Preferences and Effectiveness of Sleep Data Visualizations for Smartwatches and Fitness Bands

Alaul Islam, Ranjini Aravind, Tanja Blascheck, Anastasia Bezerianos, Petra Isenberg

Preferences and Effectiveness of Sleep Data Visualizations for Smartwatches and Fitness Bands

Alaul Islam
mohammad-alaul.islam@inria.fr
Université Paris-Saclay, CNRS, Inria,
LISN
Gif-sur-Yvette, France

Ranjini Aravind
ranjini1992@gmail.com
Université Paris-Saclay
Gif-sur-Yvette, France

Tanja Blascheck
tanja.blascheck@vis.uni-stuttgart.de
University of Stuttgart
Stuttgart, Germany

Anastasia Bezerianos
anastasia.bezerianos@lri.fr
Université Paris-Saclay, CNRS, Inria,
LISN
Gif-sur-Yvette, France

Petra Isenberg
petra.isenberg@inria.fr
Université Paris-Saclay, CNRS, Inria,
LISN
Gif-sur-Yvette, France

Figure 1: Most preferred combinations of sleep visualizations and data from our survey for different types and granularities of sleep data.

ABSTRACT

We present the findings of four studies related to the visualization of sleep data on wearables with two form factors: smartwatches and fitness bands. Our goal was to understand the interests, preferences, and effectiveness of different sleep visualizations by form factor. In a survey, we showed that wearers were mostly interested in weekly sleep duration, and nightly sleep phase data. Visualizations of this data were generally preferred over purely text-based representations, and the preferred chart type for fitness bands, and smartwatches was often the same. In one in-person pilot study, and two crowdsourced studies, we then tested the effectiveness of the most preferred representations for different tasks, and found that participants performed simple tasks effectively on both form factors but more complex tasks benefited from the larger smartwatch size. Lastly, we reflect on our crowdsourced study methodology for testing the effectiveness of visualizations for wearables. Supplementary material is available at https://osf.io/yz8ar/.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in visualization**; **Mobile devices**.

KEYWORDS

glanceable visualization, smartphones, smartwatch, fitness trackers

1 INTRODUCTION

A large number of people are interested in analyzing data about their sleep to improve their health, and overall well-being [42]. To this end people often wear smartwatches and fitness bands that come with sensors, and software that allows capturing data on sleep duration, quality, stages, or related aspects such as oxygen saturation, and heart rate. Wearables are a growing market, and as such, it is timely to investigate how best to visualize data important to the wearers of these devices.

In this paper, we focus on the visualization of sleep data on fitness trackers. Fitness trackers are wearable devices like wristbands, smartwatches, and sports watches that primarily expose data about

health, and fitness activities. Fitness trackers are increasingly capable of capturing sleep data accurately [40], and sleep tracking apps have become central for many people seeking to improve their sleep behaviors [32]. Currently, sleep data is mostly checked on companion apps on smartphones, or websites rather than on the tracker itself [33]. Checking data on the tracker, however, would allow people to quickly glance at their data post-activity like they already do for other types of tracked activities without consulting alternative devices. Fitness tracker wearers also expressed this desire for in-situ consultation of sleep data and reported wanting to see sleep data on their trackers, in particular directly after waking up [5]. Our follow-up survey confirmed this desire, and 38% of respondents mentioned checking sleep data directly after waking up. We would expect people to focus primarily on simple sleep aggregates (such as the number of hours slept) in such a quick-glance scenario. Nevertheless, we explore both simple and more complex visualizations as we expect that wearers begin to explore advanced features of their fitness trackers after extended use [22]. We have no data to explain why some device manufacturers do not expose more detailed sleep data on trackers themselves. It is possible that device manufacturers do not know the interest in this data but also that the lack of guidelines, examples, and design considerations for visualizations of detailed sleep data play a role.

Our work makes several contributions to address these concerns. First, we provide results of a survey with 108 fitness tracker wearers,[1] in which we learned how wearers track their sleep, what they wish to learn from their sleep data, and which sleep visualizations they prefer for different granularities of sleep data, and for different device form factors: smartwatch, or fitness band. Our results indicated that the most preferred visualizations were often similar for people who wore smartwatches, and those who wore fitness bands, even though the fitness band-sized visualizations were roughly half the size of the smartwatch-sized ones (Figure 1).

Our second contribution is three quantitative studies about the effectiveness of preferred visualizations across three types of form factors (smartwatch, horizontal, and vertical fitness band) to make recommendations about visualizations to use for sleep data.

We first conducted a detailed in-person pilot study followed by two crowdsourced perceptual studies that compared the preferred visualizations from the survey for weekly, and nightly sleep data under three different analysis tasks. Participants preferred, and were more confident with the smartwatch-sized visualizations but simple tasks could be performed equally effectively on both smartwatch, and fitness band form factors. Only the more complex tasks seemed to benefit from the larger display size.

Finally, we offer reflections on our study methodology, during which we balanced reaching a larger pool of crowdsourced participants with ecological validity, and direct experimental control, compared to our in-person pilot study; hoping to expand discussions on appropriate study methodologies for wearable visualization.

---

[1]Details about the questionnaire are available in the supplementary material–https://osf.io/f3vja/.

2 RELATED WORK

Research on visualizations for wearables is still sparse. Prior work with fitness trackers has focused mostly on the accuracy of tracking sensors, and recommendations to improve engagement with wearers. Our study focuses on the use, and perception of fitness data visualizations, specifically for sleep data; therefore, we discuss related research on the current use of wearables, and their visualizations.

2.1 Wearables in Use

The practice of using technology like wrist-worn wearables (e.g., smartwatches, fitness bands) is becoming more widespread. Wearers of fitness tracking devices often use visualizations to increase self-awareness, and self-knowledge that may lead to behavioral change. Previous work on fitness trackers has focused mostly on the accuracy of tracking sensors [41, 65, 68], and recommendations to improve engagement with wearers [18, 27, 61]. More recently, some studies have investigated which representations of fitness data wearers prefer for specific use cases, such as during activity, or for long-term data exploration [3, 66].

Carrion et al. [18] conducted a study with 30 teenagers using the Withings Pulse smartwatch [67], and the Misfit Shine [59] fitness tracker for a week. The main data participants used were calories burned, steps made, and sleeping time because these were related to sports, and activity contexts participants were familiar with, and related to specific goals they had set. In another similar online diary study with 34 wearers of Fitbit [16], and Jawbone [34] trackers, Asimakopoulos et al. [6] found that multiple factors motivated participants to wear a device: (i) seeing if they met their goal (movement, sleep, calories count); (ii) looking, and feeling good, improving their mood, and avoid sitting; as well as (iii) getting tips, and recommendations about their health.

Schirra, and Bentley [58] as well as Cecchinato et al. [19] conducted interviews with early adopters of smartwatches with a focus on reasons for adoption, and what tasks wearers used the smartwatches for. Later studies focused on commonly used features of smartwatches, finding that people mainly used smartwatches to monitor, and track activities, or respond to notifications in addition to timekeeping [1, 20, 46, 54]. Others looked at specific smartwatch usees such as in classrooms [56], for healthcare purposes [28, 35, 36, 45], stress detection [60], real-time eating activity detection [62], or understanding a wearer's emotional state [57].

In contrast to this stream of research, we focus on the representation of data directly on fitness trackers, and how researchers should design dedicated sleep data representations for the small-scaled display.

2.2 Current Visualizations on Wearables

The literature on smartwatch visualization is still sparse. The few publications that exist focused either on studying representations for smartwatches, or on designing representations for these small displays. For example, researchers studied low-level perceptual tasks to understand glanceability of smartwatch visualizations [10], the impact of visual parameters (e.g., size, frequency, and color) on reaction times [44], or representation preferences in an air traffic control use case [49]. Others' visualization research described

novel visualization designs specifically for smartwatches. Examples include research on representing health, and fitness data on smartwatches [3, 50], for line charts [51], temporal data [63], activity tracking more broadly [29], and even for integrating visualizations in watchstraps [38]. There were some design exploration studies on smartwatches in the past, which introduced novel representations like a border visualization [21], a woven color field [2], stripe painting [52], and avatars to represent the quantified self [48]. A recent study found a predominant display of health, and fitness data on watch faces [33], and the most frequent representation type for data was icons accompanied by text. Visualizations are still not as common as other representations such as text; even though wearers can easily, and efficiently use them to represent some of the most commonly displayed data (e.g., health, and fitness data). Most current trackers tend to display only data about the most recent activity while offloading historic data to environments with larger displays like smartphone, tablet apps, or websites. The associated environment typically offers visualizations of other long-term aggregated data in various temporal granularities like hours, days, weeks, and months.

In contrast to these works, our studies contribute visualizations that wearers prefer for sleep data, and effectiveness of preferred visualizations across three types of form factors (smartwatch, horizontal, and vertical armband) to make recommendations about visualizations to use for sleep data.

3 UNDERSTANDING SLEEP VISUALIZATION PREFERENCES

As a first step towards recommendations for sleep data visualizations on fitness trackers, we investigated which type of sleep data people collect, and use as well as the types of visualizations they would prefer. To reach a wide audience we designed an online survey based on a similar survey published as part of a EuroVis Poster [5]. Our questionnaire updated answer choices to make the types of visualizations asked about more consistent, and included a randomization of answer choices, strengthening the validity of the results. The survey material is available in the supplementary material–https://osf.io/f3vja/.

3.1 Design and Analysis

We deployed our questionnaire using Google Forms. The first part of the survey included questions about the respondents' fitness tracker, and how they analyzed their sleep data. The main part of the questionnaire consisted of questions in which participants had to select a preferred visualization out of a set of four choices for a specific type of sleep data granularity (previous night, previous week, previous month, comparison to global values), and data type (sleep duration versus sleep phases). Each group of images (duration and phases) was preceded by an open-ended question, in which participants described what they would like to learn about their sleep duration and phases, respectively. The purpose of this question was to put participants in a mindset to answer the follow-up questions in the context of their own sleep data. The visualizations shown after were inspired by those found on common commercial fitness trackers and smartphone apps. The supplementary material gives additional information about our real-world inspirations

Table 1: Sleep data that respondents would like to see on their tracker. The color-coding represents the different categories, and is a visual encoding of the percentage of answers (the higher the opacity the higher the percentage of responses).

	Last Night	Weekly Overview	Monthly Overview	Social Comparison	Not Interested
Phases	88 (81%)	41 (38%)	27 (25%)		10 (9%)
Duration	87 (80%)	60 (56%)	40 (37%)	19 (18%)	4 (4%)
Schedule	44 (41%)	53 (49%)	32 (30%)	9 (8%)	30 (28%)
Quality	87 (80%)	57 (53%)	39 (36%)	18 (17%)	9 (8%)
Metadata	88 (81%)	47 (46%)	28 (26%)	15 (14%)	13 (12%)

and chosen encodings. In addition to having common visualizations for each data type and time granularity, we always included one option that showed the data using text. We shuffled the order of visualizations for each participant taking the survey. Table 2 shows all visualizations. After each choice, participants responded whether they based their preference on the amount of data shown, the design, or both.

We open-coded the free response questions, and iteratively derived answer categories. We analyzed the fixed-choice questions using Tableau.

3.2 Procedure

After reading the consent form, and agreeing to participate in the survey, we asked participants to validate that they own a smartwatch, or fitness band, which can track sleep data, and that they look at this data. If not, we excluded them from the survey. The accepted participants proceeded to answer the remaining questions. Based on the answer they gave regarding their used device (smartwatch, or fitness band) they saw the visualizations adapted to this form factor: Square ▮, or Wide ▬ (Table 2).

3.3 Results

We recruited 132 participants via Reddit, Facebook, Twitter, and personal emails. For the analysis, we included 108 valid responses, and excluded 24 people who reported not wearing a tracker, or not checking their sleep data. It took participants between 10, and 15 minutes to answer the questionnaire.

Most (86% ▬▬) of the respondents reported wearing their fitness tracker every night while sleeping. Most participants mentioned checking their sleep data only on the phone app (80.6% ▬▬), while 4.6% ▭ checked only their fitness tracker, and 14.6% ▭ both devices.

Table 1 summarizes the answers participants gave regarding the sleep data they would like to see independent of what their current device could show. Most people were interested in data about their last night's sleep phases, duration, quality, and other metadata—all of interest to more than 80% ▬▬ of respondents. Weekly sleep data was the second most preferred type but only few participants were interested in a social comparison of sleep data (≤18% ▭).

Table 2 summarizes the percentages of preferred visualizations for different time granularities of the sleep phases, and sleep duration while the teaser (Figure 1) shows the most preferred visualizations in more detail. The differences between the two top ranked visualizations for weekly sleep duration were so low that we considered them both as the winner. Table 2 shows the data separated for Square ▪ and Wide ▭ to see differences between people using a smartwatch, and those wearing a fitness band. Across all 8 groups of visualizations, the preferred visualization was mostly (5/8) the same for both types of devices. Overall, the most preferred visualization was often an area chart (Square ▪ = 4/8, Wide ▭ = 3/8) rated first by approximately 40–50% of respondents. Bar chart-based representations were also popular, in particular for the wide form factor (Square ▪ = 2/8, Wide ▭ = 4/8). A clear winner was the hypnogram (76% ▬▭) for showing last night's sleep phases, and the horizontal bars (73% ▬▭) for the social comparison of last night's sleep phases on the wide version. Only for social comparison of sleep duration, people preferred the text version to other visualizations. In most cases, the text representation had the lowest, or second to lowest rating.

Note that while the data visualizations we showed to people were similar in color coding, the explicitly encoded information varied between designs. To understand whether people chose their "preferred" representation based on the data shown or the encoding, we asked an additional question in the survey, in which participants explained how they made their choice. On average, 49% ▬▭ of the participants reported choosing the representation considering both the displayed data and the corresponding encoding. The preference information we gathered from the survey can help to inspire further studies, in which preferences based on data or visualization can be teased apart. Our results show that participants chose those visualizations that allowed them to see and infer more information than others in most cases. One clear outlier is the text representation of the comparison of social sleep duration, which shows comparatively little information as text.

4 STUDIES COMPARING SLEEP VISUALIZATIONS

Based on our survey, we found that for five out of eight categories of sleep data, participants preferred the same type of visualization for both smartwatch, and wristband form factors. Therefore, we wanted to investigate how changes in form factor would affect the effectiveness of a visualization, and if one could expect to use the same representation type for both form factors. The main research question we investigate in the following three studies was: *How does display form factor, expressed as display size, and orientation, affect the effectiveness with which people can answer questions about sleep data?*

4.1 General Method

To investigate our research question, we conducted three studies: one in-person pilot study using a real smartwatch, and two studies conducted online as crowdsourced studies using peoples' own smartphones. For all three studies, we used the same method, measuring both completion time, and accuracy of participants with a specific task, and visualization stimulus. In each study, and task,

participants had a forced-choice between two answer options; therefore, we used a two-alternative forced choice (2AFC) design with a Type 1 procedure (yes/no-responses) [37]. For each trial, participants saw the stimulus, and answered the task by pushing one of two buttons; then they received feedback for 1000 ms, followed by the next stimulus visualization. There was no time limit to answer, but we instructed participants to answer as accurate as possible.

4.2 Stimuli

Our stimuli used in all three studies had a display size of 320 px × 320 px (smartwatch; Square ▪), 320 px × 160 px (horizontal wristband; Wide ▭), and 160 px × 320 px (vertical wristband; Tall ▯).

4.3 Data Analysis and Interpretation

To analyze the data we conducted a time, and error analysis. For all three studies, we calculated the sample mean per person, and condition for both time, and accuracy. We represent the sample mean using interval estimation with 95%-confidence intervals, which we adjusted for multiple comparisons using Bonferroni corrections [31]. We construct the confidence intervals using BCa bootstrapping (10,000 bootstrap iterations). Therefore, we can be 95% certain that the population mean is included within this given interval. We interpret the difference between two values using estimation techniques, and give the strength of the evidence about the population means as recommended in the literature [7, 8, 23, 24, 26]. A confidence interval of mean differences shows evidence if the confidence interval does not overlap with 0, which corresponds to a statistically significant result using p-value tests. The farther away from 0, and the tighter the confidence interval is, the stronger is our evidence. One can calculate equivalent p-values using the method by Krzywinski, and Altman [39].

5 IN-PERSON PILOT STUDY: BAR CHARTS OF WEEKLY SLEEP DURATIONS

As a first step, we conducted an in-person pilot study in the lab using an actual smartwatch. For this in-person pilot study, we focused on visualizations representing sleep duration because participants from the survey were most interested to see sleep duration, on average, of all temporal granularities. We focused on weekly sleep data instead of last night's sleep as a richer source of data. In addition, Cai et al. [17] have already tested the winning donut chart for last night's sleep patterns under varying display sizes, showing no negative effect at smaller display sizes.

5.1 Stimuli and Tasks

Representation Choice: The most preferred visualizations for a weekly overview of sleep durations were the area chart, and floating bar chart. We decided to use the floating bar chart, because it is the more common visualization technique for this data, and because it can encode more information including schedule, and sleep consistency. Moreover, it might be more affected by changes in tracker type due to the compression of bar width/length needed on the fitness band. A floating bar chart is a non-aligned bar chart, representing chosen preferred bedtimes, and wake-up times, i.e., the time when a person actually went to sleep, and woke up in the morning (Figure 2).

Table 2: Visualizations of sleep phases and sleep duration for smartwatch and fitness bands. Orange borders highlight the most preferred combinations of data and designs for each time granularity as well as for social comparisons.

Task Choice: The floating bar chart allowed us to add more tasks that are aligned to what respondents in the questionnaire wanted to learn from their sleep data. The tasks we investigated were:

- **T1:** On which day did you sleep longer: Saturday or Sunday?
- **T2:** Did you go to bed later than planned (22:00) on 4 or more days this week?
- **T3:** Did you sleep longer on average on the weekend days (Sat, Sun) compared to the weekdays (Mon-Fri)?

Common visualization tasks [15] inspired our tasks, which required participants to analyze temporal data. T1 is an elementary task, which a participant can answer by identifying two bars (Saturday, and Sunday), and comparing their ranges in the reference set; T2, and T3 are synoptic tasks, which require participants to gain an overview of the whole reference set [4]. For T2, participants needed to summarize the set by counting the bars in relation to a reference line (22:00). T3 required participants to identify two reference subsets (weekdays, and weekend), mentally aggregate bar length for each subset, and compare the aggregates.
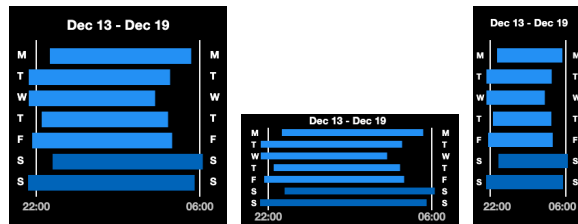
Figure 2: Stimuli for the study representing weekly sleep duration with three different form factors: Square, Wide, and Tall. Sizes, when printed without scaling, correspond to the physical sizes shown during the study.

Data Generation: To generate the data used for the visualization stimuli, we generated values for sleep duration, and schedule for each night of the week. Each bar represented a sleep duration between 370, and 501 min of sleep that started up to 30 min before to 50 min after the preferred bedtime of 22:00. For tasks that involved comparing bars, or the average of bars, we ensured that there was a difference of a least 8 px, which corresponded to 16 min of sleep duration. We colored the bars in two shades of blue to distinguish between weekdays, and weekend days. Stimuli of all three form factors had the same information, and text.

5.2 Participants

We recruited 18 participants (10 female, 8 male; 4 researchers, 14 students) with an average age of 29.27 years (SD = 6.76). Their highest degree was Bachelor (2), Master (12), or Ph.D. (4). They all had a background in HCI. Out of 18 participants, 17 reported having experience with visualizations. All participants had normal, or corrected-to-normal vision, and 2 of them reported to have a color vision deficiency. One participant owned a fitness tracker (Apple Watch), and two participants owned a fitness band (Fitbit). We compensated participants with chocolates, and tea at the end of the study.

5.3 Procedure and Apparatus

We used a within-subject design, and counterbalanced the order of the tasks, and the form factors using a Latin square. We used a Sony Smartwatch 3 with the Android Wear 2.8.0 operating system. The smartwatch's screen dimensions were 28.73 mm × 28.73 mm with a resolution of 320 px × 320 px. We attached the smartwatch to a stand at an angle of 50 degrees [11]. We adjusted the stand so that the smartwatch was placed at 20 cm height from the table surface, and roughly 28 cm viewing distance from the participant [10]. We allowed participants to adjust their seating position during the study. To familiarize themselves with the procedure, participants did 15 practice trials, followed by a 10 s break, and 30 actual trials.

Participants answered tasks by pressing one of the four arrow keys depending on the task (up/down for T1, left/right for T2, and T3) on the keyboard placed in front of them, and below the smartwatch stand. We used a Macbook Pro laptop to run a Java program, which recorded input (key presses) of participants, saved logs, and sent feedback to the smartwatch app implemented using Android programming. We connected the smartwatch, and laptop to the same WiFi hotspot, and they communicated with each other via TCP sockets.

5.4 Results

In the following, we discuss the results (both time, and accuracy) of the in-person pilot study separated by task using confidence interval estimation techniques as described in Section 4.3. Table 3, and Table 4 show the detailed results including averages, and confidence intervals.

5.4.1 Task T1: On which day did you sleep longer: Saturday or Sunday?

Completion Time: Overall, we saw that participants were slower with Wide ▬ compared to the other two form factors. We have no evidence of a difference between Tall ▮ and Square ▪. Completion times ranged on average between 766 ms–970 ms.

Accuracy: Accuracy was almost 100% for all form factors, and there was no evidence of a difference between the conditions.

5.4.2 Task T2: Did you go to bed later than planned (22:00) on 4 or more days this week?

Completion Time: Square ▪ and Wide ▬ had almost the same completion time with 1075 ms and 1074 ms. We have evidence for Tall ▮ to be slower than the other two form factors with a completion time of 1424 ms.

Accuracy: The accuracy in this task was only slightly lower than for T1 with 97–98%. We again see no evidence of a difference between form factors.

5.4.3 Task T3: Did you sleep longer on average on the weekend days (Sat, Sun) compared to the weekdays (Mon-Fri)?

Completion Time: We have weak evidence of a difference between Square ▪ (faster with 1006 ms on average), and Wide ▬ (slower with 1219 ms on average), but no evidence of a difference between the other form factors.

Accuracy: T3 had the highest error rate, but accuracy was still above 95% for each form factor. Again, we see no strong evidence of a difference between form factors, but there may be a trend for Tall ▮ to be more accurate than Wide ▬.

Summary: Accuracy was high in all three tasks (over 95%), and we saw no strong evidence of a difference across form factors according to how many errors participants made. Some differences were, however, visible according to completion time. For T1, Wide ▬ was slower than the other two. For T2, Tall ▮ was slower than the other two. In T3, we saw a potential trend for Wide ▬ being slower than Square ▪.

6 SLEEP DURATION (BAR) STUDY: BAR CHARTS OF WEEKLY SLEEP DURATIONS

After conducting the in-person pilot study, we re-ran the pilot as a crowdsourced study, during which participants had to use a smartphone. A crowdsourced study allowed us to access a broader pool of participants but with the tradeoff that we could no longer use

Table 3: *Completion Time* analysis of data from the Pilot Sleep Duration (bar) Study. Left: average completion time in milliseconds. Right: pairwise comparisons for each task, and form factor. Error bars represent 95% Bootstrap confidence intervals (CIs) in black, adjusted for three pairwise comparisons with Bonferroni correction (in red).

Task	Form factor	Average	Comparison	Difference
T1	square	766 ms [712, 823]	square – wide	−205 ms [-509, -105]
	wide	970 ms [878, 1189]	tall – wide	−190 ms [-467, -79]
	tall	781 ms [717, 861]	tall – square	15 ms [-47, 63]
T2	square	1075 ms [952, 1209]	square – wide	1 ms [-220, 186]
	wide	1074 ms [939, 1298]	tall – wide	351 ms [67, 682]
	tall	1424 ms [1214, 1715]	tall – square	350 ms [169, 553]
T3	square	1006 ms [901, 1134]	square – wide	−213 ms [-517, -17]
	wide	1219 ms [1064, 1540]	tall – wide	−122 ms [-470, 80]
	tall	1097 ms [979, 1268]	tall – square	91 ms [-69, 296]

Table 4: *Accuracy* analysis of data from the Pilot Sleep Duration (bar) Study. Left: average accuracy; right: pairwise comparisons for each task, and form factor. Error bars represent 95% bootstrap confidence intervals (CIs) in black, adjusted for three pairwise comparisons with Bonferroni correction (in red).
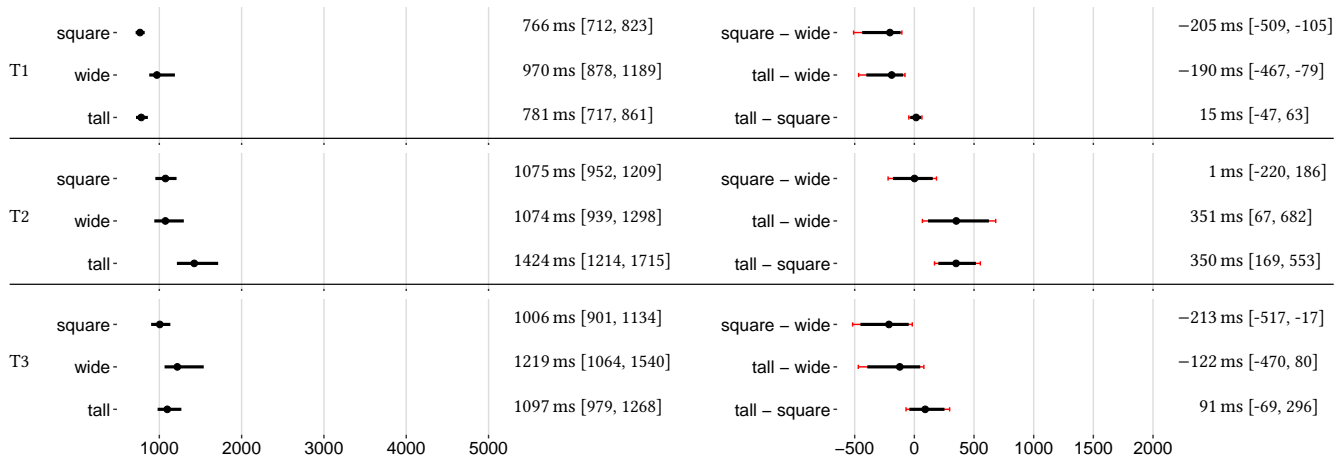
Task	Form factor	Average	Comparison	Difference
T1	square	99% [98, 100]	square – wide	1% [-1, 2]
	wide	99% [97, 99]	tall – wide	0% [-2, 2]
	tall	99% [98, 100]	tall – square	0% [-2, 1]
T2	square	98% [96, 99]	square – wide	0% [-1, 2]
	wide	98% [94, 99]	tall – wide	-1% [-4, -4]
	tall	97% [94, 98]	tall – square	-1% [-3, 2]
T3	square	97% [93, 98]	square – wide	2% [-2, 16]
	wide	95% [84, 98]	tall – wide	3% [0, 17]
	tall	98% [95, 99]	tall – square	1% [-2, 5]
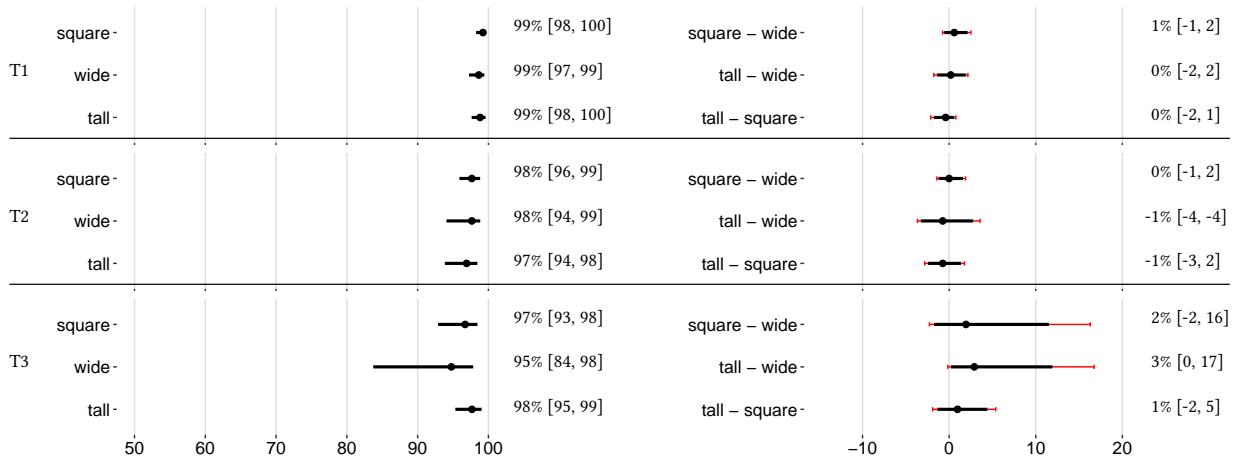
smartwatches to deploy the study. Our goal was to find out whether we could reach similar results when simulating smartwatches on smartphones in a crowdsourced setting. Due to the ongoing COVID-19 pandemic in-person studies are harder to conduct, and our results have the potential to inform the design of future perception studies for smartwatches.

6.1 Design Specifics

We used the same study design as for the in-person pilot study described in Section 5. The main difference between these two studies was that we used a between-subject design to reduce the study time per participant. The between-subjects factor was the task, so each participant saw one task but all three form factors. We preregistered the study at https://osf.io/yz8ar/.

To approximate the pixel density of smartwatches, we designed the study to be run on smartphones using a mobile web browser. We used a framework for running online studies [43] into which stimuli images were loaded. This ensured that images were displayed at the original aspect ratio of the Sony Smartwatch 3 used in the in-person pilot study. We used a display size of 213 × 213 density-independent pixels (DIP). Measuring in DIP allowed us to ensure that stimuli appear at the same physical size across screens, no matter what density those screens have. We followed the viewport size conversion of the Android developer community [25].

Participants could only participate in the study if they had a smartphone (e.g., Android mobile, iPhone) with a minimum screen resolution of 320 px × 480 px, had a web browser installed on their phone, and had a proper internet connection. We designed our

study so that at the minimum screen resolution participants did not have to scroll the web page during the study. After finishing each form factor, we asked participants about their confidence when performing the task on a 5-point Likert scale (5 = completely confident, 4 = fairly confident, 3 = somewhat confident, 2 = slightly confident, 1 = not confident at all). A post-questionnaire followed after performing the task with all three form factors. There were no time constraints for giving their answers but we asked participants to answer as quickly, and accurately as possible.

Training: Participants saw 15 random visualizations as training for each form factor. There were detailed instructions before each training describing the next type of visualization shown.

Study Trials: In total, participants completed 34 trials per form factor, with 30 trials in random order, and 4 trials as attention checks per section for a total of 102 trials per participant. We collected the given answer per trial, the correct answer, and the time taken to answer.

Exclusion Criteria: We excluded participants who reported having problems with the consent form, who reloaded the web browser, who did not finish the complete study, whose monitor specifications did not meet the study requirements, or who failed >=50% of the attention check trials. We clearly stated the exclusion criteria at the beginning of the study, notified participants if they were not able to complete the study, and terminated the study early. We also included two additional attention check questions at the end of the study asking participants to select the task they had just completed as well as which type of data they had seen during the study (sleep, as well as rainfall, or weather as distractors). If a participant failed to answer these questions, then the study was immediately terminated.

6.2 Participants

We recruited fluent English speakers using Prolific [55]. From the 175 total responses, 36 were incomplete, 2 failed the attention check questions, and 2 participated multiple times, and we discarded them from all our analyses. In total 135 participants completed our study. We compensated participants who completed the study with a reward of £2.63 for our 17-minute estimated study completion time, slightly higher than the French minimum wage requirements requested by our ethics committee. For T1, there were 39 participants, 42 participants for T2, and 54 participants for T3. It is difficult in crowdsourced studies to achieve an equal number of participants per condition but because we do not compare across tasks, the difference in participant numbers does not bias our findings.

Among participants, 45.9% ▭ were female, and 54.1% ▭ were male; 48.12% ▭ reported to be students, the rest (51.88%) did not provide their job status (Table 5). All participants had normal, or corrected-to-normal vision, and the average age was 26.43 years (SD = 7.53). Among all participants, 27.06% ▭ owned a smartwatch, 29.32% ▭ owned a fitness band, and 6.01% ▭ owned both. When asked to report on their familiarity with floating bar charts on a Likert scale (Figure 3), most participants (51.9% ▭) reported reading these charts occasionally.

6.3 Results

We describe the results per task discussing completion time, accuracy, and confidence. We separate our discussion by task due to our between-subject study design (each participant saw all form factors for one task). For the data analysis, we only consider the data that was within two standard deviations of the mean answering time. We removed as outliers around 5% of the data across all tasks mostly due to a problem with our timer implementation: the first trial of each form factor took an unexpectedly long time because the time counter included page loading time. Therefore, we removed the first trial across all tasks but fixed this problem for our subsequent crowdsourced study reported in the next section. Table 6 summarizes the results on completion time per task while Table 7 shows the detailed results of the error analysis.

6.3.1 Task T1: On which day did you sleep longer: Saturday or Sunday?

Completion Time: Participants were similarly fast with all three form factors. The average of Tall ▯ was fastest with 1300 ms, followed by Square ▪ with 1367 ms, and Wide ▬ with 1445 ms. Pairwise comparisons show no evidence of a difference between the three form factors because the corrected confidence intervals touch zero. It is nevertheless possible that there is a trend for Wide ▬ being slower than the other two because the uncorrected CIs either do not cross, or are close to 0.

Accuracy: In this task, correctness was high for all representations (all 99% correct). We do not have evidence of a difference between sizes.

Confidence: For T1 participants reported largely high confidence scores for all three form factors. Over 85% of respondents were at least fairly confident with each form factor: Square ▪ = 95% ▭ , Tall ▯ = 92% ▭ , Wide ▬ = 87% ▭ (Figure 4 top).

6.3.2 Task T2: Did you go to bed later than planned (22:00) on 4 or more days this week?

Completion Time: Answer times ranged between 1547 ms (Wide ▬) to 1933 ms (Tall ▯), and there is strong evidence for Wide ▬ being faster than Tall ▯. We did not see strong evidence of a difference between Square ▪ (with 1696 ms on average), and the other two form factors. Nevertheless, the uncorrected pairwise confidence intervals involving Square ▪ are close to 0 that may indicate a trend of a difference between Square ▪ and the other two form factors.

Accuracy: With each form factor participants were 96–97% correct. We have only some evidence that participants may be more correct with Wide ▬ compared to Tall ▯.

Confidence: Participants reported high confidence scores for all form factors. For Square ▪ and Wide ▬, 83.3% ▭ of participants were at least fairly confident, and for Tall ▯ 90.5% ▭ (Figure 4 middle).

6.3.3 Task T3: Did you sleep longer on average on the weekend days (Sat, Sun) compared to the weekdays (Mon-Fri)?

Table 5: Details about participants for the two crowdsourced studies. Number of participants (P), location, gender (F: Female, M: Male), age (M: mean, SD: standard deviation), percentage of students, percentage of participants who reported to own a smartwatch (SW), and percentage of participants who reported to own a fitness band (FB).

Study	P	Location	Gender (F/M)	Age (M/SD)	Students	SW	FB
Sleep Duration (bar)	135		(46.7% / 53.3%)	(26.56 / 7.7)	47.4%	26.6%	29.3%
Sleep Phase (hypnogram)	117		(58.1% / 41.9%)	(26.27 / 7.02)	55.5%	33.3%	24.7%
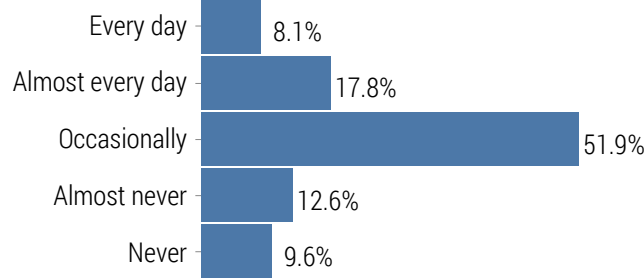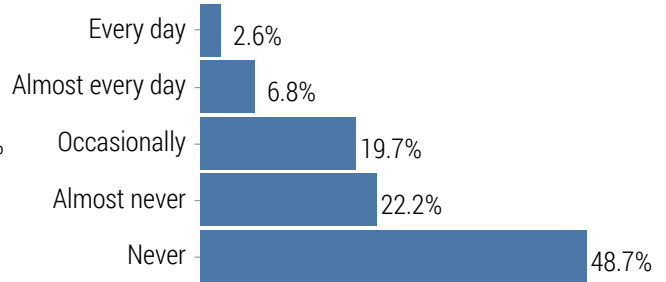
Figure 3: Analysis of demographic data from the Sleep Duration (bar) Study, and the Sleep Phase (hypnogram) Study. Left: familiarity rating of participants reading floating bar charts. Right: familiarity rating of participants reading hypnogram charts.

Table 6: *Completion Time* **analysis of data from the Sleep Duration (bar) Study. Left: average completion time in milliseconds for each task, and form factor. Right: pairwise comparisons for each task, and form factor. Error bars represent 95% Bootstrap confidence intervals (CIs) in black, adjusted for three pairwise comparisons with Bonferroni correction (in red).**

Table 7: *Accuracy* analysis of data from the Sleep Duration (bar) Study. Left: average accuracy for each task, and form factor. Right: pairwise comparisons for each task, and form factor. Error bars represent 95% Bootstrap confidence intervals (CIs) in black, adjusted for three pairwise comparisons with Bonferroni correction (in red).

		Accuracy		Pairwise comparison
T1	square	99% [97, 99]	square – wide	0.07% [-1.1, 1.0]
	wide	99% [98, 99]	tall – wide	0.3% [-0.9, 1.2]
	tall	99% [98, 99]	tall – square	0.2% [-0.7, 1.5]
T2	square	97% [95, 97]	square – wide	-0.7% [-2.2, 0.4]
	wide	97% [96, 98]	tall – wide	-1.7% [-4.1, -0.1]
	tall	96% [94, 97]	tall – square	-0.9% [-3.5, 0.5]
T3	square	96% [95, 97]	square – wide	2.7% [0.5, 4.9]
	wide	93% [92, 95]	tall – wide	1.3% [-1.1, 3.8]
	tall	95% [93, 96]	tall – square	-1.4% [-3.5, 0.4]
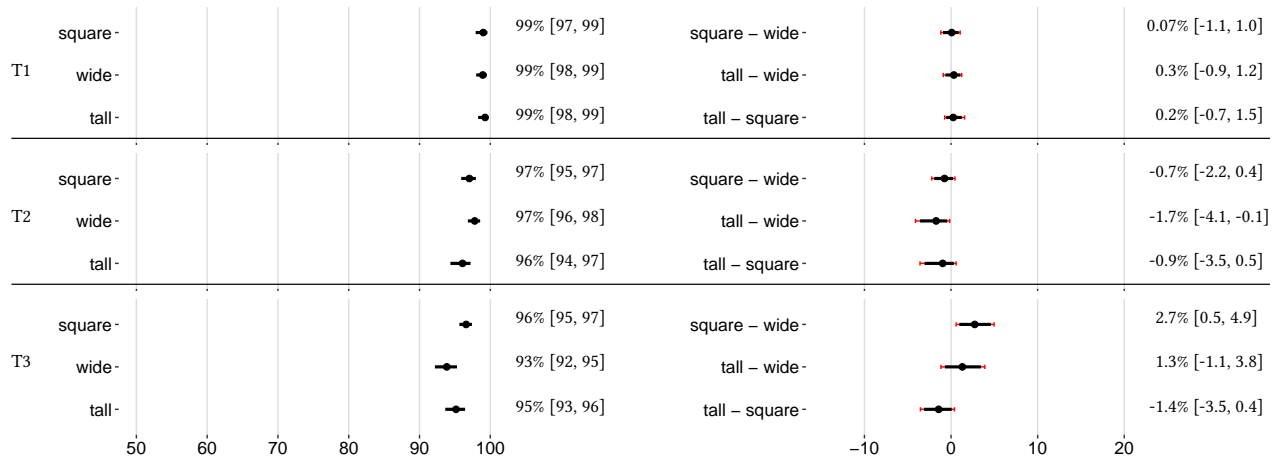
Completion Time: We saw no evidence of a difference between form factors for this task, and average completion times were similar: Tall ▮ = 1780 ms; Wide ▬ = 1857 ms, and Square ◼ = 1905 ms.

Accuracy: Correctness for this task ranged between 93% (Wide ▬), and 96% (Square ◼).

However, we have only evidence of a difference between Square ◼ (more accurate), and Wide ▬.

There is possibly a trend for Square ◼ to be more accurate than Tall ▮ as well but the CI is close to 0.

Confidence: Participants reported the highest confidence scores for Square ◼, with 89% ▆▆▆▫ being at least fairly confident. The other confidence scores of participants reporting to be at least fairly confident in their answers also remained high with 72% ▆▆▫▫ (Wide ▬), and 74% ▆▆▆▫ (Tall ▮) (Figure 4 bottom).

6.3.4 *Post-Study Questionnaire.* We asked participants to pick their preferred display form factor. For each task Square ◼ was the preferred representation (T1: 69.2% ▆▆▆▫ , T2: 73.8% ▆▆▆▫ , T3: 57.4% ▆▆▫▫). When asked to compare the two wristband form factors Wide ▬ was preferred for all tasks (T1: 51.3% ▆▆▫▫ , T2: 59.5% ▆▆▫▫ , T3: 53.7% ▆▆▫▫) but the preferences for one, or the other were not strong.

Summary: The most obvious difference between form factors in this study was present in T2, for which Wide ▬ was both faster, and more accurate than Tall ▮. Differences in time in the other two tasks are not as prominent as in the in-person pilot study, but there are trends in T1 for Wide ▬ to be slower than Square ◼ and Tall ▮. Accuracy was overall high (over 93%). We saw that in T3 Square ◼ was more accurate than Wide ▬ possibly more accurate than Tall ▮ and participants felt more confident with it. We reflect more on our methodological approach of conducting a crowdsourced study, and differences to the in-person pilot study, in Section 8.

7 SLEEP PHASE (HYPNOGRAM) STUDY: HYPNOGRAM CHARTS OF NIGHTLY SLEEP PHASES

After we identified that crowdsourced studies can help uncover insights on smartwatch perception, we expanded our exploration about whether people can use charts effectively across different form factors. Again, we focused our study on insights into sleep data, and their visualization. In this second study, we chose last night's sleep phases as one of the data sources of most interest to wearers in our survey, and because its preferred visualization was surprising to us. In particular, we were surprised by the strong preference (75.8% ▆▆▆▫) for the hypnogram chart on the wristband form factor because it is a relatively complex, and visually busy temporal chart, and we were skeptical how well it would perform in practice. In comparison, only 33.3% ▆▫▫▫ of survey respondents had picked the hypnogram as preferred for the smartwatch-sized visualization. We followed the same design procedure as in our first crowdsourced study.

7.1 Stimuli and Tasks

A hypnogram represents the stages of sleep as a function of time. The purpose of a hypnogram is to provide a general sense of nightly sleep behaviors. It differentiates between different stages of sleep on the *y*-axis: rapid eye movement sleep (REM), and non-rapid eye movement sleep (NREM), sometimes separated into different levels like light, and deep sleep, during the sleep cycle [30]. A typical hypnogram on fitness trackers shows three sleep stages: light, deep, and REM together with the time a person is awake [16]. The hypnogram shows when someone transitioned from one stage to another along the *x*-axis.

Participants in our survey reported to be interested in learning about their sleep phases. We formulated the following three tasks to target this interest, and to remain close to the types of tasks of our Sleep Duration (bar) Study:
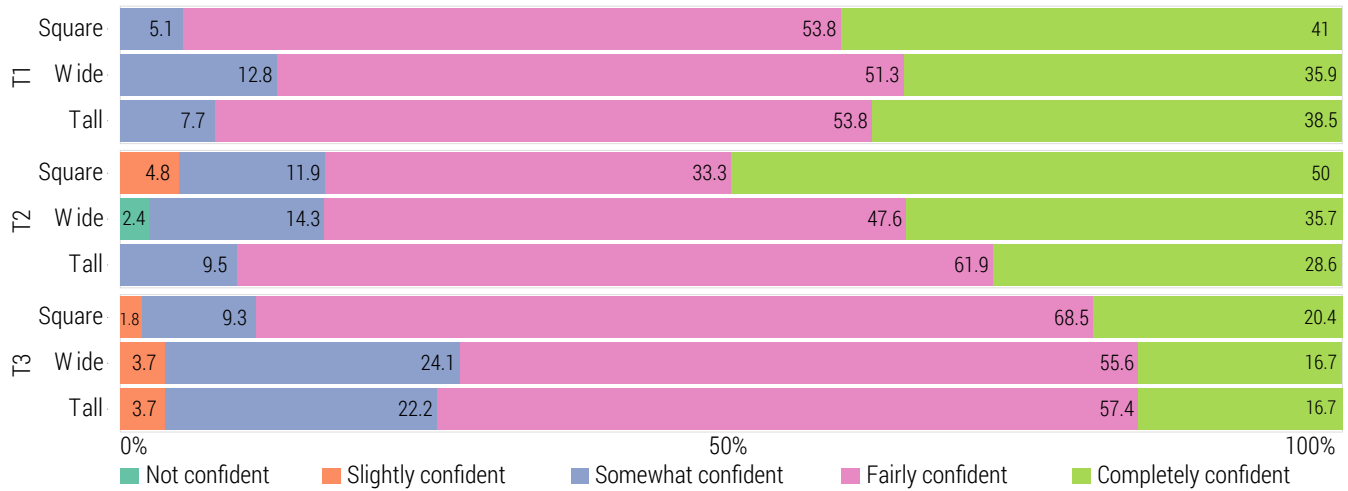
Figure 4: Percentages of confidence ratings by participants of different form factors for each task for the Sleep Duration (bar) Study.

- **T1:** Were you in the wake phase more than or equal to 4 times?
- **T2:** Did you have 4 or more transitions from REM to light sleep?
- **T3:** Did you spend more time in REM than in deep sleep?

All three tasks required participants to analyze temporal data. While T1 is an elementary task that participants can answer by counting bar peaks, T2, and T3 are synoptic tasks, which require participants to gain an overview of the whole reference set [4]. In T2, participants needed to count patterns related to a reference set from two sleep stages. T3 requires participants to identify two reference subsets (REM, and deep sleep), calculate the aggregate bar length for each subset, and compare the aggregates.

To generate synthetic data used for the stimuli, we calculated values for the four different sleep stages. On a typical night, a person goes through four to six sleep cycles. Not all sleep cycles are of the same length, but on average, they last about 90 minutes each [53]. For the stimuli, as shown in Figure 5, we used five sleep cycles, and 96 min each for an 8-hour sleep duration (22:00-06:00). For the task that involved comparing bars of the aggregates, we ensured that there was a difference of a least 8 px. Stimuli of all three form factors had the same information, and text. We colored each sleep phase differently: red-colored bars represent the *wake* sleep phase, blue-colored bars for the *REM* sleep phase, green-colored bars for the *light* sleep phase, and purple-colored bars for the *deep* sleep phase. We used the color palette Sciences Po medialab [47] generated considering color blindness. We designed the visualizations to be similar to those already found on fitness tracking apps.

7.2 Participants

We recruited all participants in the same way as for our first crowd-sourced study via Prolific [55]. Of the 162 total responses, 45 were incomplete, and discarded from all our analyses leaving us with 117 participants in total to complete our study. We compensated participants who completed the study with a reward of £2.63 for our

17-minute estimated study completion time, slightly higher than the French minimum wage requirements requested by our ethics committee. For T1, there were 45 participants, 36 participants for T2, and 36 participants for T3. We recruited 58.1% ▬▭ female, and 41.9% ▬▭ male participants; 55.5% ▬▭ of the participants reported to be students, the job status of the rest (44.5%) is unknown. All participants had normal, or corrected-to-normal vision, and the average age was 26.27 years, (SD = 7.02). Among all participants, 33.3% ▬▭ owned a smartwatch, 24.7% ▬▭ owned a fitness band, and 6.83% ▭ owned both. Participants reported not to read hypnogram charts often: Never (48.7% ▬▭) was the highest answer when rating hypnogram chart familiarity on a 5-point Likert scale. Figure 3 right, and Table 5 summarize background information about the participants.

7.3 Results

We describe the results discussing completion time, accuracy, and confidence. We separate our discussion by task due to our between-subject design. For the data analysis, we only considered the data that was within two standard deviations of the mean answering
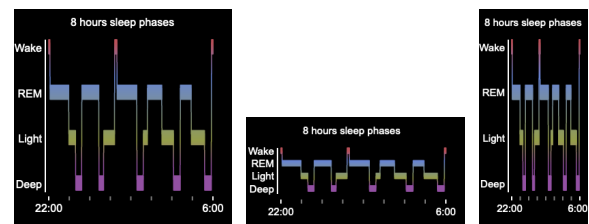

Figure 5: Stimuli for the Sleep Phase (hypnogram) Study representing nightly sleep phases (wake, REM, light, deep) using three different form factors: Square, Wide, and Tall. Sizes, when printed without scaling, correspond to the physical sizes shown during the study.

Table 8: *Completion Time* analysis of data from the Sleep Phase (hypnogram) Study. Left: average completion time in milliseconds for each task; right: pairwise comparisons for each task, and form factor. Error bars represent 95% Bootstrap confidence intervals (CIs) in black, adjusted for three pairwise comparisons with Bonferroni correction (in red).

			(left) average		pairwise	(right) comparison
T1	square		1030 ms [948, 1173]	square – wide		−15 ms [−97, 96]
	wide		1045 ms [985, 1138]	tall – wide		−75 ms [−183, 2]
	tall		970 ms [917, 1029]	tall – square		−60 ms [−227, 44]
T2	square		3081 ms [2827, 3357]	square – wide		210 ms [−129, 526]
	wide		2871 ms [2593, 3157]	tall – wide		1357 ms [817, 2018]
	tall		4228 ms [3600, 4928]	tall – square		1147 ms [491, 1777]
T3	square		1250 ms [1145, 1423]	square – wide		55 ms [−122, 209]
	wide		1194 ms [1088, 1327]	tall – wide		16 ms [−121, 164]
	tall		1211 ms [1123, 1309]	tall – square		−38 ms [−199, 100]

Table 9: *Accuracy* analysis of data from the Sleep Phase (hypnogram) Study. Left: average accuracy; right: pairwise comparisons for each task, and form factor. Error bars represent 95% Bootstrap confidence intervals (CIs) in black, adjusted for three pairwise comparisons with Bonferroni correction (in red).
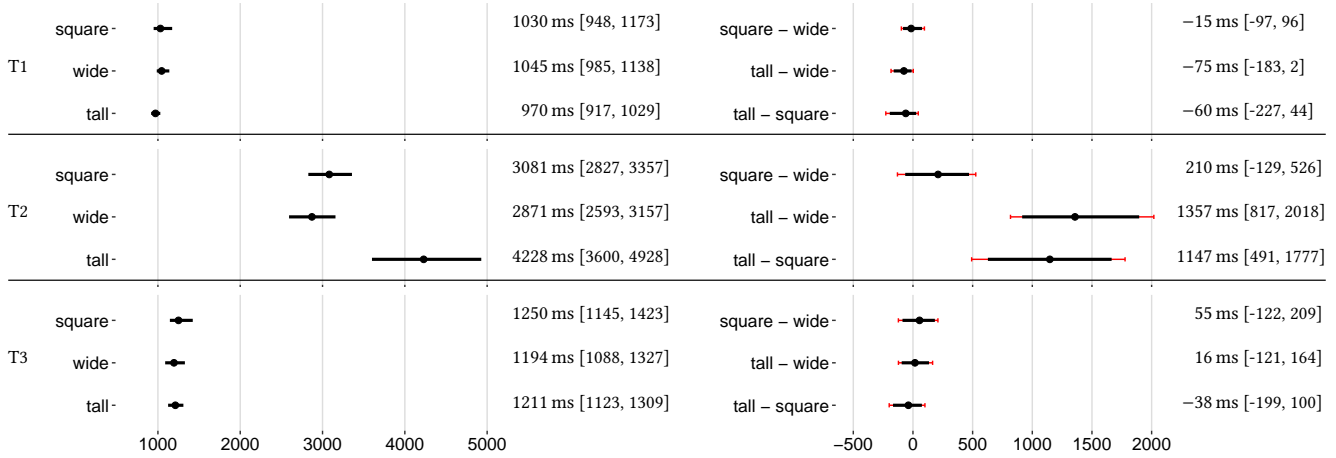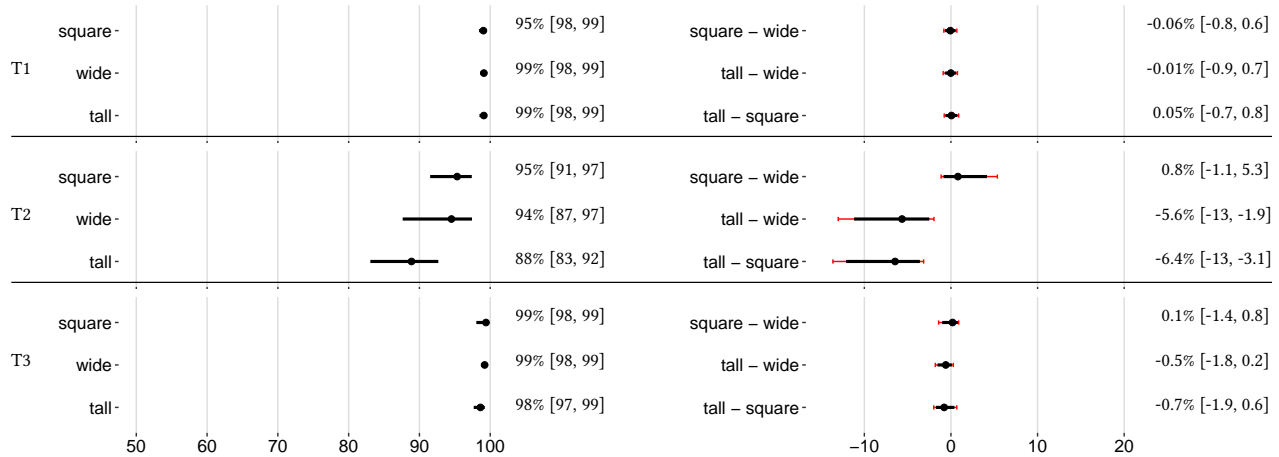
			(left) average		pairwise	(right) comparison
T1	square		95% [98, 99]	square – wide		-0.06% [-0.8, 0.6]
	wide		99% [98, 99]	tall – wide		-0.01% [-0.9, 0.7]
	tall		99% [98, 99]	tall – square		0.05% [-0.7, 0.8]
T2	square		95% [91, 97]	square – wide		0.8% [-1.1, 5.3]
	wide		94% [87, 97]	tall – wide		-5.6% [-13, -1.9]
	tall		88% [83, 92]	tall – square		-6.4% [-13, -3.1]
T3	square		99% [98, 99]	square – wide		0.1% [-1.4, 0.8]
	wide		99% [98, 99]	tall – wide		-0.5% [-1.8, 0.2]
	tall		98% [97, 99]	tall – square		-0.7% [-1.9, 0.6]

time. This means we almost uniformly removed around 5% of the data across all tasks. Table 8, and Table 9 give numerical details about the time, and accuracy results.

7.3.1 Task T1: Were you in the Wake phase more than or equal to 4 times?

Completion Time: The average completion times for Tall ▌ (970 ms), Square ▇ (1030 ms), and Wide ▅ (1045 ms) were similar. We saw no strong evidence for differences in the pairwise comparisons of techniques but there may be a trend for Tall ▌ to be slightly faster than Wide ▅.

Accuracy: Correctness was high for all representations (all 99% correct on average). We do not have evidence of a difference between form factors.

Confidence: Figure 6 top shows the confidence for all three form factors for T1. There were no participants who reported not to be confident, and confidence ratings overall were high. Most people with the Square ▇ form factor reported to be completely confident (73.3% ▇▭), followed by Wide ▅ (64.4% ▇▭), and Tall ▌ (57.8% ▇▭).

7.3.2 Task T2: Did you have 4 or more transitions from REM to light sleep?

Completion Time: T2 had the largest average completion times in our study with 2871 ms for Wide ▅ up to 4228 ms for Tall ▌. In the pairwise comparisons, we have strong evidence for this task being slower to complete with Tall ▌ than with the other two form factors (slower by over 1 s).

Accuracy: Average correctness was 94%, and 95% on average for Wide ▬ and Square ◼ respectively but dropped to 88% for Tall ▮. We have evidence that Tall ▮ was indeed less accurate than the other two form factors, but no evidence of a difference between Wide ▬ and Square ◼.

Confidence: Confidence was lowest for this task. We see the difficulties participants had with Tall ▮ confirmed in the confidence rating; 53.3% ▣ reported being only somewhat confident, or lower. Confidence ratings for Square ◼ and Wide ▬ were similar with 80.6% ▣ of respondents reporting being fairly, or completely confident.

7.3.3 Task T3: Did you spend more time in REM than in Deep sleep?

Completion Time: The average completion times for this task ranged around 1.2 s with only small differences between form factors: Wide ▬ 1194 ms, Tall ▮ 1211 ms, and Square ◼ 1250 ms. We have no evidence of a difference between form factors.

Accuracy: Participants were 98–99% correct in this task on average for all form factors. There was no evidence of a difference in the pairwise comparisons.

Confidence: The confidence scores were high for this task. Over 30% of participants reported being completely confident with each form factor, and Wide ▬ having the highest response at 47.2% ▣ . For each form factor, over 80% ▣ of respondents were at least fairly confident: 94.4% ▣ for Wide ▬, 88.9% ▣ for Square ◼, and 80.5% ▣ for Tall ▮.

7.3.4 Post-Study Questionnaire.
When asked to pick their preferred display size, participants preferred Square ◼ clearly for T1 (73.3% ▣). For T2 nobody chose Tall ▮ and the difference between Square ◼ (52.8% ▣) over Wide ▬ (47.2% ▣) was small. The pattern was similar but reversed for T3 with participants slightly preferring Wide ▬ (55.6% ▣) to Square ◼ (41.7% ▣). When asked to decide between the two wristband form factors specifically, participants showed a clear preference for Wide ▬ (T1: 86.7% ▣ , T2: 97.2% ▣ , T3: 88.9% ▣).

Summary: Most differences were visible in T2, with Tall ▮ performing worse than Square ◼ and Wide ▬ in terms of both time, and accuracy. We saw no other clear difference between the three form factors in T1, and T3. Participants tended to feel more confident with Square ◼ and Wide ▬ although confidence was high for all three form factors. These two were also the preferred form factors over Tall ▮.

8 DISCUSSION

In the following, we reflect on our study results, and the methodologies we used to study fitness tracker visualizations.

8.1 Study Results

We set out to understand how best to visualize sleep data on fitness trackers. Sleep visualizations are compelling to study because they can contain several types of data (temporal, quantitative, categorical, etc.), are relatively complex, and because many people care deeply about understanding their sleep data. From our survey, we learned that people are interested the most in short-term data (last night's sleep, or weekly sleep patterns). Surprisingly, only about a third of our respondents were interested in comparing their sleep patterns to others.

In our survey, we asked participants which types of sleep visualizations they would prefer for different types of sleep data based on the form factor of their own fitness tracker. It was interesting to see that purely text-based representations were preferred only in one case—social comparison of sleep duration data on Square ◼. More than half of our participants preferred several charts with particularly strong preferences for multiple of the Wide ▬ representations: horizontal bars for social comparison data, area charts for monthly overviews, and the hypnogram for last night's sleep phases. Due to the additional compression in the vertical direction when moving from a Square ◼ to a Wide ▬ form factor, we were surprised to see that independent participants often picked the same sleep visualizations (Figure 1) for the two form factors. The vertical compression should specifically affect vertical bar, area, or line charts, and different aspect ratios have shown in the past to affect the reading of charts [64].

Researchers can use our results to inspire the choice of dedicated sleep visualizations for fitness trackers that app developers are currently still rarely deploying on these devices. However, because the survey focused purely on visual preference we did not have evidence to recommend the use of the same types of charts on different form factors. Therefore, we conducted two follow-up studies that tested two common, and highly rated representation types for sleep duration, and sleep phase data. In our first study on sleep duration, we tested horizontal bar charts to encode weekly sleep data. We had expected that the horizontal compression of the chart, and the smaller visible differences between bars with similar start, or end positions would have a negative impact on the Tall ▮ version of the chart. Nevertheless, we included Tall ▮ as a condition because when worn it is more natural to read a wristband vertically, and wristbands often use a vertical layout in practice. Overall, we saw high accuracy with the visualizations for all tasks, and form factors. Surprisingly, the negative effect of horizontal compression only showed clearly for task completion time in Task 2, a synoptic task that required reading multiple bars, and comparing them to a reference line. Because we had designed differences to be at least 4 px for Tall ▮ and 8 px for Wide ▬ and Square ◼, the negative effect of the horizontal compression seemed to be less important than we thought. We also saw a trend for Wide ▬ being slower than the other two form factors for the elementary task that required comparing the length of two bars. This is interesting because the main difference of this form factor was the smaller bar width.

Next, we tested the hypnogram as a sleep phase visualization that our survey respondents strongly preferred for Wide ▬. We expected again that the performance for Tall ▮ would suffer due to the horizontal compression. We saw that this was only the case in Task 2, a synoptic task that required looking for specific types of transitions between phases. We had made sure that all would be visible in the Tall ▮ rendering but the fact that all transitions were visually closer together negatively affected completion time, accuracy, and confidence scores. We also saw that more than half of the respondents picked Square ◼ as their preferred display size
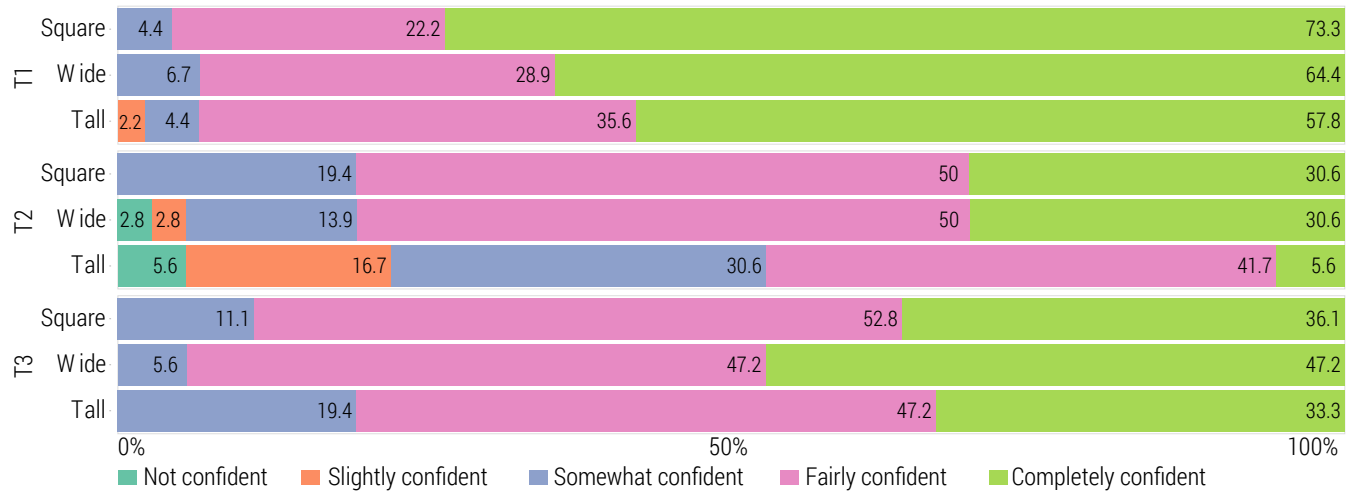
Figure 6: Percentages of confidence ratings by participants of different form factors for each task for the Sleep Phase (hypnogram) Study.

for this task, giving some evidence that despite not showing as the preferred visualization in our survey for the Square ▮, the chart worked well for most participants in our study.

8.2 Design Considerations

Based on our results, we summarize the following design considerations for sleep visualizations on fitness trackers.

Integrate sleep visualizations on fitness trackers: Overall, we saw in our survey an interest in detailed sleep data shown directly on the tracking devices themselves. Across our studies, people could solve all but one task in under 2 s with high accuracy, providing evidence that these visualizations have the potential to be glanceable, and effective at communicating sleep data to wearers [9].

Consider charts of the Tall ▮ orientation for wristbands: In our studies, we tested both vertical, and horizontal wristband sized displays. When actually worn around a wrist, a wristband-sized chart would be more quickly read if rendered using the Tall ▮ form factor because the wrist would have to be turned less. However, charts of this form factor did not perform well with respect to task completion time in more complex synoptic tasks. Practically, the differences in answer time might be outweighed by the additional time needed to turn the wrist to correctly orient charts of the Wide ▬ form factor. In addition, Tall ▮ bar, and hypnogram charts performed well according to completion time for elementary tasks, even outperforming (or trending to outperform) charts of the Wide ▬ form factor.

Consider horizontal bars for social comparison data on Wide ▬ form factors: In the survey, people strongly preferred bar representations. Due to the minimal compression in the vertical direction, which does not affect the data encoding, we expect these visualizations to be similarly accurately readable in social comparison tasks, as our horizontal bars in the present study on sleep duration.

Use horizontal bars for weekly overview data: We saw that participants were effective at both the elementary, and the two synoptic tasks. They also were confident, correct, and fast completing the three tasks (<2 s) with these charts at all form factors.

Do not use vertical hypnograms when wearers might want to understand sleep stage transitions: For T2, which required counting specific sleep stage transitions, we saw that participants needed the longest to answer (around 4 s), made the most errors, and were the least confident compared to the other form factors.

8.3 Methodological Reflection

The COVID-19 pandemic has prompted us to reconsider our research methods. Those of us who conduct in-person qualitative, or quantitative research have experienced substantial challenges in accessing study populations because of new social distancing regulations. In particular, research that requires specific types of technologies, such as smartwatches in our case, is challenged by the inability to run in-person studies. Our in-person pilot study, and the follow-up crowdsourced version of the pilot study, offer the opportunity to reflect on whether we can usefully run perception type studies for smartwatches in crowdsourced settings. Our approach simulated smartwatch-sized displays on smartphones, allowed us to target a wider audience, and make recommendations about smartwatch-sized displays. A recent study by Blascheck and Isenberg [12] already showed that running studies on a desktop computer with same sized stimuli leads to similar results, allowing researchers to run smartwatch studies on devices that are more available.

8.3.1 Summary, and Differences Between In-Person Pilot Study, and the Crowdsourced Sleep Duration (bar) Study.
The crowdsourced version of our Sleep Duration (bar) Study had a few study design differences compared to the in-person pilot study that we outlined in Section 6.1. To answer our question whether a crowdsourced study to target fitness tracker perception is viable,

we did not perform detailed statistical comparisons. Rather, we discuss how the results differ in the identified trends. In general, we would expect performance to be noisier in a crowdsourced study than in an in-person study but strong evidence should remain.

Overall, participants in the lab study had faster completion times. This can be caused by the different input methods used (keyboard vs. on-screen buttons on the smartphone), and the additional noise produced by the lack of control of the crowdsourced study environment. Several smaller effects in completion time that we observed in the in-person pilot study tended to be trends in the crowdsourced study. For example, for T1 (comparison of two adjacent bars), the evidence that Wide ▬ was slower than the other two form factors in the in-person pilot study became only a trend in the crowdsourced study; or the completion time difference between Tall ▮ and Square ▮ in T2 (comparison of multiple bars) that also became a trend in the crowdsourced study. Yet, the strong evidence of Tall ▮ being slower than Wide ▬ in T2 was present across both the in-person pilot study, and the crowdsourced study. Measuring time accurately in crowdsourced studies is a known challenge [14], and indicates that we may need a larger number of participants to reach the same level of evidence between in-person, and crowdsourced studies. Participants in both studies were accurate in their answers, with the average accuracy for T2, and T3 (comparing averages of groups of bars) being only slightly lower in the crowdsourced study (1–4 percentage points) but practically the same for T1.

Overall, we conclude that simulating fitness tracker displays on smartphones for crowdsourced perception studies worked well. We had engaged participants who performed similarly quickly as well as accurately, and strong evidence remained. Of course, we cannot perform all types of fitness tracker studies online by using smartphones. Any study that wishes to measure the impact of actually wearing the device, turning the wrist to look at the screen, or to measure the impact of contextual factors such as movement, still require field, or lab studies.

8.3.2 *Recommendations for Setting up a Crowdsourced Fitness Tracker Study.*

Here we reflect on several of our study setup decisions, and implementation details that helped to conduct a successful crowdsourced fitness tracker perception study using smartphones.

Ensuring study participants used a smartphone: On the Prolific platform we indicated that our study had to be performed using a smartphone, and wrote dedicated instructions. Because Prolific does not itself test whether participants really used a smartphone we implemented a check that tested whether participants used a mobile browser. Checking screen resolution alone is not a good test because modern smartphones have similar resolutions compared to many desktop displays still currently in use.

Ensuring correct stimuli sizing: Smartphone environments offer a way to implement visual stimuli using DIP to be rendered at similar physical sizes on viewers' smartphones. While we cannot control people's viewing distance from the screen in online perceptual studies, with DPIs we can at least ensure similar physical rendering sizes, and assume standard viewing distances for smartphones for the average participant.

Recording time: It is often argued that time is not a reliable measure in crowdsourced studies [14]. We set a time counter programmatically within the application: when participants changed the browser tab during the study the time counter would pause, and when participants returned to the tab the time counter would continue. However, issues with outliers in completion time remained, which we then removed from our analyses.

Ensuring engagement, and training: One of the main pitfalls of crowdsourced studies is reduced control in the assessment of participants' training [13]. We worked towards an increased willingness to engage with instructions by keeping them short, and splitting them across separate pages. The instructions were provided with text, and graphics before each training to be more engaging. Participants also needed to actively consent to understanding each instruction before they could see the next image. In our last non-mandatory question—"*Do you have any comment about the study, for example, concerning the clarity of the instructions, or technical issues you might have experienced? (optional)*" we asked participants to give us feedback. For the Sleep Duration (bar) Study, 28 participants gave us voluntary comments, 12 of them mentioned that the instructions were clear, and explained well to them. For the Sleep Phase (hypnogram) Study, 16 participants gave us comments, 4 of them mentioned that everything was clear, and understandable to them. Some of the comments from the participants were—"The study was enjoyable. There were no difficulties whatsoever. Everything was explained perfectly.", "The instructions were simple, and clear, and the study was very interesting.", "The instructions were clear, and well explained.", "Very clear instructions for which I was grateful.", "All was very clear, and understandable." As such, we are confident to recommend a similar study, and training setup.

Ensuring smooth study loading: We spent a lot of effort to reduce the loading time of the study web pages, and stimuli so that participants were not impeded by their internet speed. All graphics used in the study were produced in the Portable Network Graphics (PNG) format, or Graphics Interchange Format (GIF), and compressed to small file sizes. With a regular internet connection, the complete online study—a total of 32 web pages including all graphics—loaded within 2–4 seconds, and the first page within 7–15 milliseconds.

9 CONCLUSION

We investigated how to visualize sleep data on smartwatches, and fitness bands, through four different studies. In a first survey, we showed that wearers were mostly interested in weekly sleep duration, and nightly sleep phase data. Then, in an in-person pilot study, and two crowdsourced studies, we selected, and tested the effectiveness of some of the most preferred visual representations for this data—bars, and hypnograms—under both elementary, and synoptic tasks. We found that despite their reduced display real estate, Tall ▮, and Wide ▬ performed, with a few exceptions, similarly to the larger Square ▮ form factor. Accuracy was also high across tasks. This indicates that all form factors present viable platforms for displaying, and reading visualizations. Finally, we reflected on our adoption of a crowdsourced study methodology, which enabled us to reach a broader participant demographic for studies that have been traditionally lab-based.

Our work opens new questions for future research. While strong time performance trends were similar in our in-person, and crowd-sourced study, the crowdsourced study seemed to highlight more differences in accuracy. It would be interesting to investigate why this is the case, and more generally conduct a systematic comparison of how in-person, and crowdsourced study results differ for such specialized wearable devices. Moreover, both types of studies remain fairly constraint, and controlled: smartwatches, and fitness bands are worn in contexts that include movement, and changing lighting conditions that may reduce readability. It remains future work to investigate these, and other factors stemming from real-use that may affect visualization reading, and comprehension. In addition, it would be useful to test the generalizability of our findings to other types of small-scale data representations, for example, the area chart, which was also often preferred.

ACKNOWLEDGMENTS

REFERENCES

[1] M. Al-Sharrah, A. Salman, and I. Ahmad. 2018. Watch Your Smartwatch. In *International Conference on Computing Sciences and Engineering (ICCSE)*. IEEE, 1:1–1:5. https://doi.org/10.1109/ICCSE1.2018.8374228

[2] Danielle Albers, Michael Correll, and Michael Gleicher. 2014. Task-Driven Evaluation of Aggregation in Time Series Visualization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 551–560. https://doi.org/10.1145/2556288.2557200

[3] Fereshteh Amini, Khalad Hasan, Andrea Bunt, and Pourang Irani. 2017. Data Representations for In-Situ Exploration of Health and Fitness Data. In *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare* (Barcelona, Spain) *(PervasiveHealth '17)*. Association for Computing Machinery, New York, NY, USA, 163–172. https://doi.org/10.1145/3154862.3154879

[4] Natalia Andrienko and Gennady Andrienko. 2005. *Exploratory Analysis of Spatial and Temporal Data: A Systematic Approach*. Springer.

[5] Ranjini Aravind, Tanja Blascheck, and Petra Isenberg. 2019. A Survey on Sleep Visualizations for Fitness Trackers. In *Proceedings of EuroVis – Posters*. The Eurographics Association, 85–87.

[6] Stavros Asimakopoulos, Grigorios Asimakopoulos, and Frank Spillers. 2017. Motivation and User Engagement in Fitness Tracking: Heuristics for Mobile Healthcare Wearables. *Informatics* 4, 1 (2017), 5. https://doi.org/10.3390/informatics4010005

[7] Lonni Besançon and Pierre Dragicevic. 2017. The Significant Difference between p-values and confidence intervals. In *Proceedings of the Conference on l'Interaction Homme-Machine*. 53–62.

[8] Lonni Besançon and Pierre Dragicevic. 2019. The Continued Prevalence of Dichotomous Inferences at CHI. In *Extended Abstracts of the SIGCHI Conference on Human Factors in Computing Systems*. ACM.

[9] Tanja Blascheck, Frank Bentley, Eun Kyoung Choe, Tom Horak, and Petra Isenberg. 2021. Characterizing Glanceable Visualizations: From Perception to Behavior Change. In *Mobile Data Visualization*. CRC Press. To appear.

[10] Tanja Blascheck, Lonni Besançon, Anastasia Bezerianos, Bongshin Lee, and Petra Isenberg. 2019. Glanceable Visualization: Studies of Data Comparison Performance on Smartwatches. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2019), 630–640. https://doi.org/10.1109/TVCG.2018.2865142

[11] Tanja Blascheck, Anastasia Bezerianos, Lonni Besançon, Bongshin Lee, and Petra Isenberg. 2018. Preparing for Perceptual Studies: Position and Orientation of Wrist-worn Smartwatches for Reading Tasks. In *Proceedings of the Workshop on Data Visualization on Mobile Devices held at ACM CHI*.

[12] Tanja Blascheck and Petra Isenberg. 2021. A Replication Study on Glanceable Visualizations: Comparing Different Stimulus Sizes on a Laptop Computer. In *Proceedings of the Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications-Volume 3 IVAPP*. SCITEPRESS-Science and Technology Publications, 133–143.

[13] Rita Borgo, Bongshin Lee, Benjamin Bach, Sara Fabrikant, Radu Jianu, Andreas Kerren, Stephen Kobourov, Fintan Mcgee, Luana Micallef, Tatiana Landesberger, Katrin Ballweg, Stephan Diehl, Paolo Simonetto, and Michelle Zhou. 2017. *Crowdsourcing for Information Visualization: Promises and Pitfalls*. 96–138. https://doi.org/10.1007/978-3-319-66435-4_5

[14] Rita Borgo, Luana Micallef, Benjamin Bach, Fintan Mcgee, and B. Lee. 2018. Information Visualization Evaluation Using Crowdsourcing. *Computer Graphics Forum* 37 (06 2018), 573–595. https://doi.org/10.1111/cgf.13444

[15] Matthew Brehmer and Tamara Munzner. 2013. A Multi-Level Typology of Abstract Visualization Tasks. *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (2013), 2376–2385. https://doi.org/10.1109/TVCG.2013.124

[16] Sleep Stage by Fitbit. [n.d.]. REM, Light, Deep: How Much of Each Stage of Sleep Are You Getting? https://blog.fitbit.com/sleep-stages-explained/. Last visited: July, 2021.

[17] Xiwen Cai, Konstantinos Efstathiou, Xiao Xie, Yingcai Wu, Y. Shi, and Lingyun Yu. 2018. A Study of the Effect of Doughnut Chart Parameters on Proportion Estimation Accuracy. *Computer Graphics Forum* 37, 6 (2018), 300–312. https://doi.org/10.1111/cgf.13325

[18] Carme Carrion, Maurizio Caon, Stefano Carrino, Liliana Arroyo Moliner, Alexandra Lang, Sarah Atkinson, Marco Mazzola, Paolo Perego, Carlo Emilio Standoli, Conxa Castell, and Mireia Espallargues. 2015. Wearable Lifestyle Tracking Devices: Are They Useful for Teenagers?. In *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers* (Osaka, Japan) *(UbiComp/ISWC'15 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 669–674. https://doi.org/10.1145/2800835.2809442

[19] Marta E. Cecchinato, Anna L. Cox, and Jon Bird. 2017. Always On(Line)? User Experience of Smartwatches and Their Role within Multi-Device Ecologies. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)* (Denver, Colorado, USA). ACM, New York, NY, USA, 3557–3568. https://doi.org/10.1145/3025453.3025538

[20] J. Chauhan, S. Seneviratne, M. A. Kaafar, A. Mahanti, and A. Seneviratne. 2016. Characterization of Early Smartwatch Apps. In *Proceedings of the Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*. 1–6. https://doi.org/10.1109/PERCOMW.2016.7457170

[21] Yang Chen. 2017. Visualizing Large Time-Series Data on Very Small Screens. In *Proceedings of the Eurographics/IEEE VGTC Conference on Visualization: Short Papers* (Barcelona, Spain) *(EuroVis '17)*. Eurographics Association, Goslar, DEU, 37–41. https://doi.org/10.2312/eurovisshort.20171130

[22] James Clawson, Jessica A. Pater, Andrew D. Miller, Elizabeth D. Mynatt, and Lena Mamykina. 2015. No Longer Wearing: Investigating the Abandonment of Personal Health-Tracking Technologies on Craigslist. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Osaka, Japan) *(UbiComp '15)*. Association for Computing Machinery, New York, NY, USA, 647–658. https://doi.org/10.1145/2750858.2807554

[23] Andy Cockburn, Pierre Dragicevic, Lonni Besançon, and Carl Gutwin. 2020. Threats of a Replication Crisis in Empirical Computer Science. *Commun. ACM* 63, 8 (2020), 70-–79.

[24] Geoff Cumming. 2013. *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge.

[25] Android developer community. [n.d.]. Support different pixel densities. https://developer.android.com/training/multiscreen/screendensities. Last visited: August, 2021.

[26] Pierre Dragicevic. 2016. Fair Statistical Communication in HCI. In *Modern Statistical Methods for HCI*, Judy Robertson and Maurits Kaptein (Eds.). Springer, Chapter 13, 291–330.

[27] Daniel A. Epstein, Jennifer H. Kang, Laura R. Pina, James Fogarty, and Sean A. Munson. 2016. Reconsidering the Device in the Drawer: Lapses as a Design Opportunity in Personal Informatics. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Heidelberg, Germany) *(UbiComp '16)*. Association for Computing Machinery, New York, NY, USA, 829–840. https://doi.org/10.1145/2971648.2971656

[28] N. S. Erdem, C. Ersoy, and C. Tunca. 2019. Gait Analysis Using Smartwatches. In *Proceedings of the Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC Workshops)*. 1–6. https://doi.org/10.1109/PIMRCW.2019.8880821

[29] Rúben Gouveia, Fábio Pereira, Evangelos Karapanos, Sean A. Munson, and Marc Hassenzahl. c. Exploring the Design Space of Glanceable Feedback for Physical Activity Trackers. In *Proceedings of the Conference on Pervasive and Ubiquitous Computing* (Heidelberg, Germany). ACM, New York, NY, USA, 144–155. https://doi.org/10.1145/2971648.2971754

[30] Helena Hachul, Cristina Frange, Andréia Bezerra, Camila Hirotsu, Gabriel Pires, Monica Andersen, Lia Bittencourt, and Sergio Tufik. 2014. The effect of menopause on objective sleep parameters: Data from an epidemiologic study in São Paulo, Brazil. *Maturitas* 80 (11 2014). https://doi.org/10.1016/j.maturitas.2014.11.002

[31] James Higgins. 2004. *Introduction to Modern Nonparametric Statistics*. Thomson Learning.

[32] Alexandra Hosszu, Daniel Rosner, and Michael Flaherty. 2019. Sleep Tracking Apps' Design Choices: A Review. In *2019 22nd International Conference on Control Systems and Computer Science (CSCS)*. 426–431. https://doi.org/10.1109/CSCS.2019.00078

[33] A. Islam, A. Bezerianos, B. Lee, T. Blascheck, and P. Isenberg. 2020. Visualizing Information on Watch Faces: A Survey with Smartwatch Users. In *IEEE Visualization Conference (VIS)*. IEEE Computer Society Press, Los Alamitos, CA, United States, 156–160. https://doi.org/10.1109/VIS47514.2020.00038

[34] Jawbone. [n.d.]. Jawbone tracker. https://en.wikipedia.org/wiki/Jawbone_(company). Last visited: June, 2021.

[35] H. Kalantarian, N. Alshurafa, E. Nemati, T. Le, and M. Sarrafzadeh. 2015. A smartwatch-based medication adherence system. In *Proceedings of the Conference on Wearable and Implantable Body Sensor Networks (BSN)*. 1–6. https://doi.org/10.1109/BSN.2015.7299348

[36] Aida Kamišalić, Iztok Fister, Muhamed Turkanović, and Sašo Karakatič. 2018. Sensors and functionalities of non-invasive wrist-wearable devices: A review. *Sensors* 18, 6 (2018), 1714. https://doi.org/10.3390/s18061714

[37] Frederick Kingdom and Nicolaas Prins. 2010. *Psychophysics: A Practical Introduction* (1st ed.). Elsevier.

[38] Konstantin Klamka, Tom Horak, and Raimund Dachselt. 2020. Watch+Strap: Extending Smartwatches with Interactive StrapDisplays. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)* (Honolulu, HI, USA). ACM, New York, NY, USA, 1–15. https://doi.org/10.1145/3313831.3376199

[39] Martin Krzywinski and Naomi Altman. 2013. Points of Significance: Error bars. *Nature Methods* 10 (2013), 921–922.

[40] Jung-Min Lee, Wonwoo Byun, Alyssa Keill, Danae Dinkel, and Yaewon Seo. 2018. Comparison of Wearable Trackers' Ability to Estimate Sleep. *International Journal of Environmental Research and Public Health* 15, 6 (2018). https://doi.org/10.3390/ijerph15061265

[41] 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* (Barcelona, Spain) *(PervasiveHealth '17)*. Association for Computing Machinery, New York, NY, USA, 476–479. https://doi.org/10.1145/3154862.3154897

[42] 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 (Nov. 2016), 985–1000. https://doi.org/10.1007/s00779-016-0960-6

[43] Yvonne Jansen Luiz Augusto Morais, Pierre Dragicevic. [n.d.]. Framework for Running Online Experiments. https://github.com/yvonnejansen/FROE/. Last visited: August, 2021.

[44] Kent Lyons. 2016. Visual Parameters Impacting Reaction Times on Smartwatches. In *Proceedings of Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI)* (Florence, Italy). ACM, New York, NY, USA, 190–194. https://doi.org/10.1145/2935334.2935344

[45] Martin Maritsch, Caterina Bérubé, Mathias Kraus, Vera Lehmann, Thomas Züger, Stefan Feuerriegel, Tobias Kowatsch, and Felix Wortmann. 2019. Improving Heart Rate Variability Measurements from Consumer Smartwatches with Machine Learning. In *Adjunct Proceedings of the Conference on Pervasive and Ubiquitous Computing and Proceedings of the Symposium on Wearable Computers* (London, United Kingdom). ACM, New York, NY, USA, 934–938. https://doi.org/10.1145/3341162.3346276

[46] Donald McMillan, Barry Brown, Airi Lampinen, Moira McGregor, Eve Hoggan, and Stefania Pizza. 2017. Situating Wearables: Smartwatch Use in Context. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)* (Denver, Colorado, USA). ACM, New York, NY, USA, 3582–3594. https://doi.org/10.1145/3025453.3025993

[47] Sciences Po medialab. [n.d.]. i want hue. https://medialab.github.io/iwanthue/. Last visited: August, 2021.

[48] Isabella Nake, Aris Alissandrakis, and Janosch Zbick. 2016. Visualizing Quantified Self Data Using Avatars. In *ACHI 2016: The Ninth International Conference on Advances in Computer-Human Interactions*. International Academy, Research and Industry Association (IARIA), 57–66.

[49] S. M. Neis and M. I. Blackstun. 2016. Feasibility analysis of wearables for use by airline crew. In *2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC)* (Sacramento, CA). 1–9. https://doi.org/10.1109/DASC.2016.7778023

[50] Ali Neshati, Yumiko Sakamoto, and Pourang Irani. 2019. Challenges in Displaying Health Data on Small Smartwatch Screens. *Studies in Health Technology and Informatics* 257 (2019), 325–332. https://doi.org/10.3233/978-1-61499-951-5-325

[51] Ali Neshati, Yumiko Sakamoto, Launa C Leboe-McGowan, Jason Leboe-McGowan, Marcos Serrano, and Pourang Irani. 2019. G-Sparks: Glanceable Sparklines on Smartwatches. In *Proceedings of Graphics Interface (GI)* (Kingston, Ontario). Canadian Information Processing Society, 23–1. https://doi.org/10.20380/GI2019.23

[52] Shaleph O'Neill. 2016. Stripe Painting: A Method of Expressing the Experience of Cycling through 'quantified Self' Data Visualisation. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct* (Heidelberg, Germany) *(UbiComp '16)*. Association for Computing Machinery, New York, NY, USA, 600–601. https://doi.org/10.1145/2968219.2968328

[53] Araujo JF Patel AK, Reddy V. [n.d.]. Physiology, Sleep Stages. https://www.ncbi.nlm.nih.gov/books/NBK526132/. Last visited: July, 2021.

[54] Stefania Pizza, Barry Brown, Donald McMillan, and Airi Lampinen. 2016. *Smartwatch in Vivo*. Association for Computing Machinery, New York, NY, USA, 5456–5469. https://doi.org/10.1145/2858036.2858522

[55] Prolific. [n.d.]. Prolific – Online participant recruitment for surveys and market research. https://www.prolific.co/. Last visited: September, 2021.

[56] Rebecca Quintana, Chris Quintana, Cheryl Madeira, and James D. Slotta. 2016. Keeping Watch: Exploring Wearable Technology Designs for K-12 Teachers. In *Extended Abstracts of the Conference on Human Factors in Computing Systems (CHI)* (San Jose, California, USA). ACM, New York, NY, USA, 2272–2278. https://doi.org/10.1145/2851581.2892493

[57] Juan C. Quiroz, Min Hooi Yong, and Elena Geangu. 2017. Emotion-Recognition Using Smart Watch Accelerometer Data: Preliminary Findings. In *Proceedings of the Conference on Pervasive and Ubiquitous Computing and Proceedings of the Symposium on Wearable Computers* (Maui, Hawaii). ACM, New York, NY, USA, 805–812. https://doi.org/10.1145/3123024.3125614

[58] Steven Schirra and Frank R. Bentley. 2015. "It's Kind of like an Extra Screen for My Phone": Understanding Everyday Uses of Consumer Smart Watches. In *Extended Abstracts of the Conference on Human Factors in Computing Systems (CHI)* (Seoul, Republic of Korea). ACM, New York, NY, USA, 2151–2156. https://doi.org/10.1145/2702613.2732931

[59] Misfit Shine. [n.d.]. Misfit Shine Review. https://www.tomsguide.com/us/misfit-shine-fitness-tracker-review,review-1990.html. Last visited: June, 2021.

[60] Pekka Siirtola. 2019. Continuous Stress Detection Using the Sensors of Commercial Smartwatch. In *Adjunct Proceedings of the Conference on Pervasive and Ubiquitous Computing and Proceedings of the Symposium on Wearable Computers* (London, United Kingdom). ACM, New York, NY, USA, 1198–1201. https://doi.org/10.1145/3341162.3344831

[61] Katta Spiel, Fares Kayali, Louise Horvath, Michael Penkler, Sabine Harrer, Miguel Sicart, and Jessica Hammer. 2018. Fitter, Happier, More Productive? The Normative Ontology of Fitness Trackers. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI EA '18)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3170427.3188401

[62] Simon Stankoski, Nina Reščič, Grega Mežič, and Mitja Luštrek. 2020. Real-Time Eating Detection Using a Smartwatch. In *Proceedings of the Conference on Embedded Wireless Systems and Networks (EWSN)* (Lyon, France). Junction Publishing, USA, 247–252.

[63] Andrei Corneliu Suciu and Jakob Eg Larsen. 2018. Active Self-Tracking and Visualization of Subjective Experience using VAS and Time Spirals on a Smartwatch. In *Proceedings of the Data Visualization on Mobile Devices Workshop held at the ACM Conference on Human Factor in Computing Systems (CHI)*.

[64] Justin Talbot, John Gerth, and Pat Hanrahan. 2012. An Empirical Model of Slope Ratio Comparisons. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (2012), 2613–2620. https://doi.org/10.1109/TVCG.2012.196

[65] Lie Ming Tang, Margot Day, Lina Engelen, Philip Poronnik, Adrian Bauman, and Judy Kay. 2016. Daily & Hourly Adherence: Towards Understanding Activity Tracker Accuracy. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI EA '16)*. Association for Computing Machinery, New York, NY, USA, 3211–3218. https://doi.org/10.1145/2851581.2892438

[66] Lie Ming Tang and Judy Kay. 2017. Harnessing Long Term Physical Activity Data–How Long-Term Trackers Use Data and How an Adherence–Based Interface Supports New Insights. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 2, Article 26 (June 2017), 28 pages. https://doi.org/10.1145/3090091

[67] Withings. [n.d.]. Withings Pulse HR Health & fitness tracker. https://www.withings.com/us/en/pulse-hr. Last visited: June, 2021.

[68] Rayoung Yang, Eunice Shin, Mark W. Newman, and Mark S. 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* (Osaka, Japan) *(UbiComp '15)*. Association for Computing Machinery, New York, NY, USA, 623–634. https://doi.org/10.1145/2750858.2804269