Building Cognition-Aware Systems: A Mobile Toolkit for Extracting Time-of-Day Fluctuations of Cognitive Performance

TILMAN DINGLER, University of Stuttgart, Germany
ALBRECHT SCHMIDT, University of Stuttgart, Germany
TONJA MACHULLA, University of Stuttgart, Germany

People's alertness fluctuates across the day: at some times we are highly focused while at others we feel unable to concentrate. So far, extracting fluctuation patterns has been time and cost-intensive. Using an in-the-wild approach with 12 participants, we evaluated three cognitive tasks regarding their adequacy as a mobile and economical assessment tool of diurnal changes in mental performance. Participants completed the five-minute test battery on their smartphones multiple times a day for a period of 1-2 weeks. Our results show that people's circadian rhythm can be obtained under unregulated non-laboratory conditions. Along with this validation study, we release our test battery as an open source library for future work towards cognition-aware systems as well as a tool for psychological and medical research. We discuss ways of integrating the toolkit and possibilities for implicitly measuring performance variations in common applications. The ability to detect systematic patterns in alertness levels will allow *cognition-aware systems* to provide in-situ assistance in accordance with users' current cognitive capabilities and limitations.

CCS Concepts: • **Human-centered computing → Human computer interaction (HCI)**; • **Computing methodologies → Cognitive science**;

Additional Key Words and Phrases: cognition-aware systems; circadian rhythm; alertness;

ACM Reference Format:
Tilman Dingler, Albrecht Schmidt, and Tonja Machulla. 2017. Building Cognition-Aware Systems: A Mobile Toolkit for Extracting Time-of-Day Fluctuations of Cognitive Performance. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 1, Article 47 (January 2017), 15 pages.
https://doi.org/10.1145/3132025

1 INTRODUCTION

People's alertness, attention and vigilance is highly variable and subject to systematic changes across the day [13, 22]. These fluctuations—in part caused by *circadian rhythms*—impact higher level cognitive capacities, including perception, memory, and executive functions. During phases of high alertness we are able to perform tasks efficiently while during phases of low alertness we have trouble concentrating [5, 29]. An understanding and awareness of these rhythms can help schedule people's days by matching cognition-intensive and complex

This work is supported by the Future and Emerging Technologies (FET) programme within the 7th Framework Programme for Research of the European Commission, under FET grant number: 612933 (RECALL). T. Machulla and A. Schmidt were supported by the European Research Council, under grant number 683008 (Amplify).
Author's addresses: T. Dingler, A. Schmidt, T. Machulla, Institute for Visualization and Interactive Systems, University of Stuttgart, email: {tilman.dingler, albrecht.schmidt, tonja.machulla}@vis.uni-stuttgart.de.

2474-9567/2017/1-ART47
https://doi.org/10.1145/3132025

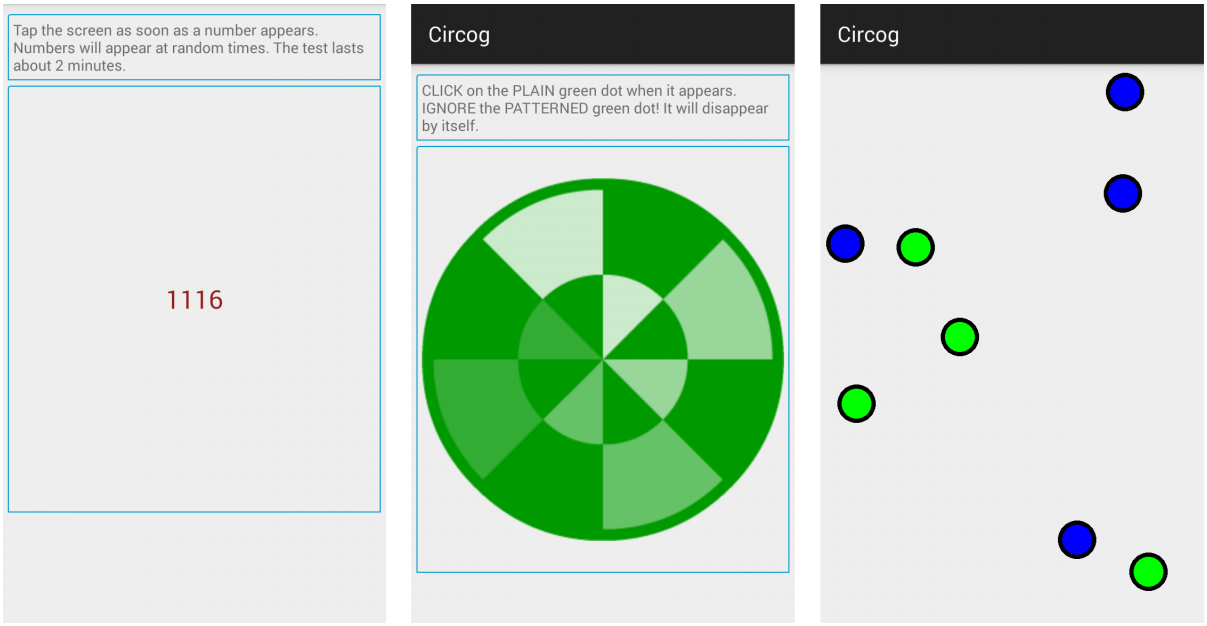

Fig. 1. Our toolkit consists of three tasks to measure alertness and cognitive performance variations across the day: a Psychomotoric Vigilance Task (left), a Go/No-Go task (middle), and a Multiple Object Tracking task (right).

tasks with times of high performance. Enabling applications to become aware of users' performance fluctuations allows them to take into account the user's current cognitive capacities.

Cognition-aware systems are computing systems capable of sensing and analyzing processes of human cognition, adapting to and aiding the user in consideration of the detected states–the so-called cognitive context [7, 8]. Such systems can be used to detect and support different aspects of mental information processing, such as user engagement, cognitive load, memory, knowledge, and learning. By identifying productive phases during the day, cognition-aware systems can suggest content or tasks that match the users' current cognitive state. Such state awareness can be used for task scheduling to support work rhythms [4], preventing interruptions [17, 25], and inducing flow states [10].

Traditional methods to assess the diurnal rhythms of alertness and associated changes in cognitive functioning include extensive lab visits, which can take up to several weeks. Other methods can be equally cumbersome or even unpleasant, such as sleep-wake protocols or physiological markers (*e.g.*, dim light melatonin onset, rectal temperature monitoring, cortisol level measurements) as well as extensive measurement of cognitive performance [16, 23]. Recent work has assessed the feasibility of using the psychomotor vigilance task (PVT [12])—the most common cognitive test to assess alertness—on mobile devices [1, 19]. The present work extends this research. In particular, our work provides the following contributions:

(1) We show that we can obtain systematic rhythms of cognitive performance from a small dataset by using sensitive statistical methods, and
(2) we release a cognitive test battery (see Figure 1) in form of a toolkit that can be used for psychological and medical research as well as for creating cognition-aware systems. By open sourcing the toolkit we enable researchers and application builders to integrate cognition-aware features into their work and allow them to sense and adapt to the user's cognitive context.

(3) We further discuss how the measures in our toolkit can be mapped to performance metrics collected by everyday applications, such as typing applications and games.

2 BACKGROUND

Our work is mainly based on previous work from the field of cognitive psychology regarding the assessment of circadian rhythms, and the notion of cognition-aware computing.

2.1 Cognition-Aware Computing

Context-aware computing traditionally entails input dimensions, such as the user's location, physical activity, task, or device environment. Bulling and Zander [8] proposed adding a cognitive dimension to the notion of context: by measuring cognitive aspects and using them as additional context cues, systems would be enabled to detect different aspects of mental information processing, such as engagement, cognitive load, memory, knowledge, and learning. The resulting cognition-aware computing systems are able to sense and adapt to the user's cognitive context. In our work we envision cognition-aware computing systems to be capable of deriving the user's cognitive context in terms of current cognitive performance capabilities. Thus, we understand cognition awareness as a subset of context awareness where alertness and cognitive performance levels are sensed and analyzed to adjust current systems to better fit the detected user state. Current systems rarely consider individuals' cognitive capacities. Instead, they generally assume a constant level of cognitive performance and rarely accommodate for variations. To provide a "window into our mind" [35] technologies need to be able to sense and infer cognitive activities, which naturally take place inside of the user. Monitoring bio-signals with the help of sensors can give us indirect clues about different cognitive states. Most of these sensors are rather invasive or require additional hardware equipment, such as Electroencephalography (EEG), Functional Magnetic Resonance Imaging (FMRI), Functional Near-Infrared Spectroscopy (fNIRS), or from eye movements [18]. Less invasive, but nonetheless disruptive techniques include self-reports through experience sampling [9] or filling in NASA-TLX surveys [15].

However, such techniques to derive cognitive processes are barely feasible for assessments in uncontrolled settings throughout the day. In an attempt to gather user data on cognitive variations in a more convenient and externally valid way, we focus in this paper on mobile devices and their capabilities to collect alertness data in-the-wild.

2.2 Measuring Alertness in-the-wild

Diurnal patterns of alertness are subject to strong individual fluctuations, which is why altertness-associated metrics, such as temperature changes and fluctuations in cognitive performance, are usually measured and averaged across a group of participants. In contrast, to determine how alert a specific individual is at a specific point in time requires comparison with a personal baseline rather than with the fluctuations measured across a group of individuals—ideally, we want an individual's typical alertness level for a particular hour of the day with small associated confidence bounds. The major issue in establishing such a baseline is that cognitive performance measurements associated with alertness levels are inherently noisy—a person might be very tired but still capable of compensating a decline in cognitive performance by allocating extra effort to the measurement task or by drinking three cups of coffee. Traditional sleep research attempts to minimize the noise by placing participants in strictly controlled environments over several weeks. Even under these circumstance, however, performance is usually so varied that averaging across many participants is required to obtain a stable baseline.

In-the-wild approaches to establishing a baseline do not have the option of controlling extraneous influences on alertness and are required to employ other means of reducing measurement and prediction noise. The most obvious is to measure over long periods of time, assuming that non-systematic fluctuations in alertness will

cancel each other out. But if these assessments are rather intrusive or time-intensive, long-term data collection is problematic for the user. A second possibility is to record factors that are known to influence alertness (such as caffeine and nicotine consumption, sleep duration and quality, day-time naps, physical activity *etc.*). These factors can then be included in mathematical models that reduce the amount of unaccounted variability in the baseline. The process of establishing a baseline can be sped up notably by fitting the empirical data with a generative model, *i.e.*, a model of the latent processes that are assumed to generate the systematic changes in a person's alertness levels. A well-chosen generative model reduces the number of data points that have to be collected before systematic changes in the data become statistically detectable. In this paper we apply such a model (see Section 2.3) to data collected across several participants to demonstrate that systematic changes in alertness can be extracted from a comparatively small data set of cognitive performance measures from 12 participants collected over on average 9 days. The eventual goal is to apply the same method to data from single participants. This, however, will require taking measurements over much longer periods of time and likely the use of implicit methods that do not interrupt the user's day.

Abdullah *et al.* [1] used the PVT to measure diurnal alertness fluctuations in-the-wild. In their study, 20 participants performed the PVT over 40 days multiple times a day. The results show that performance—when averaged across participants—can differ by up to 30% for different hours of the day (in particular between nighttime and noon). By correlating alertness levels with phone usage data they present a model to predict alertness states. Once the model has been trained with ground truth data from the PVT, it is applied in a rigid manner. Circadian rhythms of cognitive performance are, however, subject to changes depending on external factors, changing routines, and age [29]. For circadian computing technologies to become adaptive and widely accepted, they therefore need to adjust for changing patterns, *i.e.*, the models need to be updated in regular intervals. This is only feasible if alertness measurements shift from explicitly performing cognitive tasks, such as the PVT, to implicit assessments. Hence, in our work we investigate the feasibility of performance measures that can be recorded by general applications in everyday user interactions. We therefore extend the PVT metrics (*i.e.*, reaction time) by investigating measures of higher cognitive functions with the potential to be integrated in applications with the eventual goal to obtain these metric through implicit measures over extended periods of time.

In the present paper, we assess these measures explicitly. However, we i) try to minimize the disruption in participants' daily routine by using very short cognitive assessment tasks at the expense of collecting fewer data points and ii) we choose tasks that target the same cognitive capabilities and elicit the same behavioral metrics as many everyday interactions with technology. Specifically, we test a very short version of the PVT (1 minute) for its suitability to measure time-of-day dependent fluctuations in cognitive performance in-the-wild. The standard version of the PVT lasts 10 minutes [21], but multiple investigations have assessed the validity of shorter versions (between 1.5 and 3 minutes) and found it to be a reasonable substitute under conditions where the long version may be impractical [3, 24, 28]. In particular, Roach *et al.* [28] compare a 90-second version of the PVT with the 10-minute version and found that, while it loses sensitivity to diagnose sleep loss (*e.g.*, decrements in performance after 28 h of sleep loss), the response times on the short version still showed a moderate to strong correlation to the long version ($r = 0.77$). We additionally test two further cognitive tasks for their suitability for in-the-wild measurement of fluctuations in cognitive ability: a Go/No-Go task (GNG) and a Multiple Object Tracking task (MOT). By assessing the feasibility of these tasks to elicit cognitive performance measures we aim to validate general measures that can also be recorded in everyday applications, such as typing apps or games. We show how these tasks can be performed on mobile devices in a short amount of time and present their potential to eventually collect those measurements implicitly.

2.3 A Two-Process Model of Systematic Changes in Alertness Levels

According to the prevalent theory on sleep/wake regulation, variations in alertness and sleep propensity are generated by two underlying processes: sleep/wake homeostasis and a circadian process [6]. The homeostatic process manifests itself in a gradual decrease in alertness during wake periods. The longer we are awake, the stronger becomes the need for sleep. This is often referred to as sleep pressure. Alertness is further modulated by a circadian biological clock with a period length of about 24 h. Following a roughly sinusoidal pattern [20], it determines hours of the day when we experience a particularly low or particularly strong sleep drive. For many people, the alerting capability of the circadian process peaks in the late afternoon, thus partially counterbalancing the accumulated sleep pressure from the homeostatic process. This is commonly experienced as heightened alertness towards the evening after a post-lunch dip in alertness and concentration.

3 TOOLKIT FOR ASSESSING ALERTNESS

To make research on circadian rhythms of alertness widely applicable, we created and validated a test battery that can be deployed on mobile devices: it consists of three sustained attention tasks (see Figure 1). To validate this test battery, we built an Android app (Android version 4.1 or higher) that administers the tasks as well as a number of short questionnaires concerning the users' demographics, sleep and alertness self-assessments. The application prompts the user at random times during the day through notifications to complete the test battery. The time between task reminders is between 60 and 90 minutes. To respect sleep times notifications are only scheduled between 8 am and 9 pm. Additionally, users are free to launch the test application at any point in time. We included a logging mechanism that saves measurements locally on the device and transmits the logs to a remote server when a WIFI connection is available.

While there are many cognitive domains that are affected by the level of alertness (*e.g.*, memory), we focus on tasks that: 1) provide easily quantifiable metrics, 2) can later be integrated into common applications, 3) are easy to understand and commonly used in psychology, and 4) target various cognitive domains. The three tasks selected for inclusion into the toolkit are:

(1) a psychomotor vigilance task (PVT),
(2) a go/no-go task (GNG),
(3) and a multiple object tracking task (MOT).

We selected the PVT because it is the most widely used cognitive test to assess alertness. The GNG and the MOT tasks where selected for their game-like qualities and because they target more complex cognitive skills such as choice between options and keeping track of multiple objects. This makes them suitable candidates for integration into games as well as other continuous interactions with computing systems.

The Psychomotor Vigilance Task

The PVT measures the reaction time to a visual stimulus. In its original version, the task lasts 10 minutes and is thus a test of vigilance (the ability to sustain attention over time) as much as a test of psychomotor speed [12]. Shorter versions of this test have been validated [28]. During the task a visual stimulus is presented randomly every 2 to 6 seconds (see Figure 1, *left*). While the original experiment setup uses a physical button [12] to provide a response, our touchscreen implementation uses the *touch down* event as proposed by Kay *et al.* [19]. The dependent measure of the PVT is the reaction time in milliseconds.

The Go/No-Go Task

The GNG task falls into the class of choice reaction time paradigms. It uses two or more distinguishable stimuli, each associated with a unique answer option—in our case, a plain green circle, for which the participant needs to perform a speeded *touch down* gesture ("go" trial) and a patterned circle, for which this behavior needs to be

inhibited ("no-go" trial, shown in Figure 1, *middle*). Hence, this task measures reaction time, as well as executive functioning. In our implementation, we use between 8 and 12 stimuli, approximately half of which are no-go stimuli, appearing at random intervals of 1 to 8 seconds. If ignored, stimuli are shown for a maximum of 3 seconds. Therefore, the GNG task provides reaction time measures in milliseconds on correctly identified targets and the number of decoys that were reacted to due to failed inhibition, *i.e.* false alarms.

The Mutliple Object Tracking Task

The MOT is a strenuous sustained attention task that requires participants to divide their attention across multiple moving objects [27]. In our implementation 8 blue circles are shown. A subset of 4 target circles briefly flashes to indicate the objects to be tracked (see Figure 1, *right*). Then, all circles start moving in random, but linear directions. After 10 seconds the circles stop and the test person is asked to identify the target circles. This task is repeated 5 times, the performance measure is the percentage of correctly identified targets.

We wanted to assess the tasks for their utility as quick measurement tools that cause as little interruption as possible to users' daily routine. We therefore limited the first two tasks to about 1 minute each and the MOT to 2 minutes.

Open Source Toolkit

We released the app and the contained toolkit library under an open source license on *Github* [1]. The source code contains an Android project including all classes and layouts necessary to log users' performance data to a text file in JSON format. The raw measures can then be extracted from local storage and used for further analysis. The key classes comprise the three task types (PVT, GNG, and MOT) as well as the notification scheduler and logging service. By including the source code, application builders can instruct their application to either collect performance measurements with the activities provided by the toolkit or build their own based on the example log structure given. While the PVT solely collects reaction times, the GNG additionally provides false alarm rates as well as the number of missed targets and correct rejections. The MOT provides a percentage of correctly tracked targets representing multitasking capabilities.

4 VALIDATION STUDY

We conducted a user study to validate the effectiveness of the three attention tasks described regarding their ability to measure systematic fluctuations in alertness within a short duration of time. Since our goal was to measure fluctuations across the day we opted for an in-the-wild study, where participants were asked to perform the aforementioned tasks in their daily context. As dependent variables we collected task performances together with the time of day when the tasks were completed, subjective sleep and alertness assessments, and task preferences.

4.1 Procedure

We recruited 12 participants (4 female) with a mean age of 24 years ($SD = 2.67$) through university mailing lists. All participants were briefed about the purpose of the study and provided informed consent. The task order was randomized each time the app was opened. A service kept running in the background that managed the posting of notifications to remind users to perform the task sequences from time to time, up to six times a day. The prompts were shown in the notification drawer until clicked or dismissed. A click on the notification launched the task sequence. Before the first task sequence of the day, a survey was shown asking about the user's wake-up time, number of hours slept and rated quality of sleep (1=*poor*, 5=*very good*). Each task sequence was preceded by a short self-assessment regarding "*How alert are you feeling right now?*" (1=*super sleepy*, 5=*super alert*) and a checkbox labeled "*I had a caffeinated drink within the last hour*". The study ran for a total of 14 days; participants

[1] https://github.com/Til-D/circog.git

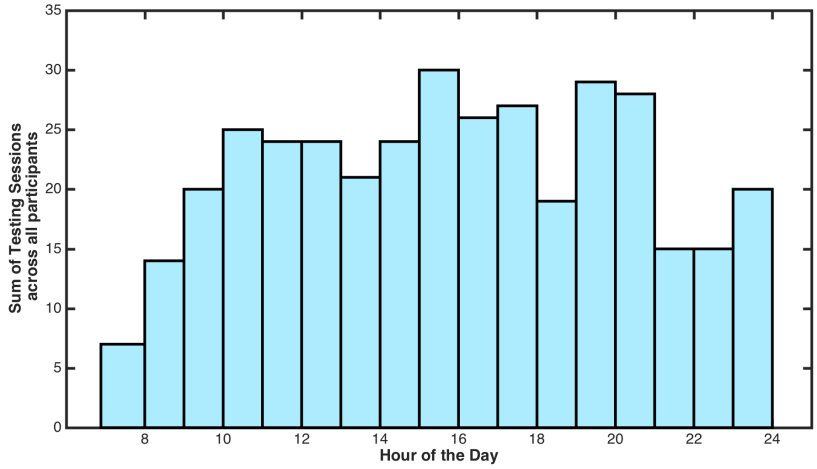

Fig. 2. Distribution of the numbers of samples entered into the analysis binned by hour of day for the MOT task. Distributions for the GNG and PVT are (close to) identical to the MOT as participants mostly performed all three tasks within one test session.

were free to start at any time by installing the app and completing the tasks for at least 7 days. We awarded 50 cents for each task sequence completed at a maximum of 6 sequences per day, resulting in up to 42 EUR. At the end of the study, we sent out a questionnaire to assess participants' subjective impressions of the different task types. For each task, we collected likert-style feedback on participants' evaluation of task difficulty, exhaustion, and fun.

4.2 Results

In the following we analyze each task according to its effectiveness to measure systematic changes in cognitive performance across the day. We further provide an assessment of the influence of caffeine, sleep, and the accuracy of participants' self-assessments, if any was found.

On average, participants performed the tasks on 9 days (SD = 3.9) with a minimum of 2 and a maximum of 13 days, resulting in a total of 367 PVT, 364 GNG, and 367 MOT tasks. We removed incomplete tasks and data points, for which the subjective alertness rating was missing. Figure 2 shows the number of data points available binned by hour of day.

Analysis Approach

To detect systematic variability in the performance data, we fitted the data consecutively with two models, one approximating the homeostatic and the other one identifying the circadian process as postulated by the two-process model of alertness fluctuations. The homeostatic process should result in a performance deterioration with time spent awake. For instance, simple reaction times (RT) in the PVT should increase throughout the day. Therefore, in a first step we examined the data for a linear trend over time. The data analysis was performed by fitting a linear mixed model to the raw data with the measure of interest as the dependent variable and the fixed factors *time of measurement (time of day in hours and minutes, with minutes converted into the decimal system)*, *self-rated alertness consumption of a caffeinated drink in the previous hour (treated as a categorical variable)*, *sleep*

	PVT	GNG		MOT
	Reaction times	Reaction times	False Alarms	Misses
Time of measurement	$\chi2(1) = 6.7$, $p = 0.009$	n.s.	$\chi2(1) = 4.3$, $p = 0.038$	n.s.
Self-rated alertness	$\chi2(1) = 5.8$, $p = 0.015$	n.s.	n.s.	n.s.
Caffeinated drink	$\chi2(1) = 10.8$, $p = 0.001$	n.s.	n.s.	n.s.
Sleep duration	n.s.	n.s.	n.s.	n.s.
Sleep quality	n.s.	n.s.	n.s.	n.s.

Table 1. Overview of the results of the linear mixed model analysis testing for the influence of the homeostatic process (time of measurement) and several other factors on the behavioral measures obtained with the PVT, GNG, and MOT tasks (non-significant effects are abbreviated as n.s.).

duration, and *self-rated sleep quality* as well as the random factor *subject*. The p-values were obtained by using a likelihood ratio test of the full model against a null model without the fixed effect of interest. If this comparison was non-significant, the fixed effect was excluded from further analysis. We did not find any significant interaction effects between these studied factors *i.e.*, non-linear interaction effects between the independent variables. For reasons of conciseness, we will only report significant effects. In other words, if we do not report on the influence of a particular independent variable, it is because it did not show a significant influence on the dependent variable. The results of this analysis are summarized in Table 1.

In a next step, we looked for variations in performance due to the circadian process. The circadian process should result in non-linear variations of cognitive performance across the day. As discussed earlier, previous laboratory studies have characterized these variations as following a roughly sinusoidal pattern. However, it has to be kept in mind that these studies differ form our study in two ways: first, they obtained measurements across the entire 24 hours of the day (*e.g.*, via a thermometer) while we were not able to take measurments at night when participants were asleep. Therefore, we cannot expect our data to follow a fully sinosoidal pattern. Second, laboratory studies impose special sleep schedules on their participants that desynchronize the homeostatic and the circadian processes. This has the advantage of allowing the measurement of the individual contribution of each process and removes possible interaction effects between the two. In contrast, in our case the two processes are synchronous in a more or less fixed manner. Currently, it is not known whether the processes add linearly or non-linearly [2].Therefore, we were not able to make a precise prediction regarding the expected pattern of performance changes caused by the circadian process in our particular data set. Hence, we predicted more generally that alertness would fluctuate non-linearly across the day after the influence of the homeostatic process is removed from the data.

To test this prediction, we first removed the linear trend (if one was found in the previous analysis) from the data. Thus, according to the two-process model, any remaining systematic variability in the data can be attributed to the circadian process (and potentially an interaction between the two processes—this cannot be distinguished further with the present experimental design). We fitted a second linear mixed model with the ordered categorical predictor variable *hour of the day* as a fixed factor (for this, observations were binned into 1-h bins) and *subject* as random factor. For each fit, we first report the results of an omnibus test (analysis of variance, ANOVA). This test indicates that there are at least two slots of one hour length, for which performance differs (*e.g.*, RT between 3

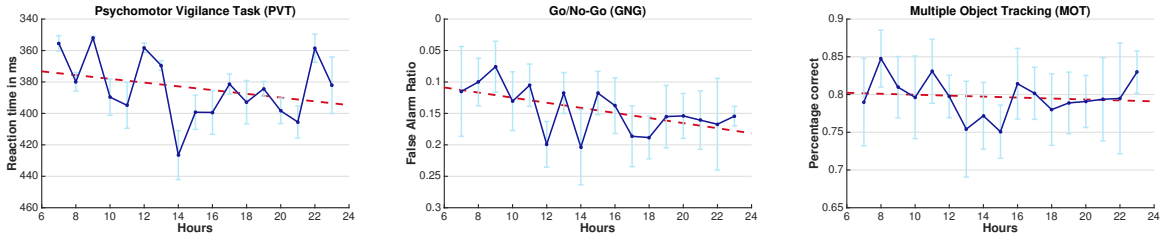

Fig. 3. Performance variations across the day in blue: mean reaction times from the Psychomotor Vigilance Task (left), mean false alarm ratio from the the Go/No-Go task (middle), and the percentage of correctly identified targets from the Multiple Object Tracking task (right). Error bars indicate the standard error of the mean. The red line depicts a linear fit to the data. The ordinate axis has been inverted for the PVT and the GNG task, to better communicate deterioration of performance. Model fitting was conducted on the raw data.

and 4 pm is significantly higher than RT between 7 and 8 pm). In a second step, we identify which time slots differ from each other. In particular, we perform a series of contrasts between successive hours (indicating whether performance changes between hour x and hour x + 1). This approach restricts the number of comparisons to the most informative ones and helps to control statistical type-I error. The level of significance was adjusted for multiple comparisons using the Holm-Bonferroni procedure.

Variability of Performance Measures across the Day

Figure 3 shows mean RTs (*left*) in the PVT, mean false alarm ratio for the GNG task (*middle*), as well as the percentage of correctly identified targets for the MOT (*right*), each as a function of time with the standard errors associated with the data falling into a time bin of one hour. Note that the data in this figure are presented for illustrative purposes, *i.e.*, to give the reader a summary impression of the data set. However, the depiction is not suited for visual analysis of statistical significance, neither in the case of linear trend analysis nor for pair-wise comparisons between time bins.

There are several reasons for this: first, the linear mixed models are not fitted on the means as depicted in the figure but rather on each participant's raw performance data. Thus, for the figure, data points collected at 3:01 and 3:59 pm would be summarized into one mean value while the analysis for the linear trend treats them as two separate values. Second, two means (*e.g.*, at 3 pm and 4 pm) cannot be visually compared as one would in case of an independent samples t-test, for example. One reason is that because of the under-constrained, in-the-wild collection of data, the time bins include different numbers of data points, from different participants. Because participants did not provide equal amounts of data points for each hour of the day (see Figure 2), the experimental design is an unbalanced within-subjects design. In this case, overlapping error bars do not indicate non-significance [11, 26].

PVT. We find that participants' RT increase by 1.9 ms (± 0.7) per hour of day (*time*: $\chi^2(1) = 6.7, p = 0.009$). An increase in RT represents slower reactions, which implies a performance deterioration reflecting the homeostatic process. After this linear relationship is removed from the data, we find a significant effect of hour ($F(16, 3175.3) = 1.8, p = 0.016$), indicating that there are further systematic differences in RT performance across the day. In particular, there is a significant increase of 54 ms from 1 to 2 pm ($t(3173.6) = 3.3, p < 0.001$). Mean RTs are highest just after 2 pm and lowest after 10 pm (399 ms vs. 316 ms). This agrees with previous reports regarding a post-lunch dip in alertness and performance [34] as well as reports of improved performance towards the evening in younger adults (which our sample predominately consists of).

We further find that consuming a caffeinated drink decreases RT on average by 17.6 ms (± 13.7 ms; $\chi^2(1) = 10.8, p = 0.001$) and that subjective alertness ratings are a significant predictor of PVT RT: RT decreases by 10.4

	PVT	GNG	MOT
Homeostatic Process	✓	✓	✗
Circadian Variations	✓	✗	✓

Table 2. Overview of tasks and their empirically measured effectiveness to detect the homeostatic process and circadian variations.

ms (± 4) per level of alertness rating ($\chi^2(1) = 5.8, p = 0.015$). In other words, participants are able to identify times of the day when their ability to maintain focus and to react quickly is impaired.

GNG. For the GNG task we analyzed RT as well as false alarm rates. False alarms are "go" responses to "no-go" stimuli. Both indices show a pattern similar to the PVT RT. Figure 3 (*upper right*) depicts the average false alarm (FA) proportion across the day.

For the RT, there is a linear increase of 0.8 ms/hour throughout the day, resulting in an overall difference of about 13 ms across a typical period awake. Additionally, there is a pronounced deterioration of performance in the early afternoon. However, in contrast to the PVT, neither the linear increase in RT is significant ($\chi^2(1) = 0.6, p = 0.44$) nor are there any significant differences in performance between consecutive hours. That is, neither the homeostatic nor the circadian processes are reflected in the RT of the GNG task. A possible reason for this finding is that the GNG task differs from the PVT task in one crucial aspect: participants attempt not only to answer as fast as possible but also as accurate as possible. These two goals compete with each other, leading to what is known as speed-accuracy trade-off: when participants are compromised in their alertness, they will either decrease their reaction speed as in the PVT or they will compensate by keeping the same speed but concede to making more false positive decisions—that is, reacting to non-target stimuli. To test whether participants attempted to stabilize reaction speed in this way, we performed a linear mixed effect model analysis with *RT* as the dependent variable and *false positive rate* as a fixed and *subject* as a random factor. We find strong indications of a speed-accuracy trade-off: as the false positive rate increases by 1%, RT decreases by 1.3 ms ($\chi^2(1) = 9.6, p = 0.002$).

For the false alarms, we find that participants become increasingly more likely to make a false alarm as the day progresses ($\chi^2(1) = 4.3, p = 0.038$, FA increase by 3% (± 2%) per hour of day). Such changes in the proportion of FA can result from changes in perceptual sensitivity or from adjustments in participants' answer patterns. Sensitivity indicates how well an observer discriminates between signal and noise (in our case, between target and non-target stimuli). It reflects an actual change in perception. Alternatively, the proportion of FA can change when humans adjust their preference for one response option over the other (*i.e.*, go vs. no-go response). To distinguish between these two possibilities, we analyzed the data using signal detection theory [14], which allows for the joined analysis of RT and the false positive rate in terms of two measures: d-prime and criterion, where d-prime measures sensitivity and the criterion is a measure of response bias, *i.e.*, the tendency to react over the tendency not to react. It is important to note that a change in criterion is a purely behavioral adjustment with no concomitant perceptual change.

Our analysis reveals that participants' response tendency (*i.e.*, criterion) does not change significantly over the day—neither is there a linear shift in criterion ($\chi^2(1) = 2.3, p = 0.129$) nor does the criterion change within successive hours. However, we find that sensitivity varies: d-prime decreases by 0.015 per hour of day ($\chi^2(1) = 4.2, p = 0.04$). This means that participants' ability to perceptually discriminate between target and non-target stimuli worsens as homeostatic sleep pressure accumulates. Rather than taking more time to correctly discriminate, participants chose to keep their reaction times stable. This could reveal a tendency to respond more impulsively as homeostatic sleep pressure increases or simply point towards the desire to finish the experiment as fast as possible.

MOT. Figure 3 (*right*) shows the average proportion of missed targets. We find no evidence for performance to worsen linearly throughout the day ($\chi^2(1) = 0.01, p = 0.91$), *i.e.*, there are no variations in the data that can be attributed to increased sleep pressure. However, we find significant differences between successive hour bins, which indicates that the MOT is sensitive to measuring circadian fluctuations. In particular, there is a significant improvement in performance between 3 and 4 pm ($z(365) = -4.1, p < 0.001$) just after performance had reached its daily low, when 25.6% (±16.6%) of targets are misidentified. Performance is best in the morning at 8 am, when only 12.9% of the targets are misidentified.

Influence of sleep quality and duration. We did not find an influence of these two variables on either of the performance measures. Participants did not show large variations in sleep quality (mean = 3.4 on a scale from 1 to 5 with a mean SD = 0.8 around participantsâĂŹ respective means) nor duration (mean = 7.4 h with a mean SD = 1 h around participants' respective means) during the measurement period. Sleep quality was largely judged positive and duration is overall well within the typical range of the general population. Hence, performance changes induced by variations in these two variables were likely too small to be detectable within our dataset.

Subjective Feedback

For evaluating participants' subjective assessments of the tasks regarding difficulty, exhaustion, and fun (0 = *totally disagree*, 6 = *completely agree*) we applied Friedman tests with post-hoc analyses using Wilcoxon signed-rank tests with a Bonferroni-corrected significance level set at $p < 0.017$.

There was a statistically significant difference in perceived **task difficulty** ($\chi^2 = 10.7, p = 0.005$) with MOT ($Mdn = 3.5, SD = 1.71$) being rated more difficult than PVT ($Mdn = 1, SD = 0.79, Z = -2.68, p = 0.007$). With regard to perceived **exhaustion**, we found a statistically significant difference ($\chi^2 = 6.9, p = 0.032$) with MOT ($Mdn = 3, SD = 1.8$) being rated as more exhausting than PVT ($Mdn = 1.5, SD = 0.7, Z = -2.413, p = 0.016$). As to which task was the most **fun** to complete we did not find any statistically significant difference, but rather a trend ($\chi^2 = 5.243, p = 0.073$) with the MOT reaching the highest rating ($Mdn = 5, SD = 2.49$). 10 out of 12 of our participants preferred the MOT (Binomial test: $p < .05$).

5 DISCUSSION

Results of our study demonstrate the feasibility to extract alertness-associated performance fluctuations across the day from comparably sparse data. To take measurements as unobtrusively as possible, we limited each task to 1-2 minutes in duration and gave participants the possibility of opting out of test sessions if timing was inconvenient within their everyday activities. To the best of our knowledge, this approach to measure and detect performance fluctuations is unique within the psychological and HCI literature, likely because it severely restricts the amount of data that can be collected in any given session and results in unbalanced experimental designs, which, in turn, result in datasets that can be challenging to analyze using inferential statistics. In spite of these factors, we were able to recover statistically significant variations in performance levels between different times of the day. In particular, statistical inference was aided by applying a biologically-plausible model of sleep/wake regulation that separates variations into a homeostatic and a circadian component as well as by the use of linear mixed modeling analysis that allows statistical testing of data collected in an unbalanced experimental design. This modeling approach extends the existing literature by providing a faster and more efficient route to the establishment of a robust prediction of time-of-day dependent performance variations.

Tasks for Measuring Cognitive Performance

The contribution of the work presented includes a demonstration of three tasks and their feasibility to extract alertness-associated performance fluctuations within a comparably short time frame. Table 2 summarizes the processes each task was able to empirically verify.

Not surprisingly, the PVT provides a robust measure of momentary alertness, a finding that is in accordance with previous publications [30, 32]. The reaction times we collected during the PVT show fluctuations that can be attributed to both increasing sleep pressure (the homeostatic process) as well as circadian rhythmicity. Typically, the PVT is conducted for 10 minutes [12] and applied regularly within fixed time windows (*e.g.*, every two hours). There are some exceptions, *e.g.*, Loh *et al.* [24] have measured for as short as 2 minutes—however, under controlled laboratory conditions and during the night, when alertness-associated fluctuations are particularly pronounced. In contrast, the PVT in our toolbox takes just about 1 minute to complete, and was administered at different times of the day at the convenience of the participants.

Our toolkit includes two further cognitive tasks—the GNG and MOT tasks, which our study has shown to be suitable for extracting homeostatic and circadian fluctuations, respectively. Thus, they are not as efficient as the PVT in terms of measuring both alertness-associated processes, most likely because less data points are collected within one session as compared to the PVT. Nevertheless, our investigation took them into consideration for three reasons: first, performing the PVT can be tedious and boring. Although it provides the most amount of data within the shortest time, it may be challenging to motivate users to perform this task several times a day over several weeks. Further, to re-assess mid-term changes in diurnal patterns the test would need to be reapplied in regular intervals. We were therefore interested in extending its metric collection (*i.e.*, reaction time) by a broader range of measures that can be taken from user interactions with everyday applications. By showing the feasibility of these measures to exhibit fluctuations of cognitive performance, application designers have more flexibility for choosing obtainable metrics and thus integrating cognition-aware features into their applications.

Second, if alertness is to be measured implicitly from users' interactions with computing systems, many natural activities will notably differ from the simple stimulus-response behavior captured by the PVT. Most interaction tasks do not only require a reflexive reaction to a stimulus, but rather decision making and performance accuracy, such as touch accuracy on smartphones or typing speed and error rates. Such tasks are better implemented as GNG tasks. In fact, our validation study demonstrates that reaction time in such choice scenarios may not be a good indicator of alertness—users can compensate decreases in psychomotor speed by being less accurate in their responses. We show that the analysis of such performance in terms of signal detection theory is more sensitive to underlying changes in alertness. Similarly, the MOT task taxes various higher cognitive functions beyond psychomotor speed, such as distributing attention across space over extended periods of time. This more closely resembles natural tasks such as navigating through a crowded environment of interacting with content on wall-sized displays.

Third, there is some indication in the literature that different cognitive functions might follow different performance profiles across the day, *e.g.* [22]—possibly because they are differently influenced by the circadian and homeostatic sleep process. For instance, impulsive responses to wrong stimuli might be more strongly affected by high sleep pressure than the distribution of attention across space, which in contrast might be more strongly impaired by circadian "lows". These different influences may also explain the slight shifts in circadian peaks detected by our different tests. While they may be due to measurement noise or noise in the participantsâĂŹ daily rhythm (measuring performance over a long period of time may result in converging peaks), Kleitman [22] found performance curves peaks at different hours across eight different neuropsychological tasks despite them following a similar general pattern. Hence, 'alertness' and 'cognitive performance' are not unitary constructs, but are rather composed of different components. The measures contained in the toolkit presented allow future research in continuous and implicit measurement of performance, which may provide us with the opportunity to investigate these components in more detail.

Relationship between Subjective and Objective Measures of Time-of-day-dependent Alertness Fluctuations

In the present study, each test session was preceded by a short self-assessment regarding how alert the participant felt. We found that the magnitude of these self-assessments fluctuates systematically with the PVT reaction times (i.e., when participants feel less alert, their reaction times increase). In contrast, this subjective measure of alertness was not significant as a predictor in the GNG- and MOT-task related measures. Does this imply that these two tasks are not subject to alertness-related fluctuations? We would argue no. One possible explanation for a non-significant result is that our analysis did not have sufficient power to detect an existing relationship.

Another explanation is that subjectively assessed alertness is likely not a highly reliable indicator of a personâĂŹs true cognitive status. This means that self-ratings do not constitute âĂIJground truthâĂİ for alertness. If they were, people would not fall asleep at the wheel or in front of the TV. In other words, misperceptions of oneâĂŹs own status are not uncommon. Momentary alertness perception can vary depending on how engaging the current activity is or on whether a person is performing an arousing physical activity such as walking stairs, which might raise perceived alertness but not necessarily improve cognitive functioning. Therefore, in walk-in laboratory studies on day-time sleepiness (i.e., studies where the participant comes into the laboratory several times a day to be tested), participants are sometimes asked to sit without activity for a couple of minutes before assessment. This serves to unify the conditions just before testing and probably results in less noisy data. In our study, we did not have this type of control over the activity the participant was currently engaging in. This might be a common problem of in-the-wild studies, as Abdullah et al. also only found a very coarse relationship between subjectively and objectively measured alertness, namely that RTs associated with alertness ratings above the median alertness were shorter than RTs associated with ratings below the median (assessed with a t test). In this regard, our study goes one step further in that our model fit allows us to conclude that for our particular measure of subjective alertness, reaction times increase on average by 10 ms for each level of the rating scale.

Lastly, while we cannot be entirely certain that GNG and MOT measure alertness-related changes, we would like to argue that this is the case for two reasons: first, both GNG and spatial tracking tasks like the MOT have been used to measure alertness in previous research [32]. Secondly, the patterns we observe are time-of-day dependent and, more importantly follow the known pattern of alertness variations across the day with peaks in the morning/evening and a low after lunch rather than the other way around. Still, it might be more prudent to speak of âĂIJtime-of-day-dependentâĂİ rather than âĂIJalertness-relatedâĂİ variations. Regardless, the eventual goal is to compensate temporary but lawful cognitive decrements in the user, independent of the cause of these impairments.

Utilizing Time-of-Day Fluctuations of Cognitive Performance

The combination of PVT, GNG and MOT in our toolkit provides a more holistic assessment of cognitive performance. We assessed the feasibility of metrics beyond reaction times (such as provided by the PVT): both, GNG and MOT, extend simple alertness assessments by targeting higher cognitive functions (such as executive control and divided attention). Using the GNG we were able to detect the homeostatic process, while the MOT measured circadian variations. This extended range of available metrics: false alarm rates (*i.e.*, errors), multitasking performance (*i.e.*, percentage of correctly tracked targets) and response inhibition can be collected by existing applications. A keyboard can assess typing errors, corrections, or speed measures, for example. In contrast to the PVT, the nature of GNG and MOT further allows them to be adapted in terms of their difficulty and the challenge they pose to the user. Games are especially suited to collect a range of metrics and different levels thereof, including spatial tasks and multitasking capabilities. Our results show the feasibility of these metrics to capture aspects of alertness and cognitive performance. By integrating them in everyday applications, such metrics can also be collected implicitly and therefore inform applications about their users' cognitive states.

Applications that take into account circadian rhythms of cognitive performance can prospectively be applied to a broad range of applications ranging from schedule alignment according to the user's internal body clock, stress prevention through sleep/wake regulation, to recommending alertness-inducing activities. While our toolkit can be applied in psychology research to collect data on cognitive performance in-the-wild, we also see potential for medical applications to preserve mental health. Stress, for instance, occurs when there is a mismatch between task requirements and the user's cognitive capabilities [33]. Chronically high levels of mental load and circadian disruptions can further lead to depression, burnout, and even diabetes [31].

6 CONCLUSION

Our research makes two contributions to the nascent field of circadian computing. We show how the combination of a biologically plausible, generative model of sleep/wake regulation with sensitive statistical modeling can improve the predictive capability of a circadian computing system—in particular, a baseline of systematic changes in alertness levels can be established on the basis of a comparatively small dataset.

Further, we present a mobile toolkit for assessing alertness and cognitive performance of users by using a combination of three tasks. Specifically, we include a very short version of the PVT, thus decreasing the disruption measurement causes to users' daily routines. Further, we added two tasks that measure variations in higher cognitive functions. Each task was successful in measuring at least one of two latent processes (circadian and/or homeostatic) known to influence alertness levels. The variety of tasks in the toolkit provides a spectrum of behavioral metrics that can be integrated into game-like applications and avoids limiting users to the rather plain PVT task.

We release the toolkit as an open source library to allow researchers in psychology and medicine, as well as application builders to create cognition-aware systems. Systems that are aware of users' performance rhythms can adapt interface complexity and information bandwidth to match the user's current state. Due to the task variety, similar metrics can be collected in existing games and applications and therefore make data collection even less obtrusive.

REFERENCES

[1] Saeed Abdullah, Elizabeth L Murnane, Mark Matthews, Matthew Kay, Julie A Kientz, Geri Gay, and Tanzeem Choudhury. 2016. Cognitive rhythms: unobtrusive and continuous sensing of alertness using a mobile phone. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 178–189.

[2] Peter Achermann and Alexander A Borbély. 2003. Mathematical models of sleep regulation. *Frontiers in bioscience: a journal and virtual library* 8 (2003), s683–93.

[3] Mathias Basner, Daniel Mollicone, and David F Dinges. 2011. Validity and sensitivity of a brief psychomotor vigilance test (PVT-B) to total and partial sleep deprivation. *Acta astronautica* 69, 11 (2011), 949–959.

[4] James Bo Begole, John C Tang, Randall B Smith, and Nicole Yankelovich. 2002. Work rhythms: analyzing visualizations of awareness histories of distributed groups. In *Proceedings of the 2002 ACM conference on Computer supported cooperative work*. ACM, 334–343.

[5] Katharina Blatter and Christian Cajochen. 2007. Circadian rhythms in cognitive performance: methodological constraints, protocols, theoretical underpinnings. *Physiology & behavior* 90, 2 (2007), 196–208.

[6] A A Borbély. 1982. A two process model of sleep regulation.. In *Human Neurobiology*, Vol. 1. Springer-Verlag, 195–204.

[7] Andreas Bulling, Daniel Roggen, and Gerhard Troester. 2011. What's in the Eyes for Context-Awareness? *IEEE Pervasive Computing* 10, 2 (2011), 48–57.

[8] A. Bulling and T. O. Zander. 2014. Cognition-Aware Computing. *IEEE Pervasive Computing* 13, 3 (July 2014), 80–83. https://doi.org/10.1109/MPRV.2014.42

[9] Sunny Consolvo and Miriam Walker. 2003. Using the experience sampling method to evaluate ubicomp applications. *IEEE Pervasive Computing* 2, 2 (2003), 24–31.

[10] Mihaly Csikszentmihalyi. 1988. The flow experience and its significance for human psychology. (1988).

[11] Geoff Cumming. 2009. Inference by eye: reading the overlap of independent confidence intervals. *Statistics in medicine* 28, 2 (2009), 205–220.

[12] David F Dinges and John W Powell. 1985. Microcomputer analyses of performance on a portable, simple visual RT task during sustained operations. *Behavior research methods, instruments, & computers* 17, 6 (1985), 652–655.
[13] N Goel, HPA Van Dongen, and DF Dinges. 2011. Circadian rhythms in sleepiness, alertness, and performance. *Principles and practice of sleep medicine* 5 (2011), 445–55.
[14] DM Green and JA Swets. 1966. Signal detection theory and psychophysics. 1966. *New York* 888 (1966), 889.
[15] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in psychology* 52 (1988), 139–183.
[16] Wytske A Hofstra and Al W de Weerd. 2008. How to assess circadian rhythm in humans: a review of literature. *Epilepsy & Behavior* 13, 3 (2008), 438–444.
[17] Eric Horvitz, Johnson Apacible, and Muru Subramani. 2005. Balancing awareness and interruption: investigation of notification deferral policies. In *Proc. UM '05*. Springer-Verlag, 5. https://doi.org/10.1007/11527886_59
[18] Marcel Adam Just and Patricia A Carpenter. 1976. Eye fixations and cognitive processes. *Cognitive psychology* 8, 4 (1976), 441–480.
[19] Matthew Kay, Kyle Rector, Sunny Consolvo, Ben Greenstein, Jacob O Wobbrock, Nathaniel F Watson, and Julie A Kientz. 2013. PVT-touch: adapting a reaction time test for touchscreen devices. In *2013 7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops*. IEEE, 248–251.
[20] GA Kerkhof and HPA Van Dongen. 2010. Effects of sleep deprivation on cognition. *Human sleep and cognition: basic research* 185 (2010), 105.
[21] William D S Killgore. 2010. *Effects of sleep deprivation on cognition* (c ed.). Progress in Brain Research, Vol. 185. Unknown Publisher. https://doi.org/10.1016/B978-0-444-53702-7.00007-5
[22] Nathaniel Kleitman. 1923. Studies on the Physiology of Sleep. *American Journal of Physiology–Legacy Content* 66, 1 (1923), 67–92.
[23] N Kleitman, S Titelbaum, and P Feiveson. 1938. The effect of body temperature on reaction time. *American Journal of Physiology–Legacy Content* 121, 2 (1938), 495–501.
[24] Sylvia Loh, Nicole Lamond, Jill Dorrian, Gregory Roach, and Drew Dawson. 2004. The validity of psychomotor vigilance tasks of less than 10-minute duration. *Behavior Research Methods, Instruments, & Computers* 36, 2 (2004), 339–346.
[25] Gloria Mark, Shamsi Iqbal, Mary Czerwinski, and Paul Johns. 2015. Focused, Aroused, but So Distractible: Temporal Perspectives on Multitasking and Communications. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*. ACM, New York, NY, USA, 903–916. https://doi.org/10.1145/2675133.2675221
[26] Michael EJ Masson and Geoffrey R Loftus. 2003. Using confidence intervals for graphically based data interpretation. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale* 57, 3 (2003), 203.
[27] Zenon W Pylyshyn and Ron W Storm. 1988. Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial vision* 3, 3 (1988), 179–197.
[28] Gregory D Roach, Drew Dawson, and Nicole Lamond. 2006. Can a Shorter Psychomotor Vigilance Task Be Usedas a Reasonable Substitute for the Ten-Minute Psychomotor Vigilance Task? *Chronobiology international* 23, 6 (2006), 1379–1387.
[29] Christina Schmidt, Fabienne Collette, Christian Cajochen, and Philippe Peigneux. 2007. A time to think: circadian rhythms in human cognition. *Cognitive neuropsychology* 24, 7 (2007), 755–789.
[30] Edward J Silva, Wei Wang, Joseph M Ronda, James K Wyatt, and Jeanne F Duffy. 2010. Circadian and wake-dependent influences on subjective sleepiness, cognitive throughput, and reaction time performance in older and young adults. *Sleep* 33, 4 (2010), 481–490.
[31] Tiinamaija Tuomi, Cecilia LF Nagorny, Pratibha Singh, Hedvig Bennet, Qian Yu, Ida Alenkvist, Bo Isomaa, Bjarne Östman, Johan Söderström, Anu-Katriina Pesonen, et al. 2016. Increased melatonin signaling is a risk factor for type 2 diabetes. *Cell metabolism* 23, 6 (2016), 1067–1077.
[32] Pablo Valdez, Candelaria Ramírez, and Aída García. 2012. Circadian rhythms in cognitive performance: implications for neuropsychological assessment. *Chronophysiol Ther* 2 (2012), 81–92.
[33] Geertje van Daalen, Tineke M Willemsen, Karin Sanders, and Marc JPM van Veldhoven. 2009. Emotional exhaustion and mental health problems among employees doing âĂIJpeople workâĂİ: The impact of job demands, job resources and family-to-work conflict. *International archives of occupational and environmental health* 82, 3 (2009), 291–303.
[34] Hans PA Van Dongen and David F Dinges. 2000. Circadian rhythms in fatigue, alertness, and performance. *Principles and practice of sleep medicine* 20 (2000), 391–9.
[35] Boris M. Velichkovsky and John Paulin Hansen. 1996. New Technological Windows into Mind: There is More in Eyes and Brains for Human-computer Interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '96)*. ACM, New York, NY, USA, 496–503. https://doi.org/10.1145/238386.238619

Received May 2017; accepted July 2017