SUPPLEMENTAL MATERIALS

Performance of Seven Consumer Sleep-Tracking Devices Compared with Polysomnography

Evan D. Chinoy[1,2]; Joseph A. Cuellar[1,2]; Kirbie E. Huwa[1,2]; Jason T. Jameson[1,2]; Catherine H. Watson[1,3]; Sara C. Bessman[1,4]; Dale A. Hirsch[1]; Adam D. Cooper[1,3]; Sean P.A. Drummond[5]; Rachel R. Markwald[1]

[1] Sleep, Tactical Efficiency, and Endurance Laboratory, Warfighter Performance Department, Naval Health Research Center, San Diego, CA; [2] Leidos, Inc., San Diego, CA; [3] Innovative Employee Solutions, San Diego, CA; [4] Eagle Applied Sciences, San Diego, CA; [5] Turner Institute for Brain and Mental Health, Monash University, Melbourne, Victoria, Australia

Corresponding Authors:

1. Rachel R. Markwald; Sleep, Tactical Efficiency, and Endurance Laboratory, Warfighter Performance Department, Naval Health Research Center, 140 Sylvester Road, San Diego, CA 92106; rachel.r.markwald.civ@mail.mil
2. Evan D. Chinoy; Sleep, Tactical Efficiency, and Endurance Laboratory, Warfighter Performance Department, Naval Health Research Center, 140 Sylvester Road, San Diego, CA 92106; evan.d.chinoy.ctr@mail.mil

Table S1: Sleep Devices Used by Each Participant

Partici-pant #	Fatigue Science Readiband	Fitbit Alta HR	Garmin Fenix 5S	Garmin Vivosmart 3	EarlySense Live	ResMed S+	SleepScore Max
1					X	X	
2					X	X	
3					X	X	
4		X			X	X	
5		X			X	X	
6		X			X	X	
7		X			X	X	
8		X			X	X	
9		X			X	X	
10		X			X	X	
11		X			X	X	
12		X			X	X	
13		X			X	X	
14		X			X	X	
15		X			X	X	
16		X			X	X	
17		X			X	X	
18		X			X	X	
19		X			X	X	
20	X	X		X			X
21	X	X		X			X
22	X	X		X			X
23	X	X		X			X
24	X		X	X			X
25	X		X	X			X
26	X		X	X			X
27	X		X	X			X
28	X		X	X			X
29	X		X	X			X
30	X		X	X			X
31	X		X	X			X
32	X		X	X			X
33	X		X	X			X
34	X		X	X			X

Table depicting the sequence of specific consumer sleep-tracking devices used by each participant in the study.

Table S2. EBE Contingency Tables: Sleep versus Wake

		PSG	
		Sleep	Wake
Actiwatch	Sleep	81439	7693
	Wake	2823	4872
Fatigue Science Readiband	Sleep	16541	1465
	Wake	966	1188
Fitbit Alta HR	Sleep	45628	2789
	Wake	2395	3328
Garmin Fenix 5S	Sleep	12086	1621
	Wake	124	358
Garmin Vivosmart 3	Sleep	18022	2291
	Wake	186	541
EarlySense Live	Sleep	38110	2957
	Wake	1633	2638
ResMed S+	Sleep	39846	2956
	Wake	3053	3124
SleepScore Max	Sleep	33949	2901
	Wake	2295	2887

Contingency tables showing epoch-by-epoch (EBE) agreement of individual sleep and wake epochs with polysomnography (PSG) for all devices. Number of epochs are the total available for analysis across all participants and nights. Epoch durations were 30 sec for Actiwatch, Fitbit Alta HR, EarlySense Live, ResMed S+, and SleepScore Max. Epoch durations were 60 sec for Fatigue Science Readiband, Garmin Fenix 5S, and Garmin Vivosmart 3.

Table S3. EBE Contingency Tables: Sleep Stages

		PSG			
		Wake	Light	Deep	REM
Fitbit Alta HR	Wake	3328	1529	253	613
	Light	2168	20559	3891	2872
	Deep	55	3167	4817	214
	REM	566	1745	94	8269
Garmin Fenix 5S	Wake	358	110	1	9
	Light	1060	4637	789	1521
	Deep	122	697	1111	117
	REM	439	1400	72	1652
Garmin Vivosmart 3	Wake	541	178	1	5
	Light	1512	7211	1266	2012
	Deep	221	962	1720	214
	REM	558	1880	86	2579
EarlySense Live	Wake	2638	1278	134	221
	Light	1950	12747	2043	3148
	Deep	185	5599	5089	195
	REM	822	2747	241	6301
ResMed S+	Wake	3124	1969	428	656
	Light	2337	16204	2874	4443
	Deep	53	4639	4787	170
	REM	566	1454	76	5199
SleepScore Max	Wake	2887	1531	319	445
	Light	2325	14001	2046	4230
	Deep	65	4054	3474	216
	REM	511	1109	82	4737

Contingency tables showing epoch-by-epoch (EBE) agreement of individual wake and sleep stage (Light, Deep, and REM sleep) epochs with polysomnography (PSG) for all the devices that output sleep stage classifications. Number of epochs are the total available for analysis across all participants and nights. Epoch durations were 30 sec for Actiwatch, Fitbit Alta HR, EarlySense Live, ResMed S+, and SleepScore Max. Epoch durations were 60 sec for Fatigue Science Readiband, Garmin Fenix 5S, and Garmin Vivosmart 3.

Table S4. Sleep Summary: Latency to Persistent Sleep (LPS)

Device	n	PSG Mean ± SD	Device Mean ± SD	Bias	Lower Limit	Upper Limit	t (p)	Effect Size	R^2 (p)
Actiwatch	102	14.7 ± 17.7	7.7 ± 14.0	-7.0	-30.3	16.4	-6.0 ***(<0.001)***	-0.44	0.12 ***(<0.001)***
Fatigue Science Readiband	42	13.9 ± 17.9	9.0 ± 10.6	-4.9	-22.3	12.5	-3.6 ***(0.001)***	-0.33	0.74 ***(<0.001)***
Fitbit Alta HR	57	14.4 ± 16.6	12.2 ± 13.8	-2.2	-29.3	24.9	-1.2 (0.223)	-0.14	0.05 (0.086)
Garmin Fenix 5S	30	14.0 ± 20.5	11.6 ± 14.1	-2.4	-42.6	37.8	-0.7 (0.518)	-0.14	0.15 ***(0.037)***
Garmin Vivosmart 3	44	13.5 ± 17.6	9.8 ± 8.4	-3.7	-31.5	24.1	-1.8 (0.083)	-0.27	0.53 ***(<0.001)***
Earlysense Live	55	15.5 ± 18.2	15.3 ± 14.3	-0.1	-31.3	31.1	-0.1 (0.949)	-0.01	0.08 ***(0.038)***
ResMed S+	54	16.3 ± 18.4	14.8 ± 16.8	-1.6	-26.1	23.0	-0.9 (0.356)	-0.09	0.02 (0.319)
SleepScore Max	44	13.6 ± 17.6	15.6 ± 14.3	2.1	-20.2	24.3	1.2 (0.227)	0.13	0.10 ***(0.037)***

Summary results for minutes of LPS, for all devices versus polysomnography (PSG). LPS was calculated as the time from bedtime to the first epoch of ten consecutive minutes of scored sleep. All nights with available LPS data for both the device and PSG are included, with the total number of nights (n) indicated in each row. Means and standard deviations (SD) are shown for PSG and each device. Bias represents the difference between the PSG and device means, with positive and negative bias values indicating the device showed an overestimation or underestimation compared with PSG, respectively. Lower and upper limits of agreement represent two SDs from the bias. Statistical significance between each device and PSG was assessed with paired t-tests and corresponding p-values. Effect sizes (Hedges' g) and proportional biases (R^2) with corresponding p-values are also shown. P-values at the $p<0.05$ level were considered statistically significant and are shown in bold and italic.

Table S5. Sleep Summary: WASO (from LPS)

Device	n	PSG Mean ± SD	Device Mean ± SD	Bias	Lower Limit	Upper Limit	t (p)	Effect Size	R^2 (p)
Actiwatch	98	50.5 ± 39.9	34.4 ± 19.2	-16.1	-79.8	47.6	-5.0 ***(<0.001)***	-0.51	0.50 ***(<0.001)***
Fatigue Science Readiband	41	52.1 ± 47.9	41.6 ± 52.3	-10.5	-115.9	94.9	-1.3 (0.208)	-0.21	0.01 (0.543)
Fitbit Alta HR	49	44.6 ± 30.9	42.7 ± 19.4	-1.9	-41.5	37.7	-0.7 (0.504)	-0.07	0.38 ***(<0.001)***
Garmin Fenix 5S	29	54.8 ± 53.6	6.8 ± 11.3	-48.0	-145.6	49.6	-5.3 ***(<0.001)***	-1.22	0.87 ***(<0.001)***
Garmin Vivosmart 3	43	53.7 ± 48.5	7.3 ± 12.5	-46.4	-129.0	36.2	-7.4 ***(<0.001)***	-1.30	0.86 ***(<0.001)***
Earlysense Live	51	47.2 ± 32.5	30.9 ± 25.2	-16.3	-69.9	37.3	-4.3 ***(<0.001)***	-0.56	0.09 ***(0.031)***
ResMed S+	51	46.2 ± 31.3	44.6 ± 34.1	-1.6	-66.3	63.2	-0.3 (0.729)	-0.05	0.01 (0.487)
SleepScore Max	42	54.8 ± 49.1	43.4 ± 33.1	-11.4	-77.9	55.1	-2.2 ***(0.032)***	-0.27	0.27 ***(<0.001)***

Summary results for total minutes of wake after sleep onset (WASO) from latency to persistent sleep (LPS), for all devices versus polysomnography (PSG). See Table S4 caption for additional table details.

Table S6. Sleep Summary: REM Latency

Device	n	PSG Mean± SD	Device Mean ± SD	Bias	Lower Limit	Upper Limit	t (p)	Effect Size	R^2 (p)
Fitbit Alta HR	57	93.1 ± 36.1	127.2 ± 55.5	34.1	-81.9	150.0	4.4 ***(<0.001)***	0.72	0.17 ***(0.001)***
Garmin Fenix 5S	30	82.2 ± 39.4	112.0 ± 45.1	29.8	-83.3	142.9	2.9 ***(0.007)***	0.69	0.02 (0.472)
Garmin Vivosmart 3	44	85.8 ± 35.6	110.3 ± 49.0	24.5	-94.8	143.7	2.7 ***(0.009)***	0.57	0.10 ***(0.041)***
Earlysense Live	54	93.3 ± 39.0	78.8 ± 45.6	-14.5	-124.0	95.0	-1.9 (0.057)	-0.34	0.02 (0.261)
ResMed S+	54	94.0 ± 36.0	116.3 ± 48.0	22.2	-87.8	132.2	3.0 ***(0.004)***	0.52	0.08 ***(0.037)***
SleepScore Max	44	86.5 ± 35.9	113.2 ± 42.3	26.8	-63.2	116.7	3.9 ***(<0.001)***	0.68	0.03 (0.264)

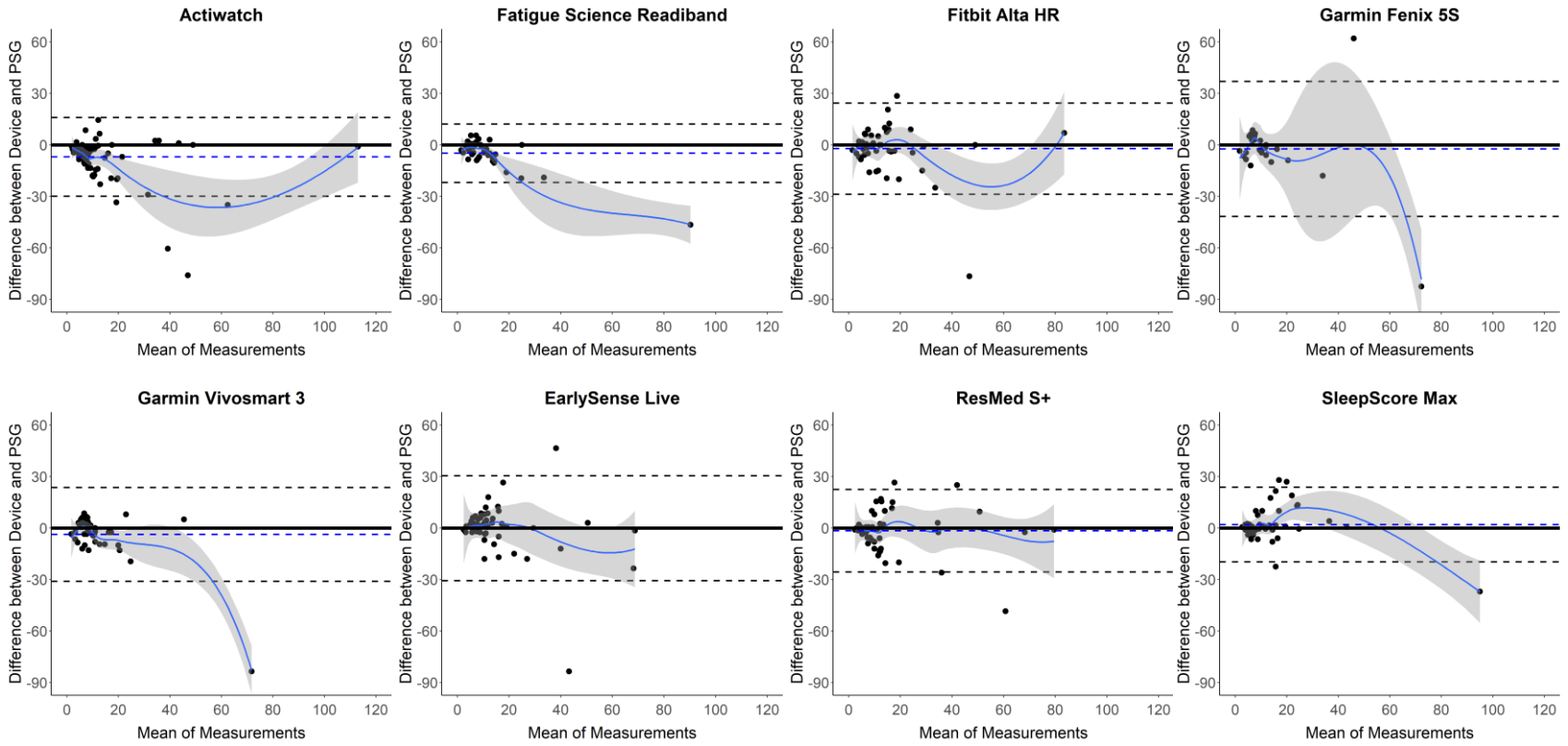
Summary results for minutes of rapid eye movement (REM) latency, for the devices versus polysomnography (PSG). REM latency was defined as the time from bedtime to the first scored epoch of REM sleep. Results are shown for all devices that output sleep stage classifications. See Table S4 caption for additional table details.

Table S7. EBE Agreement: PSG Differences from Device-Scored Epochs:

Device	Wake Epochs			Light Sleep Epochs			Deep Sleep Epochs			REM Sleep Epochs		
	Light	Deep	REM	Wake	Deep	REM	Wake	Light	REM	Wake	Light	Deep
Fitbit Alta HR	0.27	0.04	0.11	0.07	0.13	0.10	0.01	0.38	0.03	0.05	0.16	0.01
Garmin Fenix 5S	0.23	0.00	0.02	0.13	0.10	0.19	0.06	0.34	0.06	0.12	0.39	0.02
Garmin Vivosmart 3	0.25	0.00	0.01	0.13	0.11	0.17	0.07	0.31	0.07	0.11	0.37	0.02
EarlySense Live	0.30	0.03	0.05	0.10	0.10	0.16	0.02	0.51	0.02	0.08	0.27	0.02
ResMed S+	0.32	0.07	0.11	0.09	0.11	0.17	0.01	0.48	0.02	0.08	0.20	0.01
SleepScore Max	0.30	0.06	0.09	0.10	0.09	0.19	0.01	0.52	0.03	0.08	0.17	0.01

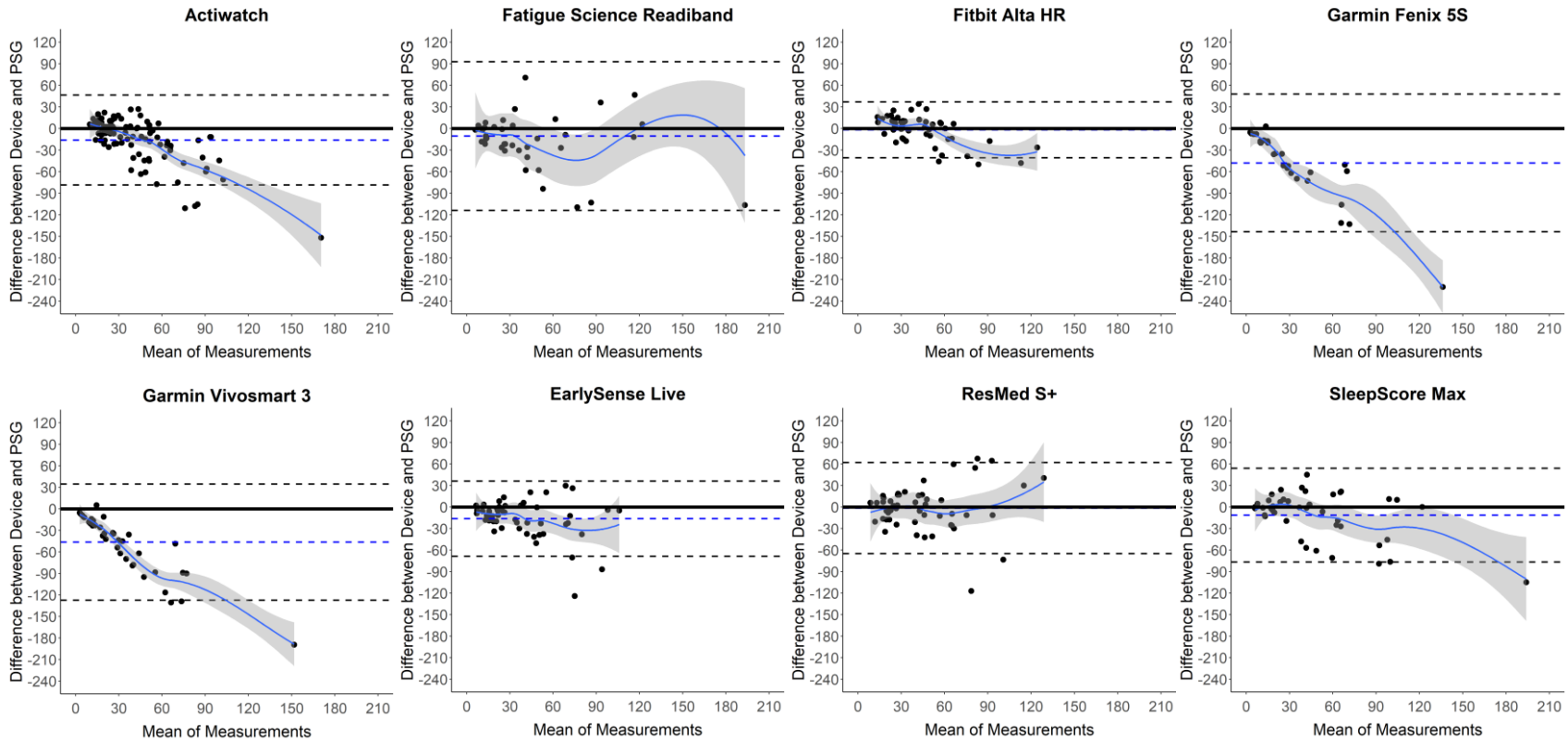
Proportions for epoch-by-epoch (EBE) differences in polysomnography (PSG) sleep stage classifications from the device-scored epochs. Device-scored classifications are the larger column categories, with the three possible PSG-scored differences under each category. Results are shown for all devices that output sleep stage classifications.

Figure S1. Bland-Altman Plots: Latency to Persistent Sleep (LPS)



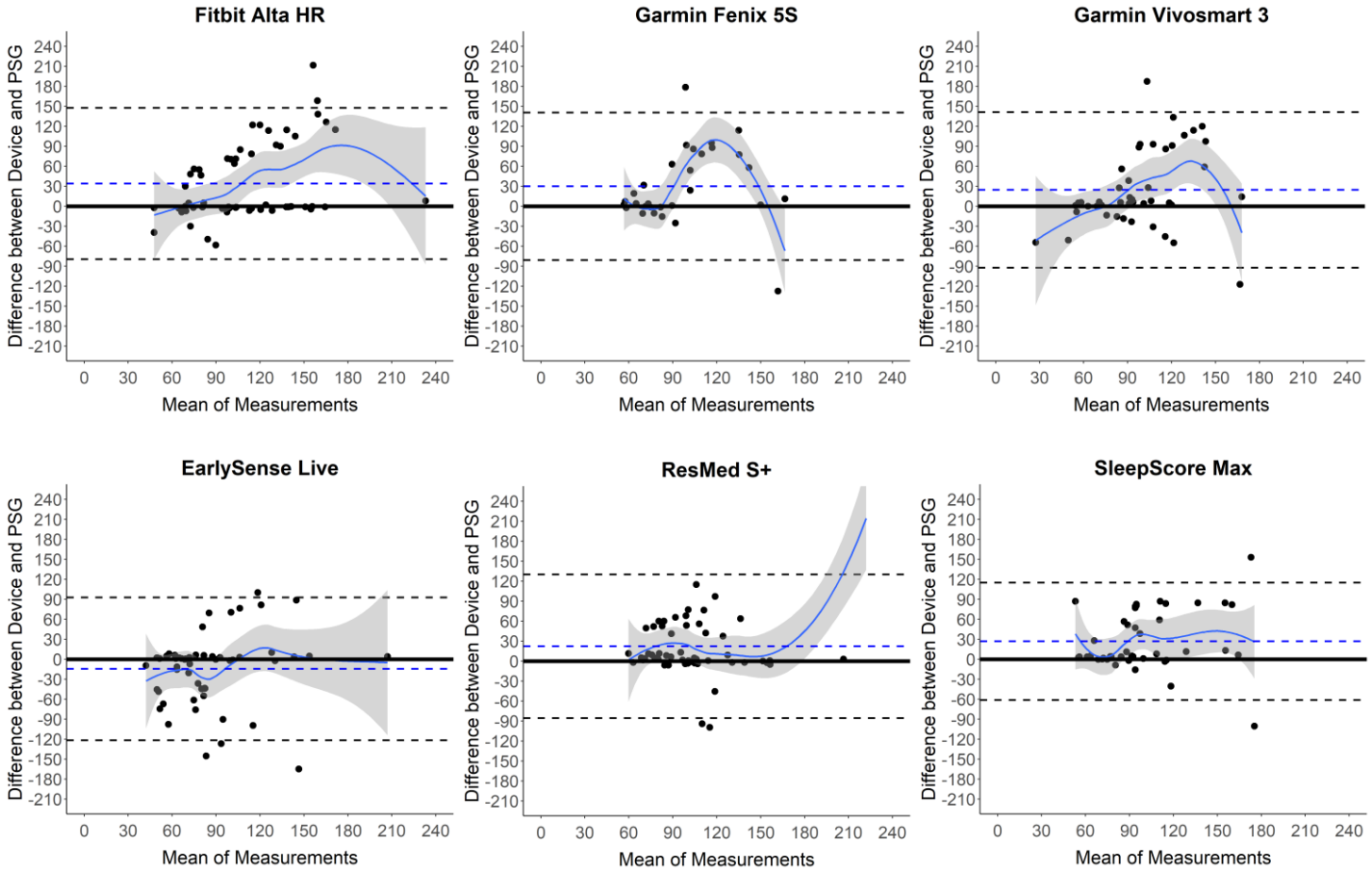
Bland-Altman plots depicting the mean bias (blue dashed line) and upper and lower limits of agreement (two standard deviations from the bias; black dashed lines) for the minutes of LPS (defined as the time from bedtime to the first epoch of ten consecutive minutes of scored sleep) for the devices compared with polysomnography (PSG). Black circles are individual nights. Solid blue curves represent the best-fit of data, with surrounding gray shaded regions representing 95% confidence bands. The solid black line at zero represents no difference, with positive and negative y-axis values indicating an overestimation or underestimation, respectively, compared with PSG.

Figure S2. Bland-Altman Plots: WASO (from LPS)



Bland-Altman plots depicting the minutes of wake after sleep onset (WASO) from latency to persistent sleep (LPS) for the devices compared with polysomnography (PSG). See Figure S1 caption for additional details on interpretation of Bland-Altman plots.

Figure S3. Bland-Altman Plots: REM Latency



Bland-Altman plots depicting the minutes of rapid eye movement (REM) latency (defined as the time from bedtime to the first epoch scored as REM) for the devices compared with polysomnography (PSG). Only devices that output data on sleep stages are depicted. See Figure S1 caption for additional details on interpretation of Bland-Altman plots.

SUPPLEMENTAL METHODS:

Supplemental Information on the Experimental Sleep Disruption Protocol:

On the experimental sleep disruption night (either Lab Visit 2 or 3, which was randomized and counterbalanced), participants were awoken by auditory tones for a brief scheduled period of every hour during the 8-hour sleep episode. The awakening periods (seven in total) occurred at exactly 1-hour intervals after bedtime and lasted for either 5-min or 10-min durations (also randomized and counterbalanced). During each awakening period, auditory tones were played by research staff using a Bluetooth smart speaker (Bose Soundlink Mini II; Bose Corp.; Framingham, MA, USA) placed on a bedside table in the research bedroom. The tones were played to keep the participant awake for as much of the scheduled awakening period duration as possible. Tones lasted 5 sec each and were played again (as needed) if the participant fell back asleep before the end of the scheduled awakening period. Research staff verified sleep and wake states by monitoring the live polysomnography (PSG) recording.

Supplemental Information on Consumer Sleep-Tracking Devices and Testing Procedures:

Due to physical and practical constraints as well as the availability of new consumer sleep-tracking device models, no individual participant used all seven devices that were tested in the overall study. Therefore, each participant was tested with a subset of the devices, and the individual devices included in the subsets were used over consecutive participants within different phases of the overall study (see Table S1). Following are the inclusive dates each device was tested: Fatigue Science Readiband: September 2018 – September 2019, Fitbit Alta HR: November 2017 – August 2018, Garmin Fenix 5S: December 2018 – September 2019, Garmin Vivosmart 3: September 2018 – September 2019, EarlySense Live: August 2017 –

August 2018, ResMed S+: August 2017 – August 2018, and SleepScore Max: September 2018 – September 2019.

Device apps used to collect sleep-tracking data were updated to the current app version at the time of testing, according to the dates listed above. App updates allow for optimal device functioning, syncing, and data export, therefore any app updates occurring over those dates were downloaded to the tablets that were used for device data collection. To our knowledge, during the phases of testing with each device there were no updates that would have impacted the performance of the device sleep-tracking algorithms. Occasional app updates may have affected the visualization of sleep-tracking data within the app or other functionality, but none of the stated app updates by the device companies indicated any changes to the algorithms or calculations of sleep-tracking metrics during the respective device testing phases. Known app versions over each testing period for each device were as follows: Fatigue Science Readiband: 1.4.19 – 2.0.14, Fitbit Alta HR: 2.61 – 2.89, Garmin Fenix 5S: 4.13.2 – 4.25.2, Garmin Vivosmart 3: 4.13.2 – 4.25.2, EarlySense Live: unknown, ResMed S+: 1.2.2 – 1.3.0, and SleepScore Max: 1.0 – 1.4.1.

The physical positions of the non-wearable devices were set according to the device company's instructions and were held constant across participants. The exact positions of non-wearable devices on the bedside table or under the mattress were measured and marked by research staff, then confirmed again to be in the correct positions at the start of each lab visit. For the wearable devices, the wrist position for each was held constant across the entire testing phase with that device, and the wrist used was determined by the hand dominance of each participant. The subsets of wrist devices tested concurrently could not all comfortably fit on one wrist (and there may be effects on the fidelity of the accelerometer and/or heart rate data if worn too high

up from the wrist), so a maximum of up to two devices were used on a single wrist at once. The Actiwatch was worn by all participants and was always worn on the non-dominant wrist. The Fatigue Science Readiband was worn on the dominant wrist, and the Fitbit Alta HR was worn on the non-dominant wrist. The only devices that were counterbalanced were the two Garmin devices. The Garmin Fenix 5S was added to the testing protocol after the Garmin Vivosmart 3 (and at the point both Garmin devices were used concurrently), so those two devices were tested on separate wrists and between participants were counterbalanced between dominant and non-dominant wrists.

Supplemental Information on PSG Recording Issues and Missing Data:

The PSG recording was successfully started at bedtime on all nights, but on four total nights the PSG either stopped recording or contained too much signal artifact in order to score sleep, for portions either in the middle and/or the very end of the night. Therefore, the PSG sleep latency measures (sleep onset latency [SOL], latency to persistent sleep [LPS], and REM latency) – which only require PSG data from the beginning of a sleep episode – were available for analysis on all nights. However, the other summary measures (total sleep time [TST], sleep efficiency [SE], wake after sleep onset [WASO], Light, Deep, and REM sleep) – which measure aspects of sleep and wake over a whole sleep episode – would not be valid with partial PSG data loss, and thus were not included for the summary analyses on nights when the PSG did not successfully record the complete 8-hour sleep episode. Although, any individual PSG epochs that were valid and available from any time on partial nights were still included in the epoch-by-epoch (EBE) analyses.

Supplemental Information on Device Recording Issues and Missing Data:

No issues occurred regarding the recording of Actiwatch data, and thus actigraphy data on all nights were available for all analyses. However, two types of issues occurred that led to occasional data loss from the consumer devices: 1) the device did not successfully record sleep-tracking data either for the whole night or for just part of the night, or 2) the timing of device epochs was misaligned from clock time (and thus was also misaligned from the PSG epochs).

For the first type of device data loss, whenever there were any valid parts of the device recording available, then those data would still be included in the final analyses (i.e., only using sleep latency measures for summary analyses if the beginning of the night was still available, and still using any available valid epochs for EBE analyses). The following are the total number of nights where: 1) the device was used for testing, 2) the device had a full missing night of sleep-tracking data, and 3) the device had a partial night of sleep-tracking data; respectively, for each device: Fatigue Science Readiband: 45, 3, 0; Fitbit Alta HR: 60, 2, 8; Garmin Fenix 5S: 33, 3, 7; Garmin Vivosmart 3: 45, 1, 3; EarlySense Live: 57, 2, 3; ResMed S+: 57, 3, 0; and SleepScore Max: 45, 2, 1. It should be noted for the Fatigue Science Readiband that the 3 full missing nights of sleep-tracking data resulted from the company accidentally deleting one of the participant user accounts right at the end of their study (before the data was exported), and thus all 3 nights for that participant were missing from all the Fatigue Science Readiband analyses but no other data loss with that device occurred. Additionally, there was one night of testing where the EarlySense Live did not pick up any REM sleep (however, the other stages were detected) and therefore that night was included in the summary analysis of total REM sleep minutes (as zero) but that night was not included in the analysis of REM latency because there was no valid REM latency that could be quantified.

For the second type of device data loss, it was occasionally found that the exported EBE data were not closely aligned in time with the start of a clock-minute (despite the device being carefully started/synced by research staff members to prevent this issue). However, this issue only affected the non-wearable devices because the sleep-tracking recording function of those devices had to be manually started by either pressing a button on the tablet app (for ResMed S+ and SleepScore Max) or would start recording whenever the Bluetooth connection between the tablet and the device was enabled (for EarlySense Live). The procedure followed by research staff for nights with the EarlySense Live was therefore to only turn on the tablet's Bluetooth signal at the start of a clock-minute. However, it is still possible that a time delay occurred between when the tablet's Bluetooth signal was turned on and when the device eventually connected to the tablet to begin the sleep recording. Aligning the exact start of the recording to the clock was especially critical for the EBE analyses, because the epoch start time of the first epoch would then dictate the timing of all subsequent epochs (in 30-sec epoch increments) for the remainder of the sleep episode. If the epoch start time for a device was misaligned by more than one-third of the epoch duration, then that device was not used for the EBE analysis for that night. Specifically, for the 30-sec epochs output by all the non-wearable devices where this was an issue, a device's EBE data were not used for that night if epochs were misaligned by more than 10 sec from the top of the clock-minute (i.e., epoch start times occurring between 11-19 sec or 41-49 sec past the start of a minute were excluded). The total number of nights affected were as follows: EarlySense Live: 7, ResMed S+: 2, SleepScore Max: 0.

Lastly, all the wearable consumer devices used in the study perform automatic and passive tracking of sleep episodes, therefore the sleep recording may only begin when the device's algorithm "detects" that the user has attempted to start their sleep episode. Likewise,

these devices will automatically end the sleep recording when the device's algorithm "detects" that the user has ended their sleep episode. Therefore, the data available for analysis in this study was affected by the interaction of the device's algorithm and the participant's behavior and physiology at bedtime, wake time, and throughout the night. For example, if the participant was physically restless, moving around or having a relatively high heart rate at bedtime (or wake time) then it may have caused the device's sleep recording to begin (or end) at a slightly later (or earlier) time – thus resulting in a sleep recording for less than the 8-hour sleep episode time in bed (TIB), impacting the analysis. Sometimes the opposite issue occurred, whereas if the participant was too sedentary at bedtime or wake time, then the device's algorithm may have "detected" that they began their sleep episode sooner than the scheduled bedtime (or ended sleep later than wake time), and may have recorded sleep for longer than the actual 8-hour TIB.

Supplemental Information on Sleep Summary Analyses:

For the summary analyses, only nights with approximately 8 hours of total sleep-tracking recording data within the scheduled sleep episode for a given device were included. We selected a duration of <470 min (i.e., <7h, 50 min) as the insufficient total recording time threshold whereby individual nights for a device were removed from the final summary analyses. Further, if a device recorded outside the scheduled 8-hour sleep episode TIB (i.e., began sleep-tracking earlier than actual bedtime and/or continued after wake time) then only the EBE data within the actual scheduled TIB were used for the summary calculations. This allowed for a standard comparison of the consumer devices and actigraphy versus PSG, and for the calculation of a wider array of sleep summary measures than may be available from a given device's app alone (sleep device apps display only a select number of sleep summary measures for the user, and

those app measures are not always the same between different devices). Thus, we sought to compare the sleep/wake summary measures that are typically analyzed in sleep research studies.