 4.**

Overall and sleep quality-stratified comparison of wake after sleep onset between the unedited commercial activity monitor and edited Actiwatch.

| CAM | CAM WASO | AW WASO | Difference | P-value | ICC (95% CI) |
|---|---|---|---|---|---|
| Fitbit Alta | | | | | |
| *Overall* | 31.4 (22.9) | 46.4 (32.9) | −14.9 (26.5) | .004 | .67 (.26–.85) |
| *PSQI ≤ 5* | 34.6 (24.8) | 50.8 (31.5) | −16.2 (28.0) | .02 | .63 (.11–.85) |
| *PSQI > 5* | 25.0 (14.6) | 37.5 (35.3) | −12.5 (24.5) | .14 | .71 (.00–.93) |
| Jawbone Up3 | | | | | |
| *Overall* | 60.1 (50.1) | 53.3 (35.6) | 6.8 (43.6) | .40 | .67 (.30–.84) |
| *PSQI ≤ 5* | 61.6 (56.9) | 55.4 (28.5) | 6.2 (42.4) | .52 | .72 (.29–.89) |
| *PSQI > 5* | 57.3 (35.1) | 49.2 (48.4) | 8.1 (48.3) | .61 | .53 (−1.0–.89) |
| Misfit Shine 2 | | | | | |
| *Overall* | 13.8 (17.9) | 37.8 (24.6) | −24.1 (24.1) | < .001 | .38 (−.19–.70) |
| *PSQI ≤ 5* | 10.9 (14.9) | 35.8 (23.9) | −24.9 (24.0) | < .001 | .27 (−.27–.65) |
| *PSQI > 5* | 19.5 (22.4) | 42.0 (26.9) | −22.5 (25.5) | .02 | .51 (−.34–.86) |
| Polar A360 | | | | | |
| *Overall* | 69.7 (32.8) | 42.0 (28.4) | 27.7 (32.5) | < .001 | .48 (−.10–.76) |
| *PSQI ≤ 5* | 74.0 (37.9) | 46.8 (29.3) | 27.2 (31.7) | .001 | .60 (−.06–.85) |
| *PSQI > 5* | 61.3 (17.9) | 32.5 (25.2) | 28.8 (35.7) | .03 | .00 (−1.0–.52) |
| Samsung Gear Fit2 | | | | | |
| *Overall* | 21.0 (15.5) | 36.9 (23.6) | −15.9 (15.8) | < .001 | .69 (−.03–.88) |
| *PSQI ≤ 5* | 20.8 (16.1) | 36.1 (22.9) | −15.3 (13.1) | <.001 | .76 (−.10–.93) |
| *PSQI > 5* | 21.3 (15.0) | 38.6 (26.2) | −17.3 (20.9) | .03 | .58 (−.27–.89) |
| Xiaomi Mi Band 2 | | | | | |
| *Overall* | 15.6 (28.1) | 43.5 (28.8) | −27.9 (28.3) | < .001 | .51 (−.14–.79) |
| *PSQI ≤ 5* | 14.7 (20.5) | 40.7 (22.9) | −26.0 (31.5) | .002 | .00 (−.67–.45) |
| *PSQI > 5* | 17.5 (40.5) | 49.2 (38.7) | −31.7 (21.4) | .001 | .79 (−.21–.96) |

Analyses for the overall sample are based on 30 participants for each CAM; n=20 for PSQI ≤ 5 and n=10 for PSQI > 5. AW: Actiwatch; CAM: commercial activity monitor; CI: confidence interval; ICC: intra-class correlation; PSQI: Pittsburgh Sleep Quality Index; WASO: wake after sleep onset in minutes. CAM WASO, AW WASO, and Difference data are reported as mean (standard deviation).

**Table 5.**

Intra-model reliability of total sleep time.

| Device | Closer | Further | Difference | P-value | Device 1 | Device 2 | Difference | P-value | ICC (95% CI) |
|---|---|---|---|---|---|---|---|---|---|
| AW Spectrum | 359.9 (82.2) | 365.3 (86.0) | −5.4 (24.6) | .51 | 362.5 (86.2) | 362.7 (82.1) | −0.2 (25.3) | .98 | .98 (.92–.99) |
| Fitbit Alta | 414.4 (60.2) | 414.6 (58.6) | −0.2 (9.1) | .95 | 411.7 (60.5) | 417.3 (58.2) | −5.6 (6.9) | .03 | .99 (.98–1.00) |
| Jawbone Up3 | 428.0 (100.6) | 402.0 (95.0) | 26.0 (67.4) | .25 | 427.7 (104.9) | 402.3 (90.4) | 25.4 (67.6) | .27 | .86 (.48–.97) |
| Misfit Shine 2 | 457.2 (110.3) | 459.0 (111.2) | −1.8 (10.2) | .59 | 455.8 (109.3) | 460.4 (112.0) | −4.6 (9.1) | .15 | .99 (.99–1.00) |
| Polar A360 | 344.1 (115.4) | 324.4 (125.1) | 19.7 (61.4) | .34 | 343.5 (115.5) | 325.0 (125.2) | 18.5 (61.8) | .37 | .93 (.74–.98) |
| Samsung Gear Fit2 | 422.8 (100.4) | 435.7 (134.4) | −12.9 (70.7) | .58 | 439.5 (134.2) | 419.0 (100.1) | 20.5 (68.7) | .37 | .91 (.65–.98) |
| Xiaomi Mi Band 2 | 451.4 (113.6) | 440.4 (114.5) | 11.0 (25.1) | .20 | 451.3 (114.1) | 440.5 (113.9) | 10.8 (25.2) | .21 | .99 (.95–1.00) |

AW: Actiwatch; CI: confidence interval; Closer: device placed on the wrist distal to the body; Further: device placed on the wrist proximal to the body; ICC: intra-class correlation; Device 1: devices used by male participants for the primary study procedures; Device 2: devices used by female participants for the primary study procedures; TST: total sleep time in minutes. All Closer, Further, Device 1, Device 2, and Difference data are reported as mean (standard deviation).

**Table 6.**

Intra-model reliability of Wake After Sleep Onset.

| Device | Closer | Further | Difference | P-value | Device 1 | Device 2 | Difference | P-value | ICC (95% CI) |
|---|---|---|---|---|---|---|---|---|---|
| AW Spectrum | 33.9 (22.9) | 37.7 (25.5) | −3.8 (8.0) | .17 | 37.5 (26.3) | 34.1 (22.0) | 3.4 (8.2) | .22 | .97 (.88–.99) |
| Fitbit Alta | 25.2 (15.0) | 27.0 (19.2) | −1.8 (5.1) | .29 | 27.5 (18.8) | 24.7 (15.5) | 2.8 (4.6) | .08 | .98 (.91–.99) |
| Jawbone Up3 | 60.6 (34.1) | 57.8 (29.9) | 2.8 (27.3) | .75 | 65.5 (32.5) | 53.9 (31.5) | 11.6 (24.8) | .17 | .80 (.13–.95) |
| Misfit Shine 2 | 14.8 (17.1) | 15.4 (16.8) | −0.6 (7.5) | .81 | 13.5 (15.5) | 16.7 (18.1) | −3.2 (6.7) | .17 | .95 (.81–.99) |
| Polar A360 | 75.0 (40.2) | 72.7 (43.2) | 2.3 (8.1) | .39 | 75.0 (41.1) | 72.7 (42.4) | 2.3 (8.1) | .39 | .99 (.97–1.00) |
| Samsung Gear Fit2 | 27.6 (24.8) | 25.1 (18.7) | 2.5 (12.4) | .54 | 31.4 (22.6) | 21.3 (20.1) | 10.1 (6.9) | .001 | .92 (.68–.98) |
| Xiaomi Mi Band 2 | 9.7 (19.0) | 7.4 (14.9) | 2.3 (6.9) | .32 | 9.7 (19.0) | 7.4 (14.9) | 2.3 (6.9) | .32 | .96 (.84–.99) |

AW: Actiwatch; CI: confidence interval; Closer: device placed on the wrist distal to the body; Further: device placed on the wrist proximal to the body; ICC: intra-class correlation; Device 1: devices used by male participants for the primary study procedures; Device 2: devices used by female participants for the primary study procedures; WASO: wake after sleep onset in minutes. All Closer, Further, Device 1, Device 2, and Difference data are reported as mean (standard deviation).