“How Does Fitbit Measure Brainwaves”: A Qualitative Study into the Credibility of Sleep-tracking Technologies

ZILU LIANG, Kyoto University of Advanced Science (KUAS)

BERND PLODERER, Queensland University of Technology (QUT)

Consumer sleep-tracking devices provide an unobtrusive and affordable way to learn about personal sleep habits. Recent research focused primarily on the information provided by such devices, i.e., whether the information is accurate and meaningful to people. However, little is known about how people judge the credibility of such information, and how the functionality and the design may influence such judgements. Hence, the aim of this research was to examine how consumers assess the credibility of sleep-tracking devices. We conducted a qualitative study with 22 participants who tracked their sleep for 3 nights with three different devices: Fitbit Charge 2, Neuroon EEG, and SleepScope, a medical sleep monitor. Based on semi-structured interviews, we found that people assess the credibility of sleep-tracking devices based not only on the credibility of sleep data per se, but also on device functionality, interface design and physical appearance. People found it difficult to judge credibility, because of the complexities of sleep stages and micro-arousals (sleep fallacy) and the black boxed nature of devices (black box fallacy), and also because of the misalignment between objective sleep measures and subjective sleep quality. We discuss the significance of design and functionality on the credibility of personal health technologies and highlight design challenges and opportunities to enhance their credibility.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**.

Additional Key Words and Phrases: Sleep-tracking, Credibility, Wearable sensors, Consumer devices, Personal informatics, Quantified self

ACM Reference Format:

Zilu Liang and Bernd Ploderer. 2020. “How Does Fitbit Measure Brainwaves”: A Qualitative Study into the Credibility of Sleep-tracking Technologies. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 1, Article 17 (March 2020), 29 pages. <https://doi.org/10.1145/3380994>

1 INTRODUCTION

Consumer sleep-tracking devices provide an affordable and unobtrusive way for individual users to gain insights into their sleep patterns. Devices like Fitbit and Neuroon provide users with information about the sleep duration, awakenings during a night, and sleep stages (light sleep, deep sleep, and REM sleep). This information is produced based on data from various sensors: devices worn on the wrist (like Fitbit) often contain accelerometers and optical heart rate sensors, whereas devices worn as a sleep mask (like Neuroon) also contain electroencephalography to record electrical activity of the brain. In principle, electroencephalography can provide more accurate insight into sleep stages, but the downside is that sensors worn on the head can be more obtrusive than wrist-worn sensors and hence interfere with sleep [77]. In other words, in choosing a consumer sleep-tracking device, users are faced

Authors' addresses: Zilu Liang, liang.zilu@kuas.ac.jp, Kyoto University of Advanced Science (KUAS), 18 Yamanouchi Gotanda-cho, Ukyo-ku, Kyoto, Japan, 615-8577; Bernd Ploderer, b.ploderer@qut.edu.au, Queensland University of Technology (QUT), Gardens Point 2 George St, Brisbane, Queensland, Australia, 4001.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2020 Association for Computing Machinery.

2474-9567/2020/3-ART17

<https://doi.org/10.1145/3380994>

with a trade-off between data quality and other factors like physical appearance and price. To complicate matters further, the accuracy of even the most advanced consumer sleep-tracking device is limited compared to medical devices [21, 55]. Medical sleep-tracking typically involves data from multiple sensors (accelerometers, heart rate, EEG, breathing) as well as personalized interpretations from sleep experts that take into account age, gender, health conditions and lifestyle [41, 81].

There is broad range of research on the information provided by sleep-tracking technologies, but these studies appear to fall into two separate areas. On the one hand, numerous studies have examined the accuracy of consumer devices compared to medical devices through controlled experiments [21, 48, 55, 57]. Such studies indicate that consumer devices tend to overestimate sleep and underestimate wake, and measurement accuracy can be associated to a person's age, subjective sleep quality and sleep patterns [54, 56]. However, these studies are based on assessments made by researchers, rather than assessments made by device users. On the other hand, studies in HCI have shown how people make sense of information about their own sleep and the challenges they face in collecting, interpreting, and acting upon such personal health information. For example, simple usability issues like battery limits and lack of comfort as well as irregular sleep patterns make it difficult to collect useful information [61]. Lack of contextual information and limited knowledge about sleep and sleep stages can make it difficult to interpret sleep information and to take actions from it [61, 77]. However, these HCI studies are not concerned with the accuracy of sleep information, or the credibility that people place in such information. Furthermore, these studies do not provide people with opportunities to compare consumer devices with medical devices.

To address this gap, this research examines how people judge the credibility of consumer sleep tracking devices in comparison with a medical device. Credibility is defined as believability, i.e., whether a person perceives a technology as trustworthy (truthful and free from bias) and as based on expertise (knowledge and competency) [27]. Credibility is of utmost importance for ubiquitous computing devices that report health and well-being data and offer recommendations to users [86], as users rely on these devices to assess their health and well-being and to explore opportunities to improve their health. A lack of credibility in such devices is likely to undermine the sustained usage and the potential benefit of these devices [16, 29, 50]. Users may even stop using a device if it is perceived non-credible [86]. Despite of the importance of credibility of sleep-tracking technologies, and self-tracking devices more broadly, there appears to be a lack of research into the perceptions of device users.

In our study, we invited twenty-two people to track their sleep for 3 nights concurrently using three devices: two consumer devices (Fitbit Charge 2 wristband and Neuroon wearable EEG eye mask) and one medical device (SleepScope, a portable medical EEG). All three devices generated similar types of information, like total sleep time (TST), number of awakenings (NAWK), and sleep stages. However, participants were faced with differences in the actual measurements provided by the different devices, which prompted them to make judgements about which device to believe in. Based on semi-structured interviews, we examined how people assessed the credibility of these three devices. Guided by Fogg and Tseng's framework of credibility [27], we sought to unpack how these judgements were influenced by different technology levels (device hardware, interface, functionality, and information), as well as whether these judgements were based on personal experience or surface level judgements.

The main contribution of this study is the empirical understanding of how people assess the credibility of sleep-tracking devices. The contribution builds on well-established HCI theories on credibility [27, 86]. Our findings show that people assess the credibility of a sleep-tracking device at different technology levels: from the quality of sleep data, to the mechanism, the dashboard design and even the physical appearance of the device. Assessments were affected by the participants' mental models of sleep-tracking, their difficulties in understanding sleep stages (sleep fallacy), and the black boxed nature of sleep tracking devices (black box fallacy). Based on these findings, we provide practical guidance for the design of credible sleep-tracking technologies.

2 RELATED WORK

2.1 Fundamentals of Human Sleep

Human sleep can be measured either subjectively or objectively along multiple dimensions [10]. Subjective sleep quality is characterized by individual's perception and satisfaction of sleep and is often measured using sleep diaries [13] or established questionnaires such as the Pittsburgh Sleep Quality Index (PSQI) [11] and Sleep disorder questionnaire (SDQ) [22]. These tools involve answering multiple questions daily or once every month. Human sleep can also be measured objectively in terms of timing, duration and ratio. Objective sleep quality is often characterized by a set of overall sleep metrics including total sleep time (TST), wake after sleep onset (WASO), sleep onset latency (SOL) and sleep efficiency (SE), and the structural organization of sleep such as sleep cycles. A sleep cycle typically consists of four sleep stages: light sleep, deep sleep, REM sleep and wake. Light sleep and deep sleep, also called non-REM sleep, are defined by the imprints on the EEG of highly rhythmic and/or large amplitude graphoelements [43]. Light sleep stage is where the sleep spindles and K-complexes emerge [30], which play important roles in consolidating brain functions related to learning, memory and in child development [2, 37, 88]. In addition, sleep spindle density has been associated with IQ [26]. Deep sleep is important for physical recovery [3], cognition [92], and for balancing sympathetic and parasympathetic nervous systems [52]. As for the role of REM sleep, it was not until recently that scientists uncovered the functional role of REM sleep in brain maturation [65] and emotion regulation [31]. Another important concept related to sleep is arousals, or awakenings. Whereas sleep arousals were previously considered as marker of sleep disruption and detrimental for sleep quality, recent studies have corrected this view and instead consider sleep arousals as an essential part of human sleep that occur spontaneously. The functional significance of arousals in sleep is to ensure the reversibility of sleep, otherwise sleep would be identical to coma [33]. Sleep arousals are further classified into cortical, subcortical and autonomic arousal [33]. It is worth noting that some arousals could be pathologic, such as the ones associated with sleep apnea and bruxism [33].

Regardless of measuring the same phenomenon, subjective sleep quality and objective sleep quality are only moderately correlated [40, 90]. The discrepancy between the two is often referred to as “sleep state misperception” [36]. Such misperception commonly exists among healthy adults [40] but could be more pronounced in pregnant women [76] and in clinical populations with Alzheimer disease, depression and sleep problems [38, 44, 62, 71, 87]. As we will show later in the findings, the discrepancy between objective measurements and subjective sleep experience adds an extra layer of complexity for participants to assess the credibility of sleep data. Participants had to reconcile such discrepancy to achieve a plausible view of their sleep.

2.2 Sleep Tracking and Monitoring Technologies

Measuring sleep structure has been the central task of all sleep tracking and monitoring technologies. Unlike other physiological signs such as heart rate and blood pressure, sleep structure cannot be directly measured but rather can only be inferred from other bio-signals. Human sleep is a complicated process accompanied by multiple streams of physiological changes in brain activities [85], limb movements [39], heart rate [89], body temperature [67], to name but a few.

In clinical settings, polysomnography (PSG) is the gold standard for measuring sleep structure and is mainly used for diagnosing breathing-related sleep disorders [4]. A PSG test involves concurrent monitoring on many physiological metrics including brainwave, oxygen level in blood, heart rate, breathing, and eye and leg movement. All these bio-signals are used to infer sleep stages. Sleep researchers have established a standard for sleep scoring [41]. They first divide the raw signals into 30s-epoch, and then assign a sleep stage epoch-by-epoch by visual evaluation. Nevertheless, sleep scoring involves many subjective judgements and therefore the average inter-rater agreement is only around 80% [81]. Wrist actigraphy is a less obtrusive alternative to PSG in clinical settings. It

can be used for longitudinal monitoring of sleep in free-living conditions and is mainly used for diagnosing sleep disorders related to circadian cycle [66, 73].

In recent years, consumer sleep-tracking technologies are becoming increasingly popular among individual users and researchers because they are unobtrusive, affordable and easy to use [48, 82]. These tools are designed in a way that people can use at home without the need for constant technical support. Like clinical sleep monitors, consumer sleep-tracking technologies broadly fall into two genres: EEG-based and actigraphy-based devices. EEG-based devices include Zeo Personal Sleep Coach, Neuroon eye mask, and SleepShepherd headband. These devices use embedded single or multi-channel EEG to measure users' brainwaves during sleep, which is then mapped to sleep stages. The genre of actigraphy-based devices includes popular activity trackers like Fitbit, smartwatches like Apple Watch, as well as smartphone applications like SleepBot and SleepAsAndroid. These devices or applications rely on embedded accelerometer to measure users' limb movement, and some wearable wristbands measure heart rate. These data are used to infer users' sleep structure and sleep quality using proprietary algorithms. Whereas older apps and devices mostly produced aggregated results on sleep such as total sleep time, total wake time, and sleep efficiency, newer apps and devices also provide information on sleep stages.

Human-computer interaction researchers have intensively studied how users interact with sleep-tracking technologies. Previous studies have qualitatively examined how users make sense of sleep data from single source [47, 60], the impact of consumer sleep-tracking devices on sleep hygiene [77], and the opportunities for sleep-tracking tools to enhance sleep hygiene [14]. There is strong evidence that sleep-tracking increases people's awareness of sleep health and support healthy sleep behaviors [61]. However, the long-term impact of sleep-tracking on sleep quality remains limited [58]. Some of the main obstacles include technology transparency, not tracking triggers, not giving actionable knowledge and not providing personalized recommendations [61]. Consequently, new sleep-tracking and analysis tools were developed to lower tracking effort [15], to explore the associations between sleep and external factors [47, 60], to support behavior changes for better sleep [5] and to support self-experimentation [17, 84].

The accuracy of consumer sleep-tracking devices has been a major concern for individual users and for researchers [5, 58, 59, 61, 77]. Many studies have examined the validity and reliability of these devices in laboratories [18, 19, 21, 63, 70] and in participants' own living environment [55]. Validation studies on previous wearable wristbands such as Fitbit Charge HR [19] and Jawbone UP [46] revealed that these devices generally underestimated wake time while overestimated sleep time and sleep efficiency in healthy adults, and their measurements of sleep stages had no correlation to medical devices [21, 48, 55, 57]. Whereas achieving better accuracy has been the main goal of many sleep-tracking technologies, it remains unknown if better accuracy increases perceived believability. This study set out to test this hypothesis.

2.3 Credibility in Human-Computer Interaction

Credibility has been well-studied in psychology for more than six decades [12] and has a relatively short history in human-computer interaction. It is a perceived quality and is not a direct measure of actual quality. Previous studies defined credibility as believability and outlined four different types of credibility: presumed credibility, reputed credibility, surface credibility, and experienced credibility. Presumed credibility relates to a person's assumptions and stereotypes. Reputed credibility relates to what a person has heard from third parties like friends or advertisements. Surface credibility relates to first impressions based on simple inspections, whereas experienced credibility is based on trying out the technology [27, 32].

With respect to what users focus on when they assess the credibility of computer products, Fogg and Tseng's systems perspective highlighted four dimensions involved in credibility assessment: device, interface, functional and information. Device credibility and interface credibility relate to the physical aspect of a device and its

user interface. Functional credibility relates to how a device achieves the functions it supposed to achieve, and information credibility relates to the perceived believability of information itself. Many factors have been shown to affect credibility assessment, such as user familiarity with the content, cognitive load, situational context, and visual design [28]. There is a rich body of research on how users judge the credibility of online information. Information credibility is defined as the extent to which one perceives the information to be believable [86]. Existing studies have identified a variety of antecedents of e-Health information credibility, including completeness of information [24], accuracy of information [6], visual appeal [79], and user traits such as demographics, motivation and beliefs [23, 69].

Despite of the importance of credibility, no study has systematically examined the credibility of personal information from sleep-tracking devices. As has been described in the previous section, studies on consumer sleep-tracking technologies have been mostly focusing on usability [14, 58, 61, 77], accuracy (both quantitative and qualitative) [21, 48, 55] and efficacy [16, 58]. Our study sets out to investigate the credibility of sleep-tracking technologies systematically in all four aspects outlined in Fogg and Tseng’s systems perspective: device, interface, functionality, and information.

3 METHOD

To investigate the credibility of sleep-tracking devices, we conducted a field study with 22 participants who tracked their sleep for 3 nights concurrently using three devices (Fitbit, Neuroon, and a medical device), followed by a semi-structured interview. By investigating how participants assess the credibility of these devices, this study seeks to offer insights into the factors that affect judgements and to pave the way for creating more credible digital health devices.

3.1 Participants

We recruited 22 participants (nine women, average age = 25 years) by distributing posters around the campus of the University of Tokyo. We excluded applicants who had diagnosed chronic conditions (including cardiovascular diseases, hypertension, and diabetes) and who could not attend the briefing and interview in person. All participants were students enrolled in the university. The ethics committee of the University of Tokyo approved this study. All participants provided written informed consent. Participants received a ¥6000 (equivalent to \$58) shopping card as an appreciation for their contribution to the study. No participant dropped out during the study.

Table 1 provides an overview of the demographic information of all participants. The ID numbers starting with N indicate novice users of sleep-tracking devices and those starting with E indicate experienced users who had used some type of tracking device before. The sleep problems listed in Table 1 were based on subjective perception and were not based on diagnosis. Seven participants were having perceived poor sleep at the time of this study as indicated by the PSQI (Pittsburgh Sleep Quality Index) [11]. Most of the problems were temporary due to the humid weather in June and had gone at the time of the interviews. The table also includes participants’ conceptions of what these devices measure, which influenced their perceptions of credibility as described in the findings section.

3.2 Sleep-tracking Devices

Participants in this study received three devices to track their sleep over 3 consecutive nights: a Fitbit Charge 2 [78], Neuroon [83], and a portable medical sleep monitor SleepScope [75]. Participants were asked to use these devices in parallel and to compare the results from each device during an interview at the end of the study. The rationale for providing multiple devices was based on a multiple source strategy reported in social psychology, which purports that having access to different sources of information (and technology) about the

Table 1. Demographic Information of Participants

ID	Sex	Age	PSQI ¹	Perceived sleep problems
N1 ²	F	22	6	Insomnia
N2	F	22	3	None
N3	M	23	11	Snore, insomnia
N4	M	30	6	Daytime sleepiness
E5	F	25	4	Snore
N6	M	26	3	None
N7	M	29	2	None
N8	M	21	7	None
E9	M	25	2	None
E10	F	28	4	Sleep paralysis, out of breath, snore
N11	M	21	7	None
N12	F	23	3	None
E13	F	23	1	None
N14	F	29	4	Temporary insomnia
E15	M	23	1	None
E16	M	24	2	Snore
E17	F	26	5	Sleep deficiency
E18	M	25	6	Temporary insomnia
E19	M	26	4	None
N20	F	32	3	None
E21	M	24	3	None
N22	M	33	4	None

¹A PSQI (Pittsburgh Sleep Quality Index) [11] greater than or equal to 5 indicates poor sleep quality.

²ID numbers starting with N indicate novice users; those starting with E indicate experienced users.

same phenomenon can increase participation engagement by allowing them to compare, contrast, and if necessary, reconcile contradictory information [34].

Fitbit Charge 2, Neuroon, and the SleepScope medical device were chosen because while they offer largely the same information about sleep, they differ in terms of other factors that may influence their credibility, i.e., in the physical manifestation of the device, interface design and amount of information, mechanism to collect and process data, and accuracy in information. At the time of the study, the Fitbit Charge 2 wristband and Neuroon were amongst the most popular actigraphy-based and EEG-based consumer devices. As summarized in Table 2, the Fitbit Charge 2 is worn on the wrist, presents information through a mobile app dashboard, and the information is based on data measured by accelerometer and optical heart rate sensor. In comparison, Neuroon is worn on the eyes, presents information through a mobile app dashboard (with different design), and the information is based on data measured by EEG. The SleepScope medical device was provided as a reference point to compare the information from the other 2 devices. We opted for a medical EEG device with interpretation from sleep experts rather than a medical wrist-worn actigraphy device, because the latter is known to be inaccurate [64]. A further consideration in choosing these devices was that they could be worn in parallel, as illustrated in Fig. 1. Other wearable EEG products such as Emotive and SleepShepherd were not selected because they were uncomfortable to wear during sleep and they could not be used concurrently with the medical device.

Table 2. Comparison of three devices in the aspects of device, interface, sensor and information

	Device	Interface	Sensor	Information
Fitbit Charge 2	Wristband	Mobile app dashboard	Accelerometer, pulse oximeter	Total sleep time (TST) Wake after sleep onset (WASO) Number of awakenings (NAWK) Sleep efficiency (SE) Ratio of 4 sleep stages Population benchmark
Neuroon	Eye mask	Mobile app dashboard	EEG (dry electrodes), EOG, accelerometer, pulse oximeter	TST, WASO, NAWK, SE Ratio of 4 sleep stages
Sleep Scope (medical device)	Electrodes attached to head; a cable connects electrodes and device	Printed report	EEG (gel-type electrodes)	TST, WASO, NAWK, SE Ratio of 5 sleep stages REM onset latency Duration of each sleep cycle Delta power, alpha power, etc.

Fig. 1. Placement of devices (left: front view; right: side view).

3.3 Data Collection and Analysis

The study started with a briefing session with each participant to explain how to use the devices and to install the necessary applications on their smartphones. Each participant received a Fitbit Charge 2 wristband, a Neuroon wearable EEG eye mask, a portable medical sleep monitor SleepScope, and accessories (batteries, electrodes and manuals). Participants were orally instructed to properly place the sensors. For example, Fitbit should be worn on non-dominant wrist and be placed a finger's width above wrist bone; the two electrodes of the medical devices

should be put in the center of forehead and behind the lobe of an ear. While we did not give any information to participants on how each device works, a few participants mentioned in their interviews afterwards that they looked up information about the devices on their own. We also provided participants with the link to a digital sleep diary to record their subjective sleep experience. The sleep diary consisted of nine questions adapted from standard sleep diary [13], such as “what time did you get into bed”, “how many times did you wake up”, “what time did you get out of bed for the day”, and “how would you rate the quality of your sleep”. Participants were instructed to fill in the diary every morning during data collection experiment. At the end of the briefing, participants were asked to fill in PSQI (Pittsburgh Sleep Quality Index) questionnaire [11] to establish a baseline of their sleep.

Participants then tracked their sleep using three devices concurrently for 3 nights in their homes. We analyzed sleep data of the second night to avoid the first-night effect [68]. The data of the third night was used only when Fitbit failed to generate sleep stage data on the second night. We retrieved Fitbit data through the Fitbit Web API. Neuroon data was manually retrieved from the dashboard, as there was no public API available at the time of this study. The medical data was sent to the company for analysis by sleep experts who downloaded the data from the device and manually interpreted the data following established sleep scoring standard [41]. A full analysis report was returned for each participant within one week.

When all sleep data were available, we conducted a one-hour semi-structured interview with each participant. Interview questions focused on participants’ prior knowledge of sleep, their subjective sleep quality (as noted in the diary), their experience with the three devices, comparison of sleep data from three devices and their evaluation of the overall credibility of each device. We did not fix the order of questions but adapted to the “flow” of the interview. During the interview, we presented to participants the sleep data measured by three devices on the same nights to encourage questions and reflection about the devices and the sleep data. We not only showed them the hypnograms from three devices plotted on the same time scale, but also a summary of the numerical data including TST, WASO, SE, and the ratio of each sleep stage. All interviews were conducted in English and were audio-recorded and transcribed verbatim.

The qualitative data from the interview transcripts was analyzed following the principles of a thematic analysis [9]. First, all authors read through the transcripts to become familiar with what the data entails and to discuss initial codes. Initial codes focused on usability concerns and the physical appearance of devices, sleep patterns and assumptions, and on the discrepancy between different sleep data and the trustworthiness of these data. The first author then coded by iteratively going through all transcripts in a software called Saturate, generating 498 unique codes. Due to the emerging codes around trustworthiness of the data and how such trust was shaped by the device, information and prior assumptions, we started to frame the data based on Fogg and Tseng’s systems perspective of credibility assessment [27]. Based on contextual information from participants and their statements, we sought to distinguish between statements based on experience, as opposed to superficial judgements based on presumption, reputation, or surface level inspections of the device. Thereafter both authors went through the excerpts and initial codes together to combine codes into overarching themes. The codes were grouped through virtual post-it notes on Padlet (<https://padlet.com/>). After several iterations, we grouped our codes into four themes: credibility in sleep information, functional credibility of sleep-tracking devices, dashboard interface credibility, and device physical appearance and credibility. These themes are described in detail in the next section.

4 FINDINGS

The findings show how the credibility of sleep-tracking devices rests on several factors: people judge credibility not only from the information that they receive but also based on their perceptions of the device functionality as well as based on their interactions with the device and interface. It is important to note that people’s judgements

Table 3. Identified Factors Affecting Credibility Assessment at Each Technology Level

Technology level	Factors
Information credibility	Sleep fallacy Validation with medical device Agreement with prior sleep knowledge and assumptions Agreement with subjective and desirable sleep experience
Functional credibility	Black-box fallacy EEG sensing
Interface credibility	Comprehensiveness of data Granularity/fragmentation of data
Device credibility	Device fitness Up-to-date appearance

may not always be accurate, because judgements could be based on presumptions, reputation and surface level judgements in addition to actual experience. Furthermore, some participants found it difficult to interpret sleep stages (sleep fallacy), and the black boxed nature of devices made it difficult for them to understand how information was sensed and processed (black box fallacy). Table 3 provides an overview of the findings about credibility assessments at different technology levels. In what follows, we first present the discrepancies between sleep data from different devices, followed by our findings about how people assess the credibility of devices in light of these discrepancies.

4.1 Discrepancies in Sleep Data

The findings highlighted discrepancies between data from different devices as well as discrepancies between device data and how people subjectively assess their sleep. As described in section 2, we expected such discrepancies. They are an important premise for this study because the lack of a single and accurate sleep assessment prompts people to compare and assess the credibility of different devices (which will be described in sections 4.2-4.5).

4.1.1 Conflicting Data from Three Devices. As illustrated in Fig. 2, the medical device and Fitbit were closely aligned for almost all participants on total sleep time (TST). However, 15 out of 22 participants had conflicting data on TST measured by Neuroon compared to the medical device. This was largely because the Neuroon eye mask easily slid off participants’ faces, which for 16 of the participants led to poor sleep data. Participants recognized these conflicts and commented that “*Neuroon cannot be fixed on the face so the data is not reliable*” (E18).

The discrepancies became more obvious when comparing the number of awakenings (NAWK) between devices and experiences recorded by participants in their sleep diary. As illustrated in Fig. 3, all the participants were faced with discrepancies between their subjective experience and the device data. Participants typically recorded only a small number of awakenings, which largely aligned with the number of awakenings counted by Fitbit and, though to a lesser degree, with the count provided by Neuroon. For this metric, there was a discrepancy between the participants’ subjective experience from their sleep diary and the data from the medical device, with the medical device data presenting a larger number of awakenings for all participants. The medical data was more closely aligned with a second Fitbit metric that presents the sum of the number of awakenings and the number of brief periods of restlessness (less than 2 minutes), but neither data reflected the subjective experiences of participants. Participants mentioned that the discrepancy on awakenings may be attributed to the differences in how this metric was defined.

Fig. 2. Comparison of total sleep time (TST) with three devices and sleep diary.

Fig. 3. Comparison of number of awakenings (NAWK) measured with three devices and sleep diary.

"I think the definition of awakening is different. I counted the awakenings that I recognized that I was awake. But maybe the medical device counted something different." (E13)

"I think the difference is the threshold of the awakening. It's because I moved too much during my sleep. I wasn't awake, but I moved, so the devices categorized it as awakening." (E16)

Discrepancies were also found among devices on sleep stage ratios. For example, Fig. 4 illustrates deep sleep ratio measured with three devices. Compared to Fitbit and Neuroon, the medical device showed much lower ratio of deep sleep for almost all participants. E16 commented that *"I was shocked to see that I had no deep sleep on the medical report"* and N22 said *"I'm surprised that I seldom had deep sleep. I thought it was natural to have deep sleep, in all sleep cycles, at least 3 times in one night. But according to the data, I didn't have much deep sleep"*.

Fig. 4. Comparison of deep sleep ratio measured with three devices.

4.1.2 Subjective Meaning of Sleep Quality. A second challenge in interpreting the sleep data is that human sleep can be measured using multiple metrics. In this study we asked participants to choose the most important and least important sleep metric for them. The results shown in Fig. 5 demonstrates that most participants (16 out of 22) consider feeling fresh in the morning as an indicator of having slept well. Other important sleep metrics include total sleep time (5), sleep efficiency (4) and sleep onset latency (4).

Fig. 5. Distribution of the most important and least important sleep metric for participants.

We also found that some participants considered more than one metric as critical in assessing their sleep quality. The main reason is that human sleep can be measured along multiple dimensions, and these dimensions are not totally independent from one another. For example, N1 and E12 considered wake up freshness and sleep efficiency equally critical as these two metrics are correlated.

"Sleep efficiency is important for me...also wake up freshness. I think these two are correlated." (N1)

"Bedtime will affect how fresh I feel when I wake up." (E19)

On the other hand, 6 out of 22 participants considered the number of awakenings as least important, followed by total sleep time (3), sleep onset latency (3) and wake after sleep onset (3). These metrics were considered not important partly because participants had no problem in these dimensions of sleep quality. For example, N7

states that *"Number of awakenings is not important for me because I can easily get back to sleep."* While wake up freshness was often considered as an important sleep metric, E5 and N14 expressed their concern of relying on this subjective measure that could potentially be affected by many factors other than sleep quality. *"During exam period, I usually woke up feeling fresh despite of short sleep hours. It was because I was motivated, not because my sleep quality was good. Feeling could be tricky."* (N11) In addition, E13, E17, E18, E21 found it hard to discard any sleep metric as they were all important. *"All these metrics are important to a certain degree."* (E13)

4.2 Credibility in Sleep Information

Information credibility relates to perceived believability of information about sleep length, sleep stages, etc. In judging this data, participants relied on the authority of the medical device information to validate information from Neuroon and Fitbit. Furthermore, participants judged information based on its plausibility in light of their knowledge and experience of sleep, even if their knowledge was limited.

4.2.1 Sleep Fallacy. Human sleep is a complex phenomenon that can be measured by many metrics along multiple dimensions. Interpreting the full set of metrics requires extensive domain expertise, which is often challenging for users of sleep-tracking technologies. Although participants in our study intuitively understood basic sleep metrics like total sleep time (TST) and time awake after sleep onset (WASO), they at times misinterpreted sleep stages. Nine of 22 participants considered deep sleep the most important to feel refreshed, while five participants doubted the significance of light sleep. Some participants also had unrealistic expectations on the amount of deep sleep that healthy adults could possibly get, which deviated significantly from scientific findings that healthy adults typically spend 16-20% of total sleep time in deep sleep stage [74].

"I still don't get a clear idea of light sleep. How does light sleep differ from deep sleep? Is it really as important as deep sleep for me?" (N6)

"I thought that sleep would be mostly deep sleep, at least 50%. That's what I would guess without any knowledge of how sleep works." (E21)

Some participants had an incorrect understanding of the functional role of REM sleep. REM sleep plays an important role in brain maturation and emotional memory processing [39]. It is an essential component of normal human sleep regardless of age. Nevertheless, seven participants had negative impression on REM sleep, and they associated having dreams with tiredness or stress. N4 and E13 thought dreaming had negative impact on overall sleep quality, while N3 and E15 clearly stated that they preferred not having dreams at all.

"Is REM necessary? I do not know whether it is good or not to have dreams..." (N3)

"In my understanding, if you are dreaming, either you are too tired, or you are sleeping too much. If I'm dreaming, that means my mind is not sleeping. I'm thinking of something else." (N4)

Another aspect of sleep fallacy relates to micro-arousals. Recent findings in sleep science indicate that micro-arousal is an essential part of human sleep and is important in ensuring the reversibility of sleep, otherwise sleep would be identical to coma [33]. In healthy adults, the average duration is around 15 seconds, and the number increases with age [8]. In our study, we found a common misunderstanding amongst participants who associated sleep disruptions solely with adverse consequences, because participants were not aware of the need for normal and healthy periods of micro-arousal.

"The medical report shows even more awakenings than the Fitbit. Do I have any problem?" (E18)

"Is it normal to be awake this often?" (E19)

Sleep fallacy affects the perceptions of information credibility. For example, some participants gave high credence to devices that produce better results—such as longer sleep or less awakenings—because they assumed that those are indicative of good sleep.

"I would imagine more deep sleep is good. So I trust Neuroon because it gave better data." (N14, Fitbit only showed three deep sleep bouts in the first half of night, while Neuroon showed extra two deep sleep bouts in the second half of night)

4.2.2 Validation with Medical Device Information Increases Reputed Credibility. A major consideration in judging the information from consumer sleep-tracking devices was if it aligned with the information provided in the report from the medical device. It is not surprising that by introducing this study to participants as a comparison between two different consumer devices and a medical device, that our participants perceived the medical device as having the highest level of accuracy. After all, labelling a device as 'medical' and having data interpreted by professional sleep experts enhances the reputation of the sleep information received. Hence, we found that the medical device indeed served as an authority for many participants that they first used to validate the information from other devices.

"It's already tested before it was released to be used in the hospital or for medical treatment, so I think it's the most reliable." (E19)

"When you say that it's a medical device, there is a little bit of bias in my feeling and maybe because I study medical related subjects, I more rely on authority's device than the commercial ones." (N20)

Subsequently, people believed in information from consumer devices when it aligned well with the information from the medical device.

"Because both Fitbit and the medical device say that I was awake here, I would trust the data." (N2)

"I trust Fitbit over Neuroon because it detected deep sleep in the second half of my sleep, which is consistent with the medical device." (N14)

4.2.3 Agreement with Prior Sleep Knowledge and Presumptions Increases Credibility. Our analysis suggests that many participants also assessed the credibility of sleep data by comparing the data with their prior knowledge. The data that agreed with what they presumed thus gained more credence than that deviated from their knowledge. For instance, a few participants knew that going to bed late would lead to poor sleep and that deep sleep was more likely to occur in the first half of night, and they were satisfied with the device that produced supportive data.

"Maybe I should trust the medical device on deep sleep, because I went to bed really late and I might not get that much deep sleep right after the exam. Fitbit showed too much deep sleep." (N11)

"As far as I know, usually we have good rest between 10 to 2 am. According to the data, I had deep sleep after 10 and before 2. I think I'm very satisfied with this data. It is true." (N20)

However, some participants had limited knowledge and erroneous presumptions. For them, data validity was less important than the agreement of data to what they assumed, especially on deep sleep. Being presented with data of deep sleep ratio by medical device (0.4%), Fitbit (17%) and Neuroon (24%), N14 reluctantly gave high credence to consumer devices despite that her rationale was in favor of the medical device. *"My brain is telling me the medical device is correct, but I don't want to trust medical device (because the numbers are not good). Probably just take average of Fitbit and Neuroon."* (N14). Similarly, N4 favored Fitbit over medical device for deep sleep. *"As for deep sleep, I trust Fitbit over the medical device because there's some deep sleep between 2-4 am as I expected."* (N4).

4.2.4 Agreement with Subjective and Desirable Sleep Experience Increases Credibility. Another important reference for credibility assessment of sleep data was the subjective sleep experience of participants, as noted in their diaries. Participants leveraged notes on dreams and sleep onset latency as reference points, and they were inclined to give more credence to device data that aligned with their notes and recollections of their sleep experience.

"The REM is just before waking up, so it corresponds to my image that I often wake up from dreams. I trust Fitbit over Neuroon because it's more consistent with my feeling." (E16)

"Deep sleep is only 0.6% in the medical report. According to my feeling, I felt good when I woke up. Before I saw the results, I trusted medical device. But now I choose Fitbit because it's closer to my feeling." (E18)

Although subjective experience of sleep was frequently used as a reference point, participants acknowledged that feelings could be incomplete, inaccurate and unreliable, stating that "we do not consciously experience sleep stages" (11 out of 22 participants), "memory is unreliable" (8), and "aawakenings could be too brief to be noticed" (6).

Participants have expectations of their sleep data based on their subjective sleep experience. The desirability of sleep data was an important consideration especially when sleep experience and sleep data did not align. Particularly participants who were confident of their sleep quality were inclined to give high credence to devices that produced desirable numbers. For example, N7 considered the one long awakening on Fitbit more believable than the 18 brief awakenings indicated by medical device and stated that "I think I should believe in Fitbit because the numbers are better. I told you I aim at better numbers, because I generally sleep well." (N7). Medical data with exaggerated number of awakenings or low deep sleep ratio were very likely to be discarded as non-credible.

"In terms of awakenings, I trust Fitbit because it follows my definition of awakening. I don't know what the medical device measures. Maybe my brain was awake, but my eyes were not awake." (N12)

"The medical report only showed very brief deep sleep. Compared with Fitbit, it's unbelievable because I felt good after I woke up." (E13)

"Normally, one would believe medical data, but so many awakenings reduced my willingness to believe it." (N22)

Participants highlighted a number of considerations in reconciling sleep data and subjective experience. We identified three strategies that participants adopted to reach a plausible view of their sleep. As summarized in Table 4, eight participants chose to believe their subjective sleep experience, while six participants favored device data. The rest chose to combine both: believing subjective experience on perceivable sleep metrics such as awakenings while relying on devices for non-perceivable metrics such as deep sleep and REM sleep. The choice of strategy had little to do with personal history of sleep disorder or prior self-tracking experience. That said, it is possible that good sleepers were more inclined to favor their subjective experience over data, as the average PSQI of group A (3.25 ± 1.67) was lower (indicating better subjective sleep) than that of group B (4.63 ± 2.82) and group C (4.67 ± 2.42), though the difference was not statistically significant according to Kruskal-Wallis test ($\chi^2 = 5.50, p = 0.24$).

4.3 Functional Credibility of Sleep-tracking Devices

Functional credibility relates to how a device achieves the functions it is supposed to achieve. Judging functional credibility of sleep-tracking devices was difficult for many participants, because they perceived the sleep-tracking devices as a black box that conceals how data is collected and processed. Hence, many participants relied on surface cues to judge the credibility: devices that collect EEG signals were seen as more credible in inferring sleep stages.

4.3.1 Black Box Fallacy. Sleep-tracking devices are complicated technologies that remain a black box for end users in terms of what these devices measure and how the measures are processed and mapped to sleep stages. Firstly, we found that many participants were not familiar with what exactly each device measured, regardless of whether they were experienced users or novice users (see Table 5). Only around half of the participants (12 out of 22 participants) knew exactly what the medical device measured, other participants either had no idea or thought the medical device measured limb movements like Fitbit. "In terms of technical mechanism, I'm not sure how the medical device measures movements." (N20) Some participants thought the medical device used a multimodal sensing scheme (though the electrode behind ear was just a ground electrode). "The medical device uses two electrodes, which made me feel it's more reliable." (N14)

Table 4. Strategies Participants Adopted to Reconcile Discrepancies between Data and Subjective Experience

Strategies	Participants	Quotes
(A) Favor subjective experience	Group A (8 participants) N1, N7, E9, N12, E13, E17, E21, N22	<p><i>“I believe my feeling and I doubt the records here. I feel my sleep is better than average, based on the time it takes me to fall asleep, and the fact that I don’t wake up during sleep.”</i> (E13)</p> <p><i>“If the device says something different from how I feel, I would say the device got broken.”</i> (E17)</p> <p><i>“The clinical data is telling me all those sleep cycles, but I don’t know what I can get out of it. The only thing that I definitely know is I felt rested in the morning.”</i> (N22)</p>
(B) Combine subjective experience and device data	Group B (8 participants) N3, N6, E10, N14, E16, E18, E19, N20	<p><i>“I would believe my experience, but for REM and deep sleep, there is no other way but to believe the medical data.”</i> (N3)</p> <p><i>“I will take average of the data from all devices for light sleep and REM. For awakenings I believe my feeling, and for deep sleep I should believe the medical data.”</i> (N14)</p> <p><i>“I only believe my feeling on the awakenings. Other than that, I believe the Fitbit data.”</i> (E19)</p>
(C) Favor device data	Group C (6 participants) N2, N4, E5, N8, N11, E15	<p><i>“I don’t think clearly during sleep, so I can’t really believe my experience. Maybe machines are more reliable and trustable.”</i> (N2)</p> <p><i>“I will believe the medical data because I’m not really certain about the quality of my sleep.”</i> (E15)</p>

On the other hand, two participants (N8, N11) thought Fitbit relied on the same mechanism of Neuroon and the medical device (measuring brainwaves), while 3 participants (N4, E16, N20) thought Neuroon detected limb movements like Fitbit.

“So the Fitbit senses brainwaves by detecting the blood flow I think... Oh, movements. Then I guess movements reflect brainwaves because the Fitbit data is quite similar to the medical data.” (N8)

Black box fallacy led some participants to arrive at wrong conclusions on the inner workings of these devices. For example, E17 and E21 assumed that the lights in Neuroon were optical sensors used to track users’ cornea. Since they thought this extra modality of input could enhance measurement accuracy, they gave more credence to Neuroon than Fitbit. *“I think the Neuroon uses good principle because there are the lights that track your pupils.”* (E21). On the contrary, good knowledge on how devices work is essential for participants to form realistic expectations on the devices.

“Fitbit didn’t detect the first REM. Maybe it was too short. Also, there are two long awakenings on Fitbit while the other devices didn’t detect them. I guess I was moving my arm or something during sleep. It seems quite logical because your arm is more likely to move than your head, so maybe that’s because of this.” (N20)

Secondly, even for participants who understood how devices collected data, it was often not clear how these signals related to sleep stages. This was particularly relevant for Fitbit, as participants were unclear how movement and heart rate data could be turned into sleep stages without having brainwave measures.

“I’m not quite sure of how Fitbit measures REM sleep. During REM sleep, our body doesn’t move, so how do wristbands measure it?” (N6)

Table 5. Participants' Perception on What Devices Measure

ID	Fitbit ¹	Neuroon ²	Medical device ³
N1 ⁴	No idea	Brainwaves	Brainwaves
N2	No idea	No idea	Brainwaves
N3	Movement	No idea	No idea
N4	Movement	Movement	Brainwaves
E5	No idea	No idea	No idea
N6	Movement	Brainwaves	Brainwaves
N7	No idea	Brainwaves	Brainwaves
N8	Brainwaves	No idea	No idea
E9	Movement & heart rate	No idea	No idea
E10	No idea	No idea	No idea
N11	Brainwaves & heart rate	Brainwaves	Brainwaves
N12	No idea	No idea	No idea
E13	No idea	Brainwaves	Brainwaves
N14	No idea	No idea	No idea
E15	No idea	Brainwaves	Brainwaves
E16	Movement	Movement	No idea
E17	Heart rate	No idea	No idea
E18	Heart rate	Brainwaves	Brainwave
E19	Movement	Brainwaves	Brainwaves
N20	Movement	Movement	Movement
E21	Movement & heart rate	Brainwaves	Brainwaves
N22	No idea	Brainwaves	Brainwaves

¹Fitbit measures movement and heart rate.

²Neuroon measures brainwave, eye and limb movement, and blood oxygen saturation.

³Medical device measures brainwave.

⁴ID numbers starting with N indicate novice users; those starting with E indicate experienced users.

"I see, the mechanism is on the movement of the body. Then how can it recognize what is light sleep and deep sleep? We don't move in REM sleep? I thought I didn't move in all the three stages." (N7)

Consequently, several participants questioned the credibility of the Fitbit device. Even when Fitbit agreed quite well to the medical device, participants attributed it to other reasons. Some believed that REM sleep stage might be easy to measure by any device, and thus good measurement accuracy on that sleep stage did not justify the advantage of Fitbit. One participant (N12) even considered the good agreement as coincidence.

"REM sleep is easy to measure, right? All three devices have the same results." (N11)

"So the Fitbit measures movement and heart rate. What becomes the significance of that data? I know your brainwave is supposed to tell you something about your sleep stages, but how can you measure the REM cycle and light sleep and deep sleep with heartbeat or movement? Maybe if I knew how heart rate is tied to brainwave, I will be convinced, but at this point, to me it's just like a coincidence that Fitbit gave similar results as the medical device." (N12)

4.3.2 EEG Sensing Increases Presumed Credibility. As participants had limited knowledge on how sleep-tracking devices work, they presumed that devices relying on EEG sensors worn on the head are more credible than

devices worn on the wrists. Hence, a major consideration in judging the functional credibility of sleep-tracking devices was whether it included EEG signals, even though the exact nature of EEG sensing was not understood. Consequently, Neuroon and the medical device were in general considered more trustworthy than Fitbit.

"Because the medical device puts sensors here on the head, it could catch more information, better than using here (the wrist)." (E15)

"I think Neuroon should be more accurate for deep sleep because it measures brainwaves, but for Fitbit it measures only movement." (E19)

4.4 Dashboard Interface Credibility

Interface Credibility relates to the way a device visualizes information and the interaction experience. Fitbit and Neuroon allowed participants to view sleep metrics through a mobile dashboard interface, as illustrated in Fig. 6 (a) and (b). Information from the medical device was presented to participants through a printed report that incorporated sleep stages as interpreted by sleep experts, as illustrated in Fig. 6 (c). In interacting with these dashboards and the report, we found that credibility was judged based on surface level appearances like the amount of detail offered as well as by the granularity of the information. There were no comments about other obvious differences in these examples, like the different colors used in these visualizations (e.g., whether an alarming color like red is better suited to denote awakenings, poor signals, or REM sleep), the immediacy of the dashboard compared to the report received after several days, or about the degree of interactivity that the Fitbit and Neuroon dashboards offered compared to the printed report.

4.4.1 More Detail May Both Increase and Decrease Surface Credibility. Based on simple inspection, participants noted that the medical report gave more comprehensive data than the consumer devices. For example, sleep was classified into five stages (wake W, light sleep N1, light sleep N2, deep sleep N3, and REM sleep R) in the medical report but was only classified into four stages on the consumer dashboards (wake, light sleep, deep sleep, and REM sleep). In addition, the medical report offered extra information about the number of sleep cycles, the duration of each sleep cycle, arousal index, and metrics of brainwave signal spectrum such as alpha, beta, and delta power.

A few participants considered the medical device most credible due to the extra amount of details given in the medical report. "I think I believe the medical device because the data is in more detailed than the consumer devices." (E16). Others considered too much detail unnecessary, and they gave more credence to Fitbit because the data was easier to understand and thus more trustworthy.

"The medical data is more into details. It takes time to understand the data. If I just want something that I can check regularly or compare with previous days, Fitbit is more appropriate." (E5)

"I believe Fitbit over Neuroon not because it's close to the medical report, but because it is easy to understand, and the movement is reasonable." (N11)

4.4.1 More Detail May Both Increase and Decrease Surface Credibility. Based on simple inspection, participants noted that the medical report gave more comprehensive data than the consumer devices. For example, sleep was classified into five stages (wake W, light sleep N1, light sleep N2, deep sleep N3, and REM sleep R) in the medical report but was only classified into four stages on the consumer dashboards (wake, light sleep, deep sleep, and REM sleep). In addition, the medical report offered extra information about the number of sleep cycles, the duration of each sleep cycle, arousal index, and metrics of brainwave signal spectrum such as alpha, beta, and delta power.

A few participants considered the medical device most credible due to the extra amount of details given in the medical report. "I think I believe the medical device because the data is in more detailed than the consumer devices." (E16). Others considered too much detail unnecessary, and they gave more credence to Fitbit because the data was easier to understand and thus more trustworthy.

(a) Sleep data presented on Fitbit dashboard

(b) Sleep data presented on Neuroon dashboard

(c) Sleep data presented in a medical report

Fig. 6. Interface of three devices: (a) Fitbit dashboard, (b) Neuroon dashboard, and (c) medical report.

"The medical data is more into details. It takes time to understand the data. If I just want something that I can check regularly or compare with previous days, Fitbit is more appropriate." (E5)

"I believe Fitbit over Neuroon not because it's close to the medical report, but because it is easy to understand, and the movement is reasonable." (N11)

4.4.2 Higher Data Granularity May Both Increase and Decrease Surface Credibility. A second surface level cue for judging the dashboard credibility was the degree of granularity of the data. In the medical report, the hypnogram of a whole night's sleep is featured by high granularity (and thus more fragmentations), while the hypnograms in Fitbit and Neuroon dashboards look smoother and more continuous. On the one hand, many participants found it hard to believe in highly fragmented plots. As illustrated in Fig. 7, the hypnogram of E10 contains many brief segments of deep sleep. She commented, *"I have a hard time believing small things moving like that in the medical report, because I don't know if it makes much sense to have a deep sleep of 30 seconds and then completely change. I feel I can believe things that are longer and more consistent like what's shown on Fitbit."* (E10, Fig. 7). For the same reason, N6 did not believe medical data for awakenings and light sleep because these parts were too fragmented.

Fig. 7. The hypnogram of E10 contains many brief segments of deep sleep (highlighted in dashed red circles). The fragmentation of the data on the interface undermined the credibility of the medical device for E10.

"I'm trying to think on how fragmented the data is. I trust the medical device for deep and REM sleep, but not for awake and light sleep. These parts are just too messy." (N6)

On the other hand, two participants considered the medical device the most credible exactly because of the high data granularity in the plot. The quotes below suggest that participants here did not necessarily consider the detailed content provided but rather used the layout as a heuristic cue to judge the believability of the devices.

"Medical data are plotted by points while Fitbit data are plotted by lines. I think the graph that is plotted by points is more accurate than that plotted by lines because the sampling rate is higher. Fitbit just take two or three points and it couldn't get accurate data in between." (N7)

"It's hard to detect deep and REM sleep because they are only small portion of sleep. So sampling rate could have strong impact on the results of REM and deep sleep." (N11)

4.5 Device Physical Appearance and Credibility

Device credibility relates to the physical aspect of a device such as physical appearance and hardware structure. We found that physical attractiveness indeed affected surface credibility. For example, one participant gave higher credence to Neuroon simply because it looked more up-to-date than the medical device. *"In general, I would trust Neuroon even more than the medical device, because it's newer. The medical device looks old because the display is like those on cellphones eight years ago."* (N11).

The shape of a device also influenced perceptions of credibility. Devices that were easy to wear and whose sensors were firmly placed received more credence especially after participants had gained first-hand experience with the devices. The three devices used in this study were very different in shape. The Fitbit device was a lightweight wristband worn on the non-dominant arm. Neuroon was an eye mask worn on the face and was fixed by a rubber band. In the study, the eye mask easily slid off participants' faces, leading to deteriorated signals and thus poor data quality. Sixteen out of the 22 participants complained that Neuroon eye mask slid off their faces. Many of them gave low credence to Neuroon after the experiment even if they initially ranked it better than Fitbit. The medical device consisted of a body cuboid of size 95mm by 64mm, which did not need to be attached to the user. Instead, it connected to two electrodes that were glued to the head, one on the forehead and the other behind an ear. Despite the inconvenience of the wires, the fact that the two electrodes were adhesive, and that the device was not worn on the body lent the medical device more credibility than the consumer devices.

"The medical device had contact to my forehead and neck all the time, but Neuroon may have slid off my face, and Fitbit may have gotten some errors when I changed my position." (N11)

"I'm just wondering what happened to the signal quality. Maybe it didn't fit my facial shape. Maybe somehow the mask was not touching where it was supposed to touch." (E17)

5 DISCUSSION

This paper presents an understanding of how people assess the credibility of sleep-tracking devices. Building on well-established HCI theory on credibility [27], our findings reveal that credibility is based on assumptions, personal experience of sleep (and time awake), and comparing data from consumer devices with data from the medical device. The credibility of information per se is not the sole factor in judging a device: functional, interface, and device aspects also influence judgements. These judgements are often superficial, rather than based on experience and deep understanding, and to some extent inaccurate. The credibility framing highlighted two fallacies in sleep-tracking. Firstly, like others [61, 77], we observed sleep metrics and sleep stages do not align with the way people experience and understand their sleep. We also observed that participants interpreted sleep disruptions as a concern, because they were unaware that micro-arousals are an essential and healthy part of human sleep [33]. Both of these misconceptions lead to sleep fallacies. Secondly, technologies conceal the way data is collected and processed, which makes it difficult to develop a better understanding of sleep and results in a black-box fallacy.

The novelty of this contribution lies in that we used three devices of distinct design, mechanism and information accuracy to tease out the factors that affect users' credibility assessment at different technology levels. This methodology allowed us to uncover aspects that are key to different technology levels of credibility and that are not possible to be disentangled using one device, such as device functionality (related to presumed and surface credibility), dashboard interface design (related to surface credibility) and physical appearance (related to surface and experienced credibility). The medical device provided a reference point for participants to evaluate the accuracy of consumer devices, thus allowing us the probe whether accuracy is key to information credibility as has been suggested in other context such as online health systems and car navigation systems [27].

This study sets a precedent in studying the credibility of sleep-tracking technologies. In what follows we discuss the findings along the four technology levels of credibility assessment, the two types of fallacies we observed in sleep-tracking, and the ways in which people aligned their knowledge and experience with sleep data. Based on these discussions, we highlight design opportunities to strengthen credibility.

5.1 Four Levels of Credibility Assessment

Sleep-tracking research has been dominantly focused on users' perception of sleep information per se, especially whether the information is accurate [59, 61, 77, 94] and meaningful to people [14–16, 47, 60]. This study confirms

that the perceived accuracy of sleep data itself indeed is a main aspect that participants considered when assessing the credibility of a sleep-tracking device, as several participants gave high credence to medical data and to consumer devices that agreed with the medical device. This finding echoes the work by Yang and colleagues that users of self-tracking devices took accuracy into account when they assessed their data [45]. Without having medical data as reference point, therefore, previous sleep-tracking studies generally assumed that better measurement accuracy led to higher credibility, as it would increase the trustworthiness of sleep data offered by these devices [20, 21, 55].

Nevertheless, our study further found that data accuracy, despite of being a key concern in sleep research, is not necessarily a key concern when people assess the credibility of technology. In the domain of sleep medicine, sleep diaries or questionnaires that measure subjective sleep experience are routinely ranked lower than any kind of sleep-tracking devices—either contactless devices, contact devices or PSG—with respect to accuracy [42]. However, our study shows that approximately a third of the participants gave higher credence to their subjective experience over data, and another third acknowledged the importance of both subjectivity and objectivity. Data accuracy that is high-valued in clinical sleep research thus may not necessarily reflect what people value in sleep-tracking in daily life. To this end, our study provides complementary insights into the role of accuracy in sleep-tracking. In addition to medical data, participants also relied on their prior sleep knowledge and subjective sleep experience to assess the believability of the sleep data. The availability of multiple reference points added an extra layer of complexity. When these reference points contradicted one another, especially when subjective sleep experience deviated from device data, participants had to reconcile the differences by either choosing one over the other or by combining or averaging over the two.

Second, in addition to the perceived quality of sleep data, our analysis showed that participants also considered the credibility of the device mechanism. The mechanism of a device mostly affected presumed and surface credibility of a device. Due to the black-box nature of sleep-tracking devices, most participants had no understanding of how exactly the devices work. Nevertheless, they relied on heuristic cues such as the placement of sensors and were prone to give credence to devices that rely on EEG sensing.

Third, our finding echoes previous studies on e-health system that the comprehensiveness and granularity of data presented on dashboard interface affected surface credibility [25, 79, 80]. Other factors such as color scheme, the immediacy of data, and the degree of interactivity did not seem to affect participants’ assessment of interface credibility in this study. The potential impact of these factors is left for further investigation. Interestingly, no consensus was found regarding the effect of data comprehensiveness and granularity. For a few participants, comprehensive data with high granularity in the medical report gained reputed credibility. Nevertheless, the majority of participants considered comprehensive data overwhelming to digest on daily basis, and they were reluctant to believe data in high granularity as they assumed fragmented data were not believable. Hence, Fitbit gained more credence over medical report for its clear and simple interface design and for the smoothened plots in the dashboard. Compared to Neuroon, Fitbit was also considered more credible because it offered extra information (e.g. personal and population benchmark) to help users gain insights into their sleep quality [93]. These findings indicate that a balance between comprehensiveness and simplicity is the key to interface credibility in sleep-tracking. It is thus critical to present the right amount of sleep data that is easy to interpret and digest for each user.

Fourth, we found that conveying credence through physical appearance is important for consumer sleep-tracking devices, as participants did “judge a book by its cover”. Among the two consumer devices, many participants initially gave more credence to Neuroon than Fitbit due to its resemblance to the medical device (e.g. the device was worn on the head to detect brainwaves) or because it looked more up to date. Nevertheless, they later changed their opinion because Neuroon often slid off their faces and led to deteriorated signal quality. As Neuroon had problem fitting to various types of facial geometry, many participants concluded that Fitbit

was more credible among the two consumer devices. These findings show that physical appearance of a device affected not only surface credibility but also experienced credibility.

5.2 Two Types of Fallacies in Sleep Tracking

Another key finding of this study is the two types of fallacy in participants' mental model of sleep-tracking: sleep fallacy and black-box fallacy. These fallacies can negatively affect the ways in which users assess credibility and how they interpret sleep data.

Sleep fallacies mainly occurred due to participants' misunderstanding of sleep continuity (e.g. number of awakenings) and sleep stages. Brief awakenings, though often being considered as a symbol of sleep disruption and thus associated to poor sleep, are a normal part of human sleep [33]. Many participants were doubtful about the excessive number of awakenings shown in the medical report, especially when they did not consciously experience those awakenings and they were confident of their sleep quality. In contrast, consumer devices tended to ignore brief arousals and only marked awakenings that lasted longer than 2 minutes, which contributed to increased credibility. Additionally, participants often had a limited understanding of the functional role of different sleep stages and the normal amount of each sleep stage that healthy adults can possibly get. In line with the finding in [77], we found that participants placed undue emphasis on deep sleep. Furthermore, many participants had unrealistic expectation of how much deep sleep they could possibly get (higher than 50%). In contrast, the functional importance of REM was not recognized: some participants wrongly assumed that REM contributes little to body restoration; others associated having dreams with tiredness or stress.

Technology transparency has been considered important in sleep tracking [58, 61, 77]. However, we found that participants often failed to assess the functional credibility of the three sleep-tracking devices through sensible analysis, because the devices remained a black box that prevented participants from constructing a correct conceptual model of how they work [15, 77]. Some participants wrongly asserted that Fitbit detected brainwave through blood flow, while others assumed that the medical EEG measured body movements. Due to lack of domain knowledge, it was difficult for participants to understand the sensing modalities from the physical structure of device, like sensors and lights. Without knowing how the bio-signals relate to sleep stages, participants found it hard to believe Fitbit data despite of its agreement with the medical device. Consequently, credible device information was considered as non-credible. Therefore, helping users understand how sleep-tracking devices work is critical to gaining credibility, as has been shown in previous studies about people's trust on machines [72].

5.3 (Mis)alignment between Subjective Sleep Experience and Objective Sleep Data

In this study, good alignment on total sleep time was found—for many participant—between Fitbit and the medical device (regardless of relying on different mechanisms), and between device data and what participants recorded in their sleep diary. However, significant discrepancies were found for awakenings and sleep stages (especially deep sleep).

The reason for such discrepancies boils down to the tension between what users consciously experience about their sleep (i.e. the known) and what users are not able to consciously experience during sleep (i.e. the unknown). The fifteenth century Philosopher Nicholas of Cusa argued that knowledge is always interpreted within the context of a combination of known and unknown [7]. This is especially true in sleep tracking where people may know about prolonged awakenings and total sleep, but where they also have to make sense of the objective data that they cannot consciously experience and know, such as sleep stages and brief awakenings marked by EEG signals.

Our findings show that participants adopted different strategies to reach a plausible view on these data. For awakenings, they were inclined to ignore brief arousals that were not consciously perceived and only relied on

long awakenings that they remembered or noted in their diary to judge the credibility of device data. Participants were sensitive to excessive number of awakenings, which echoes findings from related work that sleep quality essentially means sleep continuity for many people [1]. Consequently, medical device data lost credibility because it counted in any kind of awakenings, both long awakenings that participants knew about as well as micro-arousals that they were not aware of. Fitbit and Neuroon, on the other hand, gained credibility by only recording major awakenings longer than 2 minutes. Neuroon totally ignored shorter awakenings, while Fitbit marked them as restlessness. Although Fitbit restlessness is inferred from wrist movement rather than from brainwave signals, it agreed reasonably well to micro EEG arousals in medical data as indicated by the quantitative comparison. Some participants attributed the discrepancies on awakenings to the difference in how this metric was defined in different settings. Indeed, the interchangeable use of “awakening” and “arousal” may account for the misalignment of medical data and subjective experience in terms of sleep continuity.

For sleep stages, on the other hand, many participants tended to judge the credibility of the unperceivable objective measures based on subjective sleep satisfaction—something they were certain of. Even though the correlations between subjective feeling and objective measures are typically only moderate [40, 90], many participants assumed a proportional relationship between how restorative they felt about their sleep and the ratio of deep sleep. Consequently, they questioned the believability of extremely low ratio of deep sleep indicated by the medical report and favored data from consumer devices that showed a higher ratio. This phenomenon can be best explained by the confirmation bias theory in psychology, where people favor information that confirms their preexisting beliefs to alleviate cognitive dissonance [91].

The misalignment between objective sleep measures and subjective sleep quality—a phenomenon known as sleep misperception—has been well-documented in sleep research [35, 36]. From a human-computer interaction perspective, we narrowed down possible reasons to the inconsistency in metric definition (i.e. awakening vs arousal) and to the undue mapping from subjective sleep experience to objective measures. Addressing the misalignment between what devices say and how people feel about their sleep is important as it affects not only how users assess the believability of sleep data but also sustained usage of sleep-tracking technologies [51]. Despite of the tension between objectivity and subjectivity in sleep tracking, it is probably less appropriate to consider the solution as a dichotomy where we need to choose one or the other than bringing the two together. Objective sleep quality and subjective sleep quality measures different aspects of the same phenomenon and complement each other. The mission of sleep tracking technologies is to achieve a better alignment between the two by addressing the reasons for misalignment.

5.4 Design Challenges and Opportunities

Based on these observations, we highlight the following design challenges and opportunities to enhance the credibility of sleep-tracking technologies.

Design challenge 1: Support the construction of sleep knowledge. Our findings show that sleep fallacy was common and a major reason for confirmation bias in sleep-tracking. A key opportunity for design is to provide users with an opportunity to construct new knowledge, which builds on the knowledge and experience of users, expands on sleep data, addresses potential fallacies and biases, and provides insight into evidence-based sleep knowledge. We recommend sleep-tracking devices offer evidence-based information on how sleep metrics are defined, the functional roles of micro arousal and sleep stages, the physiological changes accompanying each sleep stage, as well as the National Sleep Foundation Recommendations [74] to benchmark good sleep (i.e. the recommended range for objective sleep measures). Being able to construct new knowledge is critical for users to assess the trustworthiness of sleep data against scientific knowledge as well as to align objective measures and subjective sleep experience.

Design challenge 2: Increase the transparency of device mechanism and validity. Technology transparency has been considered important in sleep tracking [58, 61, 77]. It is thus critical to help users form an appropriate mental model of how sleep-tracking devices work and the validity of different types of devices. Some of the key information include what bio-signals are measured and the significance of these signals in sleep staging. Plotting bio-signals together with sleep data to emphasize the signal patterns that are used to infer sleep stages may also increase perceived trustworthiness. Since each type of sleep-tracking technology has its target user group [64], we recommend device makers to also inform users the strength and limitations of the device based on established evidence to help them form realistic expectations [21, 48, 55].

Design challenge 3: Identify objective measures that are relevant and knowable to each user. Our findings show that the key to interface credibility is to strike a balance between comprehensiveness and simplicity. Previous study found that users abandoned tracking devices if they perceived the collected data as not useful [51]. Instead of showing the same full set of sleep measures to all users on daily basis, an alternative strategy is to adapt the dashboard content to the sleep pattern and needs of each user through user profiling. For example, users can be sub-grouped based on clinical/physiological characteristics or based on association of objective measures and subjective sleep quality rating [49]. Our finding show that participants used different metrics to assess their sleep quality. Therefore, sleep-tracking technologies may focus on objective measures that users relate to or incorporate conscious experiences and knowledge that people have about their sleep. One approach is to develop new algorithms to predict subjective sleep quality based on multiple nights of objective measures, or to incorporate diary data, as findings from sleep studies show that several nights of actigraphy measurements are needed to reach a better match to subjective sleep quality [53]. In addition, metrics demonstrating high variations over a time period (e.g. one month or half a year) should be prioritized over more constant metrics, because these metrics has more clinical significance [53]. These strategies will not only help users to narrow down their focus on the most relevant metrics but also contribute to the alignment of objective measures and subjective sleep quality.

5.5 Recommendations for Future Research

The limitations of this study point out several directions for future research. First, this study provides insight into experience related to initial credibility of sleep-tracking devices developed over three nights, but it did not look at how credibility may be lost and regained over time [16, 29, 61]. Our findings pave the way for future longitudinal studies to investigate how the credibility of sleep-tracking devices may be lost, to what extent it can be regained, and what factors are involved. Second, the possible effects of gender, age, motivation for sleep-tracking and prior experience with sleep tracking was not explored in the current study due to the homogeneity of the study cohort (e.g., all college students). While investigating the effect of these factors was not the focus of this study, future research may examine the importance of these factors for the development of credibility and how they may be used to profile and model users. Third, this study did not delve into the possible effects of interface design aspects such as color scheme, the immediacy of data, and the degree of interactivity as they were out of the main scope. Further studies are needed to examine how these aspects may affect perceived interface credibility of sleep-tracking devices.

6 CONCLUSION

This study has examined how people assess and perceive the credibility of sleep-tracking technologies. Our findings provided insights into credibility at four different technology levels: information quality, device functionality, dashboard interface design, and physical appearance of a device. Credibility assessments were based on first-hand experience with three sleep-tracking devices as well as on superficial judgements and assumptions. We identified two blind spots—sleep fallacy (incomplete and erroneous sleep knowledge) and black-box fallacy

(incorrect understanding of device mechanism)—which pose challenges in assessing credibility. These findings offer two implications. First, enhancing the credibility of sleep-tracking technologies should seek solutions at different technology levels ranging from information accuracy (as has been generally assumed in previous studies) to mechanism, interface design and device design. Second, because the details of sleep stages and the inner workings of black-boxed devices were not well understood, and because of the potential misalignment between objective data and subjective sleep experience, people’s credibility assessment may deviate from the real level of trustworthiness of the devices. Therefore, addressing these two fallacies and the misalignment between objective data and subjective sleep experience is central to the design of credible sleep-tracking technologies that help people gain insights into their sleep by allowing them to establish connections between their lived experience, their data and scientific evidence.

ACKNOWLEDGMENTS

This work is supported by the Japan Society for the Promotion of Science (JSPS), under grant JSPS KAKENHI 16H07469 and 19K20141. The authors would like to thank the participants for their contribution to this study.

REFERENCES

- [1] Torbjorn Akerstedt, Ken Hume, David Minors, and Jim Waterhouse. 1994. The meaning of good sleep: a longitudinal study of polysomnography and subjective sleep quality. *J Sleep Res* 3, 3 (1994), 152–158. <https://doi.org/10.1111/j.1365-2869.1994.tb00122.x>
- [2] Florin Amzica and Mircea Steriade. 2002. The functional significance of K-complexes. *Sleep Medicine Reviews* 6, 2 (2002), 139–149. <https://doi.org/10.1053/smr.2001.0181>
- [3] Nayyab Asif, Razia Iqbal, and Chaudhry Fahad Nazir. 2017. Human immune system during sleep. *American Journal of Clinical and Experimental Immunology* 6, 6 (2017), 92–96.
- [4] Lauren Babcock, Kelsey Riggs, Vernon Rowe, John Hunter, Arlene O’Shea, Doug Schell, Dorsey Paul, Mark Varona, Shelia Miller, James Barnett, and Snehal Kadia. 2018. Polysomnography is crucial to the diagnosis of sleep disordered breathing syndromes in patients with hypermobility spectrum disorder (P1.099). *Neurology* 90, 15 Supplement (2018).
- [5] Jared S. Bauer, Sunny Consolvo, Ben Greenstein, Jonathan W Schooler, Eric Wu, Nathaniel F Watson, and Julie A. Kientz. 2012. ShutEye: encouraging awareness of healthy sleep recommendations with a mobile, peripheral display. In *Proceedings of the 2012 SIGCHI Conference on Human Factors in Computing Systems*. ACM Press, New York, NY, 1401–1410. <https://doi.org/10.1145/2207676.2208600>
- [6] Gretchen K. Berland, Marc N. Elliott, Leo S. Morales, Jeffrey I. Algazy, Richard L. Kravitz, Michael S. Broder, David E. Kanouse, Jorge A. Muñoz, Juan-Antonio Puyol, Marielena Lara, Katherine E. Watkins, Hannah Yang, and Elizabeth A. McGlynn. 2001. Health information on the Internet: accessibility, quality, and readability in English and Spanish. *JAMA* 285, 20 (2001), 2612–2621. <https://doi.org/10.1016/j.physbeh.2016.03.006>
- [7] J. Bornemark. 2018. The limits of ratio: an analysis of NPM in Sweden using Nicholas of Cusa’s understanding of reason. In *Metric Culture: Ontologies of Self-tracking Practices*, B. Ajana (Ed.). Emerald Publishing Limited, UK, 235–253.
- [8] M. Boselli, L. Parrino, A. Smerieri, and M. Terzano. 1998. Effect of age on EEG arousals in normal sleep. *Sleep* 21, 4 (1998), 351–357.
- [9] Virginia Braun and Victoria Clarke. 2015. Using thematic analysis in psychology. *Qual Res Psychol* 3, 2 (2015), 77–101. <https://doi.org/10.1191/1478088706qp0630a>
- [10] Daniel J. Buysse. 2014. Sleep Health: Can We Define It? Does It Matter? *Sleep* 37, 1 (2014), 9–17. <https://doi.org/10.5665/sleep.3298>
- [11] Daniel J. Buysse, Charles F. Reynolds, Timothy H. Monk, Susan R. Berman, and David J. Kupfer. 1989. The Pittsburgh sleep quality index: a new instrument for psychiatric practice and research. *Psychiatry Res* 28, 2 (1989), 193–213. [https://doi.org/10.1016/0165-1781\(89\)90047-4](https://doi.org/10.1016/0165-1781(89)90047-4)
- [12] Walter Weiss Carl I. Hovland. 1951. The influence of source credibility on communication effectiveness. *Public Opinions Quarterly* 15, 4 (1951), 635–650. <https://doi.org/10.1086/266350>
- [13] Colleen E. Carney, Daniel J. Buysse, Sonia Ancoli-Israel, Jack D. Edinger, Andrew D. Krystal, Kenneth L. Lichstein, and Charles M. Morin. 2012. The consensus sleep diary: standardizing prospective sleep self-monitoring. *Sleep* 35, 2 (2012), 287–302. <https://doi.org/10.5665/sleep.1642>
- [14] Eun Kyoung Choe, Sunny Consolvo, Nathaniel F Watson, and Julie A. Kientz. 2011. Opportunities for computing technologies to support healthy sleep behaviors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM Press, New York, NY, 3053–3062. <https://doi.org/10.1145/1978942.1979395>
- [15] Eun Kyoung Choe, Bongshin Lee, Wanda Marie Pratt Matthew Kay, and Julie A. Kientz. 2015. SleepTight: low-burden, self-monitoring technology for capturing and reflecting on sleep behaviors. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM Press, New York, NY, 121–132. <https://doi.org/10.1145/2750858.2804266>

- [16] Nediya Daskalova, Bongshin Lee, Jeff Huang, Jessica Lundin, and Chester Ni. 2018. Investigating the effectiveness of cohort-based sleep recommendations. *Proc ACM Interact Mob Wearable Ubiquitous Technol* 2, 3 (2018), Article 101. <https://doi.org/10.1145/3264911>
- [17] Nediya Daskalova, Danae Metaxa-Kakavouli, Adrienne Tran, Nicole Nugent, Julie Boergers, John McGeary, and Jeff Huang. 2016. SleepCoach: A Personalized Automated Self-Experimentation System for Sleep Recommendations. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. ACM Press, New York, NY, 347–358. <https://doi.org/10.1145/2984511.2984534>
- [18] Massimiliano de Zambotti, Fiona C. Baker, and Ian M. Colrain. 2015. Validation of sleep-tracking technology compared with polysomnography in adolescents. *Sleep* 38, 9 (2015), 1461–1468. <https://doi.org/10.5665/sleep.4990>
- [19] Massimiliano de Zambotti, Fiona C. Baker, Adrian R. Willoughby, Job G. Godino, David Wing, Kevin Patrick, and Ian M. Colrain. 2016. Measures of sleep and cardiac functioning during sleep using a multi-sensory commercially-available wristband in adolescents. *Physiological Behavior* 158 (2016), 143–149. <https://doi.org/10.1016/j.physbeh.2016.03.006>
- [20] Massimiliano de Zambotti, Stephanie Claudatos, Sarah Inkelis, Ian M. Colrain, and Fiona C. Baker. 2015. Evaluation of a consumer fitness-tracking device to assess sleep in adults. *Chronobiology International* 32, 7 (2015), 1024–1028. <https://doi.org/10.3109/07420528.2015.1054395>
- [21] Massimiliano de Zambotti, Aimee Goldstone, Stephanie Claudatos, Ian M. Colrain, and Fiona C. Baker. 2018. A validation study of Fitbit Charge 2 compared with polysomnography in adults. *Chronobiology International* 35, 4 (2018), 465–476. <https://doi.org/10.1080/07420528.2017.1413578>
- [22] Alan B. Douglass, Robert F. Bornstein, German Nino-Murcia, Sharon Keenan, and et al. 1994. The Sleep Disorders Questionnaire: I. creation and multivariate structure of SDQ. *Sleep* 17, 2 (1994), 160–167. <https://doi.org/10.1093/sleep/17.2.160>
- [23] Mohan Dutta-Bergman. 2003. Trusted online sources of health information: differences in demographics, health beliefs and health-information orientation. *J Med Internet Res* 5 (2003), e21. <https://doi.org/10.2196/jmir.5.3.e21>
- [24] Mohan J. Dutta-Bergman. 2006. The impact of completeness and web use motivation on the credibility of e-health information. *Journal of Communication* 54, 2 (2006), 253–269. <https://doi.org/10.1111/j.1460-2466.2004.tb02627.x>
- [25] Gunther Eysenbach. 2008. *Credibility of health information and digital media: new perspectives and implications for youth*. Digital Media, Youth, and Credibility. Number 123-154. The MIT Press, Cambridge, MA. <https://doi.org/10.1162/dmal.9780262562324.123>
- [26] Zhuo Fang, Valya Sergeeva, Laura B. Ray, and Jeremy Viczko. 2017. Sleep spindles and intellectual ability: epiphenomenon or directly related? *Journal of Cognitive Neuroscience* 29, 1 (2017), 167–182. https://doi.org/10.1162/jocn_a_01034
- [27] B. J. Fogg and Hsiang Tseng. 1999. The elements of computer credibility. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM Press, New York, NY, 80–87. <https://doi.org/10.1145/302979.303001>
- [28] Krisandra S. Freeman and Jan H. Spyridakis. 2004. An examination of factors that affect the credibility of online health information. *Technical Communication* 51, 2 (2004), 239–263.
- [29] Cecilie Gaziano. 1988. How credible is the credibility crisis? *Journalism Mass Communication Quarterly* 65, 2 (1988), 267–278. <https://doi.org/10.1177/107769908806500202>
- [30] L. D. Gennaro and M. Ferrara. 2003. Sleep spindles: an overview. *Sleep Medicine Review* 7, 5 (2003), 423–440. <https://doi.org/10.1053/smr.2002.0252>
- [31] S. Groch, I. Wilhelm, S. Diekelmann, and J. Born. 2013. The role of REM sleep in the processing of emotional memories: Evidence from behavior and event-related potentials. *Neurobiology of Learning and Memory* 99 (2013), 1–9. <https://doi.org/10.1016/j.nlm.2012.10.006>
- [32] Albert C. Gunther. 1992. Biased press or biased public? Attitudes towards media coverage of social groups. *The Public Opinion Quarterly* 56, 2 (1992), 147–167.
- [33] Péter Halász, Mario Terzano, Liborio Parrino, and Róbert Bódizs. 2004. The nature of arousal in sleep. *J Sleep Res* 13, 1 (2004), 1–23. <https://doi.org/10.1111/j.1365-2869.2004.00388.x>
- [34] Stephen G. Harkins and Richard E. Petty. 1981. Effects of source magnification of cognitive effort on attitudes: an information-processing view. *Journal of Personality of Social Psychology* 40 (1981), 401–413. <https://doi.org/10.1037/0022-3514.40.3.401>
- [35] Allison G. Harvey, Kathleen Stinson, Katriina Whitaker, Damian Moskowitz, and Harvinder Virk. 2008. The subjective meaning of sleep quality: a comparison of individuals with and without insomnia. *Sleep* 31, 3 (2008), 383–393.
- [36] Allison G. Harvey and Nicole Tang. 2012. (Mis)Perception of sleep in insomnia: a puzzle and a resolution. *Psycho Bull* 138, 1 (2012), 77–101. <https://doi.org/10.1037/a0025730>
- [37] Nora Hennies, Matthew A. Lambon Ralph, Marleen Kempkes, James N. Cousins, and Penelope A. Lewis. 2016. Sleep spindle density predicts the effect of prior knowledge on memory consolidation. *Journal of Neuroscience* 36, 13 (2016), 3799–3810. <https://doi.org/10.1523/JNEUROSCI.3162-15.2016>
- [38] Vanessa Herbert, Daniel Pratt, Richard Emsley, and Simon D. Kyle. 2017. Predictors of nightly subjective-objective sleep discrepancy in poor sleepers over a seven-day period. *Brain Sciences* 7, 3 (2017), Article 29. <https://doi.org/10.3390/brainsci7030029>
- [39] James A. Horne. 2015. Human REM sleep: influence on feeding behavior, with clinical implications. *Sleep Medicine* 16, 8 (2015), 910–916. <https://doi.org/10.1016/j.sleep.2015.04.002>
- [40] Fan-Chi Hsiao, Pei-Jung Tsai, Changwei W Wu, Chien-Ming Yang, Timothy Joseph Lane, Hsin-Chien Lee, Ling-Chun Chen, We-Kang Lee, Lu-Hsin Lu, and Yu-Zu Wu. 2018. The neurophysiological basis of the discrepancy between objective and subjective sleep during

- the sleep onset period: an EEG-fMRI study. *Sleep* 41, 6 (2018), zsy056. <https://doi.org/10.1093/sleep/zsy056>
- [41] C Iber, Sonia Ancoli-Israel, A. L. Chesson, and Stuart F. Quan. 2017. For the American Academy of Sleep Medicine. The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications. *Darien, IL: American Academy of Sleep Medicine* Version 2.4 (2017).
- [42] Vanessa Ibáñez, Josep Silva, and Omar Cauli. 2018. A survey on sleep assessment methods. *PeerJ* 6 (2018), e4849. <https://doi.org/10.7717/peerj.4849>
- [43] Andreas A. Ioannides, Lichan Liu, Vahe Poghosyan, and George K. Kostopoulos. 2017. Using MEG to understand the progression of light sleep and the emergence and functional roles of spindles and K-complexes. *Frontiers in Human Neuroscience* 11 (2017), Article 313. <https://doi.org/10.3389/fnhum.2017.00313>
- [44] Marta Jackowska, Samantha Dockray, Hilde Hendrickx, and Andrew Steptoe. 2011. Psychosocial factors and sleep efficiency: discrepancies between subjective and objective evaluations of sleep. *Psychosomatic Medicine* 73, 9 (2011), 810–816. <https://doi.org/10.1097/PSY.0b013e3182359e77>
- [45] Ing-Marie Jonsson, Helen Harris, and Clifford Nass. 2008. How accurate must an in-car information system be? Consequences of accurate and inaccurate information in cars. In *Proceedings of CHI 2008*. ACM Press, New York, NY, 1665–1674. <https://doi.org/10.1145/1357054.1357315>
- [46] Seung-Gul Kang, Jae Myeong Kang, Kwang-Pil Ko, Seon-Cheol Park, Sara Mariani, and JiaWeng. 2017. Validity of a commercial wearable sleep tracker in adult insomnia disorder patients and good sleepers. *Journal of Psychosomatic Research* 97 (2017), 38–44. <https://doi.org/10.1016/j.jpsychores.2017.03.009>
- [47] Matthew Kay, Eun Kyoung Choe, Jesse Shepherd, Benjamin Greenstein, Nathaniel Watson, Sunny Consolvo, and Julie A. Kientz. 2012. Lullaby: a capture access system for understanding the sleep environment. In *Proceedings of the ACM Conference on Ubiquitous Computing*. ACM Press, New York, NY, 226–234. <https://doi.org/10.1145/2370216.2370253>
- [48] Bhanu Prakash Kolla, Subir Mansukhani, and Meghna P. Mansukhani. 2016. Consumer sleep tracking devices: a review of mechanisms, validity and utility. *Expert Review of Medical Devices* 13, 5 (2016), 497–506. <https://doi.org/10.1586/17434440.2016.1171708>
- [49] Andrew D. Krystal and Jack D. Edinger. 2018. Measuring sleep quality. *Sleep Medicine* 9, Supplement 1 (2018), S10–S17. [https://doi.org/10.1016/S1389-9457\(08\)70011-X](https://doi.org/10.1016/S1389-9457(08)70011-X)
- [50] Tarcan Kumkale, Dolores Albarracin, and Paul J. Seignourel. 2010. The effects of source credibility in the presence or absence of prior attitudes: implications for the design of persuasive communication campaigns. *J Appl Soc Psychol* 40, 6 (2010), 1325–1356. <https://doi.org/10.1111/j.1559-1816.2010.00620.x>
- [51] Amanda Lazar, Christian Koehler, Theresa Jean Tanenbaum, and David H Nguyen. 2015. Why we use and abandon smart devices. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp’15)*. ACM Press, New York, NY, 635–646. <https://doi.org/10.1145/2750858.2804288>
- [52] Julia Lechinger, Dominik Philip Johannes Heib, Walter Gruber, Manuel Schabus, and Wolfgang Klimesch. 2015. Heartbeat-related EEG amplitude and phase modulations from wakefulness to deep sleep: interactions with sleep spindles and slow oscillations. *Psychophysiology* 52, 11 (2015), 1441–1450. <https://doi.org/10.1111/psyp.12508>
- [53] Sakari Lemola, Thomas Ledermann, and Elliot M. Friedman. 2013. Variability of sleep duration is related to subjective sleep quality and subjective well-being: an actigraphy study. *PLoS ONE* 8, 8 (2013), e71292. <https://doi.org/10.1371/journal.pone.0071292>
- [54] Zilu Liang and Mario Alberto Chapa-Martell. 2018. Not all errors are created equal: influence of user characteristics on measuring errors of consumer wearable devices for sleep tracking. *EAI Endorsed Transactions on Pervasive Health and Technology* 18, 15 (2018), e4. <https://doi.org/10.4108/eai.24-7-2018.159404>
- [55] Zilu Liang and Mario Alberto Chapa-Martell. 2018. Validity of consumer activity wristbands and wearable EEG for measuring overall sleep parameters and sleep structure in free-living conditions. *Journal of Healthcare Informatics Research* 2, 1-2 (2018), 152–178. <https://doi.org/10.1007/s41666-018-0013-1>
- [56] Zilu Liang and Mario Alberto Chapa-Martell. 2019. Accuracy of Fitbit wristbands in measuring sleep stage transitions and the effect of user-specific factors. *JMIR Mhealth Uhealth* 7, 6 (2019), e13384. <https://doi.org/10.2196/13384>
- [57] Zilu Liang and Mario Alberto Chapa-Martell. 2019. Combining numerical and visual approaches in validating sleep data quality of consumer wearable wristbands. In *Proceedings of the 17th Annual IEEE International Conference on Pervasive Computing and Communications*. 777–782. <https://doi.org/10.1109/PERCOMW.2019.8730805>
- [58] Zilu Liang and Bernd Ploderer. 2016. Sleep tracking in the real world: a qualitative study into barriers for improving sleep. In *Proceedings of the 28th Australian Conference on Computer-Human Interaction*. ACM Press, New York, NY, 537–541. <https://doi.org/10.1145/3010915.3010988>
- [59] Zilu Liang, Bernd Ploderer, and Mario Alberto Chapa-Martell. 2017. Is fitbit fit for sleep-tracking?: sources of measurement errors and proposed countermeasures. In *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare*. ACM Press, New York, NY, 476–479. <https://doi.org/10.1145/3154862.3154897>
- [60] Zilu Liang, Bernd Ploderer, Wanyu Liu, Yukiko Nagata, James Bailey, Lars Kulik, and Yuxuan Li. 2016. SleepExplorer: a visualization tool to make sense of correlations between personal sleep data and contextual factors. *Personal Ubiquitous Comput* 20, 6 (2016), 985–1000.

- <https://doi.org/10.1007/s00779-016-0960-6>
- [61] Wanyu A. Liu, Bernd Ploderer, and Thuong Hoang. 2015. In bed with technology: challenges and opportunities for sleep tracking. In *Proceedings of the Annual Meeting of the Australian Special Interest Group for Computer Human Interaction*. ACM Press, New York, NY, 142–151. <https://doi.org/10.1145/2838739.2838742>
 - [62] Hannah G. Lund, Bruce D. Rybarczyk, Paul B. Perrin, David Leszczyszyn, and Edward Stepanski. 2013. The discrepancy between subjective and objective measures of sleep in older adults receiving CBT for comorbid insomnia. *Journal of Clinical Psychology* 69, 10 (2013), 1108–1120. <https://doi.org/10.1002/jclp.21938>
 - [63] Janna Mantua, Nickolas Gravel, and Rebecca M. C. Spencer. 2016. Reliability of sleep measures from four personal health monitoring devices compared to research-based actigraphy and polysomnography. *Sensors* 16, 5 (2016), 646. <https://doi.org/10.3390/s16050646>
 - [64] Miguel Marino, Yi Li, Michael N. Rueschman, J. W. Winkelman, J. M. Ellenbogen, J. M. Solet, Hilary Dulin, Lisa F. Berkman, and Orfeu M. Buxton. 2013. Measuring sleep accuracy, sensitivity, and specificity of wrist actigraphy compared to polysomnography. *Sleep* 36, 11 (2013), 1747–1755. <https://doi.org/10.5665/sleep.3142>
 - [65] Gerald A. Marks, James P. Shaffery, Arie Oksenberg, Samuel G. Speciale, and Howard P. Roffwarg. 1995. A functional role for REM sleep in brain maturation. *Behavioural Brain Research* 69, 1-2 (1995), 1–11. [https://doi.org/10.1016/0166-4328\(95\)00018-O](https://doi.org/10.1016/0166-4328(95)00018-O)
 - [66] Jennifer L. Martin and Alex D. Hakim. 2011. Wrist actigraphy. *CHEST* 139, 6 (2011), 1514–1527. <https://doi.org/10.1378/chest.10-1872>
 - [67] Antonio Martinez-Nicolas, Marc Guaita, Joan Santamaria, Josep M. Montserrat, Maria Angeles Rol, and Juan Antonio Madrid. 2017. Circadian impairment of distal skin temperature rhythm in patients with sleep-disordered breathing. *Sleep* 40, 6 (2017), zsx067. <https://doi.org/10.1093/sleep/zsx067>
 - [68] Catherine McCall and Vaughn McCall. 2012. Objective vs subjective measurements of sleep in depressed insomniacs: first night effect or reverse first night effect? *J Clin Sleep Med* 8, 1 (2012), 59–65. <https://doi.org/10.5664/jcsm.1664>
 - [69] Harrison McKnight and Charles Kacmar. 2007. Factors and effects of information credibility. In *Proceedings of the ninth international conference on electronic commerce*. 423–432. <https://doi.org/10.1145/1282100.1282180>
 - [70] Lisa J. Meltzer, Laura S. Hiruma, Kristin Avis, Hawley Montgomery-Downs, and Judith Valentin. 2015. Comparison of a commercial accelerometer with polysomnography and actigraphy in children and adolescents. *Sleep* 38, 8 (2015), 1323–1330. <https://doi.org/10.5665/sleep.4918>
 - [71] Els I. S. Most, Samir Aboudan, Philip Scheltens, and Eus J.W. Van Someren. 2012. Discrepancy between subjective and objective sleep disturbances in early- and moderate-stage Alzheimer disease. *Am J Geriatric Psychiatry* 20, 6 (2012), 460–467. <https://doi.org/10.1097/JGP.0b013e318252e3ff>
 - [72] Bonnie M. Muir. 1987. Trust between humans and machines, and design of decision aids. *International Journal of Man-Machine Studies* 27, 5-6 (1987), 527–539. [https://doi.org/10.1016/S0020-7373\(87\)80013-5](https://doi.org/10.1016/S0020-7373(87)80013-5)
 - [73] V. Natale, D. Leger, M. Martoni, V. Bayon, and A. Erbacci. 2014. The role of actigraphy in the assessment of primary insomnia: a retrospective study. *Sleep Medicine* 15, 1 (2014), 111–115. <https://doi.org/10.1016/j.sleep.2013.08.792>
 - [74] M. Ohayon, E. M. Wickwire, M. Hirshkowitz, and et al. 2017. National Sleep Foundation’s sleep quality recommendations: first report. *Sleep Health* 3, 1 (2017), 6–19. <https://doi.org/10.1016/j.sleh.2016.11.006>
 - [75] J.T. Quigley. 2014. *This sleep-tracking medical device is so good it’s being tested by Japan’s astronauts*. Retrieved February 6, 2020 from <https://www.techinasia.com/sleepwell-sleep-scope-japan-medtech-startup-eeg-alternative>
 - [76] Leontien M. Van Ravesteyn, Joke H.M. Tulen, Astrid M. Kamperman, Monique E. Raats, A.J. (Tom) Schneider, Erwin Birnie, Eric A.P. Steegers, Witte J.G. Hoogendijk, Henning W. Tiemeier, and Mijke P. Lambregtse-van den Berg. 2014. Perceived sleep quality is worse than objective parameters of sleep in pregnant women with a mental disorder. *J Clin Sleep Med* 10, 10 (2014), 1137–1141. <https://doi.org/10.5664/jcsm.4118>
 - [77] Ruth Ravichandran, Sang-Wha Sien, Shwetak N. Patel, Julie A. Kientz, and Laura R. Pina. 2017. Making sense of sleep sensors: how sleep sensing technologies support and undermine sleep health. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI’17)*. ACM Press, New York, NY, 6864–6875. <https://doi.org/10.1145/3025453.3025557>
 - [78] CR Consumer Reports. 2020. *Fitbit Charge 2 fitness tracker*. Retrieved February 6, 2020 from <https://www.consumerreports.org/products/fitness-tracker/fitbit-charge-2-387694/overview/>
 - [79] David Robins and Jason Holmes. 2008. Aesthetics and credibility in web site design. *Information Processing Management* 44, 1 (2008), 386–399. <https://doi.org/10.1016/j.ipm.2007.02.003>
 - [80] David Robins, Jason Holmes, and Mary Stansbury. 2010. Consumer health information on the web: the relationship of visual design and perceptions of credibility. *Journal of the American Society for Information Science and Technology* 61, 1 (2010), 13–29. <https://doi.org/10.1002/asi.21224>
 - [81] Richard S. Rosenberg and Steven Van Hout. 2013. The American Academy of Sleep Medicine inter-scorer reliability program: sleep stage scoring. *J Clin Sleep Med* 9, 1 (2013), 81–87. <https://doi.org/10.5664/jcsm.2350>
 - [82] Anita Valanju Shelgikar, Patricia F. Anderson, and Marc Stephens. 2016. Sleep tracking, wearable technology, and opportunities for research and clinical care. *CHEST* 150, 3 (2016), 732–743. <https://doi.org/10.1016/j.chest.2016.04.016>

- [83] Sleeptrackers.io. 2020. *Neuroon: Intelligent sleep mask and light therapy device*. Retrieved February 6, 2020 from <https://sleeptrackers.io/neuroon-mask>
- [84] Sara Taylor, Akane Sano, Craig Ferguson, Akshay Mohan, and Rosalind W. Picard. 2018. QuantifyMe: An open-source automated single-case experimental design platform. *Sensor* 18, 4 (2018), 1097. <https://doi.org/10.3390/s18041097>
- [85] Marton Toth, Istvan Kondakor, and Bela Faludi. 2016. Differences of brain electrical activity between moderate and severe obstructive sleep apneic patients: a LORETA study. *Journal of Sleep Research* 25, 5 (2016), 596–604. <https://doi.org/10.1111/jsr.12403>
- [86] Shawn Tseng and B.J. Fogg. 1999. Credibility and computing technology. *Commun. ACM* 42, 5 (1999), 39–44. <https://doi.org/10.1145/301353.301402>
- [87] Kounosuke Tsuchiyama, Haruo Nagayama, Kiyomi Kudo, Katsuhiko Kojima, and Kumiko Yamada. 2003. Discrepancy between subjective and objective sleep in patients with depression. *Psychiatry and Clinical Neurosciences* 57, 3 (2003), 259–264. <https://doi.org/10.1046/j.1440-1819.2003.01114.x>
- [88] Peter P. Ujma, Piroska Sandor, Sara Szakadat, Ferenc Gombos, and Robert Bodizs. 2016. Sleep spindles and intelligence in early childhood: developmental and trait-dependent aspects. *Developmental Psychology* 52, 12 (2016), 2118–2129. <https://doi.org/10.1037/dev0000233>
- [89] Geert J. M. van Boxtel, Pierre J. M. Cluitmans, Roy J. E. M. Raymann, and Martin Ouwerkerk. 2018. Heart rate variability, sleep, and the early detection of post-traumatic stress disorder. *Sleep and Combat-Related Post Traumatic Stress Disorder* (2018), 253–263. https://doi.org/10.1007/978-1-4939-7148-0_22
- [90] Anna Westerlund, Ylva Trolle Lagerros, Goran Kecklund, and John Axelsson. 2016. Relationships between questionnaire ratings of sleep quality and polysomnography in healthy adults. *Behavioral Sleep Medicine* 14, 2 (2016), 185–199. <https://doi.org/10.1080/15402002.2014.974181>
- [91] Ryan W. White and Eric Horvitz. 2015. Belief dynamics and biases in web search. *ACM Transactions on Information Systems* 33, 4 (2015), Article 18. <https://doi.org/10.1145/2746229>
- [92] Kristine A. Wilckens, Fabio Ferrarelli, Matthew P. Walker, and Daniel J. Buysse. 2018. Slow-wave activity enhancement to improve cognition. *Trends in Neurosciences* 41, 7 (2018), 470–482. <https://doi.org/10.1016/j.tins.2018.03.003>
- [93] Huahai Yang, Yunyao Li, and Michelle Zhou. 2014. Understand users’ comprehension and preferences for composing information visualizations. *ACM Transactions on Computer-Human Interaction (TOCHI)* 21, 1 (2014), Article 6. <https://doi.org/10.1145/2541288>
- [94] Rayoung Yang, Eunice Shin, Mark Webster Newman, and Mark Steven Ackerman. 2015. When fitness trackers don’t ‘fit’: end-user difficulties in the assessment of personal tracking device accuracy. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM Press, New York, NY, 623–634. <https://doi.org/10.1145/2750858.2804269>