Validation of sleep-2-Peak: A smartphone application that can detect fatigue-related changes in reaction times during sleep deprivation

Jean-François Brunet¹ · Dominique Dagenais¹ · Marc Therrien^{2,3} · Daniel Gartenberg⁴ · Geneviève Forest¹

Published online: 8 September 2016
© Psychonomic Society, Inc. 2016

Abstract Despite its high sensitivity and validity in the context of sleep loss, the Psychomotor Vigilance Test (PVT) could be improved. The aim of the present study was to validate a new smartphone PVT-type application called *sleep-2-Peak* (s2P) by determining its ability to assess fatigue-related changes in alertness in a context of extended wakefulness. Short 3-min versions of s2P and of the classic PVT were administered at every even hour during a 35-h total sleep deprivation protocol. In addition, subjective measures of sleepiness were collected. The outcomes on these tests were then compared using Pearson product-moment correlations, *t* tests, and repeated measures within-groups analyses of variance. The results showed that both tests significantly correlated on all outcome variables, that both significantly distinguished between the alert and sleepy states in the same individual, and that both varied similarly through the sleep deprivation protocol as sleep loss accumulated. All outcome variables on both tests also correlated significantly with the subjective measures of sleepiness. These results suggest that a 3-min version of s2P is a valid tool for differentiating alert from sleepy states and is as sensitive as the PVT for tracking fatigue-related changes during extended wakefulness and sleep loss. Unlike the PVT, s2P does not provide feedback to subjects on each trial. We discuss how this feature of s2P raises the possibility that the

performance results measured by s2P could be less impacted by motivational confounds, giving this tool added value in particular clinical and/or research settings.

Keywords Vigilance · Sleep Loss · Alertness · Performance · PVT · Smartphone application · Sustained attention

Night shift work often causes poor sleep quality and shortens sleep, which may result in somnolence, difficulties in executing required tasks, or even errors and accidents while working (Dinges, 1995; Philip & Åkerstedt, 2006). Unfortunately, night workers have few means to objectively assess their level of vigilance, and therefore cannot assess their level of efficacy in the field (Lamond, Dawson, & Roach, 2005). Most of the tools available are designed for research or require sophisticated or expensive equipment (Balkin et al., 2004). The Psychomotor Vigilance Test (PVT), developed in the 1980s by Dinges and Powell (1985), is a good example of such a tool. Indeed, the PVT is currently the most popular validated reaction time (RT) test for sleep research and has been the gold standard in the field for many years (Dinges & Powell, 1985; Khitrov et al., 2014). However, the PVT is available and validated on only a very limited number of devices, including the original PVT-192 and a few computer-based versions. This RT test is designed to measure alertness and is sensitive to sleep deprivation, extended wakefulness, and circadian misalignment (Balkin et al., 2004; Basner & Dinges, 2011, 2012; Basner, Mollicone, & Dinges, 2011; Basner & Rubinstein, 2011; Loh, Lamond, Dorrian, Roach, & Dawson, 2004; Thorne et al., 2005).

The advantages of the PVT are widely documented, especially due to its high sensitivity and validity in the context of sleep loss (Basner & Dinges, 2011, 2012). For example, many studies have confirmed that the RTs

✉ Geneviève Forest
genevieve.forest@uqo.ca

¹ Laboratoire du Sommeil, Département de Psychoéducation et de Psychologie, Université du Québec en Outaouais, Gatineau, Québec, Canada J8X 3X7

² Neuro Summum, Gatineau, QC, Canada

³ Neurologie, CISSS de l'Outaouais, Gatineau, QC, Canada

⁴ George Mason University, Fairfax, VA, USA

obtained on the PVT during chronic and partial sleep deprivation reflect changes in neurocognitive performance consistent with theories of the cognitive functions of sleep (Dorrian, Rogers, & Dinges, 2005; Gunzelmann, Moore, Gluck, Van Dongen, & Dinges, 2011). The PVT also has good ecological validity; that is to say, it can reflect the performance of an individual in their daily functioning (Dorrian et al., 2005). Despite numerous advantages, the PVT has been often considered too long (10 min) for clinical use (Loh et al., 2004), the price of the PVT-192 too high for widespread community accessibility (Lamond et al., 2005; Thorne et al., 2005), and the size of the PVT-192 too big for some research protocols that require a pocket-sized unit (Lamond et al., 2005).

More recently, efforts have been made to create portable, shorter, and more affordable assessment tools that will offer alternatives to the classic PVT (Basner et al., 2011; Basner & Rubinstein, 2011; Lamond et al., 2005, 2008; Roach, Dawson, & Lamond, 2006; Thorne et al., 2005). In recent years, smartphones and tablets have offered new opportunities for the development of tasks alternative to the PVT. In 2014, 1.301 billion smartphones were delivered (ZDNet, 2015), and in June 2015, Apple stated that more than 100 billion apps have been downloaded from the App Store (Ingraham, 2015). This new technology platform represents an easy and accessible way to reach many people and to create customized tools to assess fatigue-related RT changes in a variety of work and clinical settings.

Sleep-2-Peak (s2P) is a smartphone app developed by Therrien and Gartenberg in 2012 (Proactive Life LLC, 2012). This PVT-type task is designed to track changes in RTs over the course of the day, which are related to the amount of sleep obtained by the individual (Proactive Life LLC, 2012). The app allows its users to do several tests during the day and is able to advise on the best bedtime and wake-up time to optimize human performance. Its cost is low, and it can be used on the two major smartphone platforms (iOS and Android; Proactive Life LLC, 2012). This app also allows both the user and the experimenter to have easy access to all data. However, some features of s2P differ from those of the classic PVT, such as the touch screen, the timing of stimulus presentations, the type of stimulus, and the absence of feedback. Therefore, validation studies are needed to determine whether s2P is a good alternative to the PVT for measuring variations in vigilance.

The aim of the present study was to validate the s2P app and determine its ability to assess fatigue-related changes in alertness. To do this, a total sleep deprivation (TSD) protocol was used and s2P performance was compared to the performance on the classic PVT, in addition to other subjective measures of alertness.

Method

Subjects and protocol

Twenty-two subjects (12 men and ten women), aged 18–27 years old, participated in this study. Recruitment was done through social media and posters at the Université du Québec en Outaouais in Gatineau (Québec, Canada), where the study was conducted. The protocol was approved by the ethics committee of the university in accordance with the declaration of Helsinki. Exclusion criteria were the presence of a neurological, psychiatric, hormonal, or sleep disorder; taking psychotropic drugs or tricyclic birth control pills; a history of alcohol or drugs abuse; the presence of an uncorrected vision disorder; and being a smoker. Also, only individuals who kept a regular sleep–wake schedule, slept regularly from 7 to 9 h at night, had a habitual bedtime between 10 PM and midnight and a habitual wake time between 6 AM and 8 AM and nonnappers were allowed to participate. People who had worked night shifts in the last three months or who had jet lag (one week of recovery for each 1 h of jet lag) were also excluded.

All subjects were pretrained on a PVT-192 (PVT) and on a fourth-generation iPod touch (iOS 7.1) that had s2P installed (version 1.2; released on October 28, 2013). The training consisted of three practice sessions on each device to familiarize the subject with both the RT tests and the equipment. Subjects were then monitored using actigraphy and a sleep diary for 7 days before the TSD protocol was applied. During this period, all subjects were asked to maintain their regular sleep–wake habits and daily activities and to abstain from taking naps, doing intense physical exercise after 6 PM, and taking nonprescribed drugs (except Tylenol). Coffee consumption was restricted to a maximum of one coffee per day (ingested before noon), and alcohol was forbidden for three days before the TSD session.

On the seventh day, starting at 8 AM (on average, about an hour after waking up in the morning), subjects completed subjective measures of sleepiness (the Stanford Sleepiness Scale [SSS], the Karolinska Sleepiness Scale [KSS], and the visual analogue scale [VAS]) and both vigilance tests (s2P and PVT) in a counterbalanced design at every even hour until 8 PM at home. Then, the testing sessions continued at the sleep laboratory under close supervision from 10 PM until 6 PM the next day, for a total of about 35 h of consecutive waking. During each testing session, a 5-min break was scheduled between the two RT tests. Each subject had to fill out a testing diary in which he confirmed the time and the order of completion of both RT tests. During the night at the laboratory, subjects were free to engage in various activities, such as reading, board games, watching movies, or using the Internet. Meals and snacks were provided to control consumption of stimulants (caffeine, sugars). A safe departure from the

lab was scheduled with each subject (they either took the bus or were driven home by the experimenter).

Outcomes measures

Sleep-2-Peak Sleep-2-Peak is an app running on Apple (iPod Touch and iPhone) and Android (Blackberry, Samsung, etc.) mobile operating systems. In the present study, the s2P app (version 1.2; released on October 28, 2013) was installed on a fourth-generation iPod Touch (iOS 7.1) from the Apple company, with a screen size of 7.5 cm × 5.0 cm. The app is designed to track changes in RTs over the course of the day and can be used to relate these changes to the components of sleep (Proactive Life LLC, 2012). The user can retrieve all of the data and graphs obtained with the app by e-mail. The task involves tapping on the screen as quickly as possible with the dominant index finger when the stimulus, a sun, appears on the device screen (see Fig. 1). The sun's size is 34 mm in diameter, including 3-mm sunrays (28-mm diameter without sunrays). The sun is centered horizontally and distanced from the top of the screen at one third of the vertical screen dimension. The stimulus size and location are always the same, independent of the mobile device or screen size used. The specific instructions were "Hover the index finger of your dominant hand close (1 cm) to the screen. Tap as quickly as possible on the sun when it appears." Subjects were also asked to hold the iPod in their nondominant hand at lower-abdomen level and to sit upright in a chair in a quiet room without distractions. They were asked to keep their arms free from the armrests and to keep both feet on the ground.

Fig. 1 Picture of the stimulus, a sun, that appears on the device screen in the s2P application. The subject had to tap the sun as quickly as possible with the dominant index finger

No immediate feedback on the subject's RTs were presented after each trial, but the subject viewed their average RT at the end of each session.

S2P offers the flexibility of adjusting the duration of the testing session from 10 s (one trial) to several minutes: up to 60 trials (10 min) on the Android version, and up to 999 trials (166.5 min) on iOS. Considering the fact that a 3-min version of the PVT instead of the classic 10-min version has shown to be a promising tool to differentiate alert from sleepy individuals (Basner & Dinges, 2011), 3-min versions of both s2P and the PVT were selected for the present study. For software and programming reasons, the interstimulus intervals (ISIs) in s2P are set randomly from 4 to 15 s, which differs from the PVT (with ISIs of 2 to 10 s). Data on the touch responsiveness delay are integrated in the app so that the RTs obtained are accurate.

All standard outcome variables were extracted from s2P in order to compare them to the PVT classic outcome variables (Basner & Dinges, 2011): the numbers of lapses and false starts, the mean RT, the reciprocal response time ($RRT = 1/RT$), the 10 % fastest RT, and the 10 % slowest RRT. Only mean RTs over 100 ms were entered in the analyses; those falling below 100 ms were considered false starts. Mean RTs longer than 500 ms were counted as lapses.

Psychomotor vigilance test The PVT is currently the gold standard for objectively measuring alertness and was used in this study as the main tool for validation of the s2P app. The PVT was performed on the PVT-192 device (Ambulatory Monitoring Inc., Ardsley, NY) in a 3-min version. The subjects were told to maintain their dominant index finger on the button (a 1-cm black square on the lower part of the device) and to press as quickly as possible when a red stimulus counter appeared on the small screen (located on the upper part of the device). Subjects were asked to hold the device in their nondominant hand at lower-abdomen level, and to sit upright in a chair in a quiet room without distractions. They were asked to keep their arms free from the armrests and to keep both feet on the ground. Pressing the button automatically stopped the counter, and thus displayed for a 1-s period the RTs of the person. The ISIs varied from 2 to 10 s. The same variables were calculated as those extracted for s2P: the numbers of lapses and false starts, the mean RT, the RRT, the 10 % fastest RT, and the 10 % slowest RRT. The same parameters were set for false starts (<100 ms) and lapses (>500 ms).

Subjective measures of alertness In addition to comparing s2P's performance to outcomes on the classic PVT, the changes in RTs on s2P following sleep loss were compared to various subjective measures of sleepiness. To do this, three widely used questionnaires were administered along with the PVT and s2P: the SSS, the KSS, and a VAS.

The SSS (Hoddes, Zarcone, Smythe, Phillips, & Dement, 1973) is a validated test of subjective sleepiness

with a 7-statement scale ranging from 1 *feeling active, vital, alert, or wide awake* to 7 *No longer fighting sleep, sleep onset soon; having dream-like thoughts*. The subject was told to choose the value corresponding to the statement that best fit how they felt at the current moment; thus, the dependent variable varied from 1 to 7 (MacLean, Fekken, Saskin, & Knowles, 1992). The KSS (Åkerstedt & Gillberg, 1990) is also a validated test measuring the current degree of sleepiness, but it is on a 9-point scale (from 1 *very alert* to 9 *very sleepy, great effort to keep awake, fighting sleep*). States 1, 3, 5, 7, and 9 are labeled, and the intermediate states are only noted. The subject chooses the number that best fits his or her level of sleepiness (Åkerstedt & Gillberg, 1990). A VAS is often used to assess the current degree of sleepiness. On a straight 100-mm line, ranging from *not sleepy at all* to *extremely sleepy*, the individual stated with a dash their current level of sleepiness. The distance in millimeters between the beginning of the scale and the location of the drawn line reflected the level of sleepiness and was considered the dependent variable (Herbert, Johns, & Doré, 1976).

Data analysis

The data analysis and statistical procedure presented here are based on Basner, Mollicone, and Dinges (2011). All analyses were generated using SPSS version 21 and Mathematica version 8. For various reasons (technical problems, late morning awakening, personal obligation, etc.), test bouts were missing for some of the subjects. A total of 333 pairs of PVT and s2P test bouts, out of 396 possible, were included in the final analysis. To take into account multiple comparisons, a conservative significance level of $p = .001$ was used unless otherwise specified.

To verify whether the PVT and s2P had similar sensitivities to the TSD protocol (from 8 AM on the first day to 6 PM the next day), the strength of the relationship between the two devices was determined. Thus, Pearson product-moment correlations between each device's outcomes were conducted (mean performance score from 8 AM to 6 PM for each dependent variable). A significant positive correlation would indicate that both tests measured substantially the same construct, whereas a nonsignificant positive or a negative correlation would indicate a discrepancy between the measures.

The main concern in validation studies for PVT-type tasks is to determine whether the new test is as sensitive to effectively detect cognitive decline as the original PVT (Basner & Dinges, 2012; Basner et al., 2011; Lamond et al., 2008; Loh et al., 2004; Thorne et al., 2005). To evaluate the ability of s2P to differentiate between the sleep-deprived and alert states of the subjects, the test bouts from 8 AM to 10 PM were averaged to reflect the “alert” state, whereas the test bouts from 12 AM to 6 PM were averaged to reflect “sleepiness.” A

similar cutoff had been used in previous studies (Basner et al., 2011) and is based on research showing that performance of the PVT usually begins to decline after 16 h awake (Van Dongen, Maislin, Mullington, & Dinges, 2003).

We used a one-sample t test to determine whether there was a significant difference between the sleep-deprived state and the non-sleep-deprived state, and then calculated the effect sizes for those analyses. The effect size can be interpreted as small (>0.2 and <0.5), medium (≥ 0.5 and <0.8), and large (≥ 0.8) according to Cohen (1988). As a measure of effect size precision, we calculated 95 % non-parametric bootstrap confidence intervals based on 1,000,000 samples (Efron & Tibshirani, 1993).

To verify whether both devices were able to track fatigue-related changes during the TSD, a repeated measures within-groups analysis of variance (ANOVA) of Device (s2P vs. PVT) $\times$ Test Time (test bouts from 8 AM to 6 PM the next day) was calculated on all outcomes variables. To eliminate possible systematic differences in the performance on each device due to confounding factors (hardware, operating mode, etc.), the mean RTs on each device were also presented centered on the average performance in the “alert” state. The centering method had previously been used in other validation studies (Basner et al., 2011; Lamond et al., 2008; Roach et al., 2006). In this case, we again calculated, for each moment, 95 % nonparametric bootstrap confidence intervals based on 1,000,000 samples. The outcomes between both devices were then compared using t tests for paired samples for every test bout from 12 AM to 6 PM. To accommodate multiple calculations, we adjusted the p values using the false discovery rate method (Curran-Everett, 2000).

To verify that the performance on s2P varied in accordance with changes in subjective measures of sleepiness, Pearson product-moment correlations were conducted between each outcome variable on both devices and each dependent variable of the subjective measures (SSS total score, KSS total score, VAS score).

Results

The results presented in Table 1 show that both devices were significantly correlated on all outcomes variables. The strongest relationship was for the mean RTs, and the weakest was for the false starts. For all of the other variables, the relationships were moderate to high.

Table 2 presents within-subjects, one-sample t tests performed between the non-sleep-deprived state, identified as the alert state (test bouts from 8 AM to 10 PM), and the sleep-deprived state, identified as the sleepy state (test bouts from 12 AM to 6 PM the next day) for s2P and the PVT for every outcome variable. Regardless of the device, all variables significantly distinguished between the two states, with the

Table 1 Pearson product-moment correlations between each device's outcomes (mean performance score from 8 AM to 6 PM for each outcome variable)

	sleep-2-Peak		PVT		<i>r</i> Values	<i>p</i> Values
	Mean	(SD)	Mean	(SD)		
Number of lapses	0.850	(1.305)	0.810	(1.791)	.457	<i>p</i> < .001
False starts	1.010	(2.744)	0.67	(1.007)	.196	<i>p</i> < .001
Mean RTs	269.050	(36.515)	254.190	(35.750)	.628	<i>p</i> < .001
RRT	3.784	(0.499)	4.007	(0.527)	.596	<i>p</i> < .001
10 % fastest RT	205.371	(29.346)	193.371	(25.500)	.500	<i>p</i> < .001
10 % slowest RRT	2.865	(0.478)	2.816	(0.451)	.405	<i>p</i> < .001

exception of the number of false starts on s2P (see Table 2). The effect sizes for all outcomes on each device, based on 95 % nonparametric bootstrap confidence intervals of 1,000,000 samples, are presented in Fig. 2. The effect sizes are generally higher for the PVT than for s2P. For both devices, the effect size is lowest for the number of false starts, and highest for the 10 % slowest RRT. The largest discrepancy between the devices was observed for the number of lapses (the effect size is 52 % less for s2P). Overall, with the exception of the number of false starts, even if there seems to be a slight advantage for the PVT over s2P, the effect sizes were medium to large for both tests (see Fig. 2).

The results from the ANOVA are presented in Table 3. Significant main effects were found for device on the outcomes of mean RTs, RRT, and 10 % fastest RT. Significant main effects of test time were found for number of lapses, mean RTs, RRT, 10 % fastest RT, and 10 % slowest RRT. Importantly, we found no interactions between device and time test, suggesting that the devices were similar in their ability to detect changes in alertness across the sleep

deprivation protocol. Both devices significantly varied through the sleep deprivation manipulation, yet there was also a systematic difference between the devices, in which RTs were consistently lower on the PVT (see Fig. 3). Note that Fig. 3 shows every outcome across time, along with 95 % nonparametric bootstrap confidence intervals based on 1,000,000 samples.

To eliminate this systematic difference in performance, due to possible confounding factors (differences in hardware, operating mode, etc.), the mean RTs on each device were centered on the average performance in the alert state. The *t* tests for each time test in the sleepy state when centered on the mean of the alert state showed no significant difference between the devices on each outcome, after correcting for the false discovery rate (see Fig. 4). Figure 4 shows the results for every outcome across time, centered on the average performance in the alert state, along with 95 % nonparametric bootstrap confidence intervals based on 1,000,000 samples.

For all s2P outcomes, significant correlations were found for all outcomes and the SSS, the KSS, and the

Table 2 One-sample *t* tests between the alert state (test bouts from 8 AM to 10 PM) and the sleepy state (test bouts from 12 AM to 6 PM the next day) for s2P and the PVT for every outcome variable

Device	Outcomes	Alert State		Sleepy State		<i>t</i> Values	<i>p</i> Values
		Mean	(SD)	Mean	(SD)		
s2P	Number of lapses	0.560	(0.426)	1.020	(0.879)	2.945	<i>p</i> < .001
	False starts	0.479	(0.607)	1.279	(2.311)	1.714	<i>p</i> = .101
	Mean RTs	253.419	(21.454)	279.629	(29.051)	7.692	<i>p</i> < .001
	RRT	3.991	(0.330)	3.642	(0.380)	7.558	<i>p</i> < .001
	10 % fastest RT	197.618	(17.829)	210.872	(24.451)	4.357	<i>p</i> < .001
	10 % slowest RRT	3.071	(0.333)	2.728	(0.264)	6.667	<i>p</i> < .001
PVT	Number of lapses	0.196	(0.230)	1.187	(0.981)	5.528	<i>p</i> < .001
	False starts	0.516	(0.567)	0.819	(0.720)	3.261	<i>p</i> = .004
	Mean RTs	234.898	(18.725)	270.078	(38.097)	5.158	<i>p</i> < .001
	RRT	4.296	(0.341)	3.838	(0.392)	7.222	<i>p</i> < .001
	10 % fastest RT	182.870	(14.356)	199.213	(21.410)	5.613	<i>p</i> < .001
	10 % slowest RRT	3.056	(0.303)	2.680	(0.250)	6.115	<i>p</i> < .001

Fig. 2 Effect sizes for all outcomes for s2P and the PVT, based on 95 % nonparametric bootstrap confidence intervals of 1,000,000 samples. The relative decrease in effect size from PVT to s2P is indicated as a percentage above the bars for each outcome metric. “Nb” stands for “Number”

VAS (see Table 4). Except for false starts, all PVT outcomes also correlated significantly with the SSS, the KSS, and the VAS (see Table 4).

Discussion

In this study we validated a new tool for measuring alertness called sleep-2-Peak, a smartphone app developed by Therrien and Gartenberg (Proactive Life LLC, 2012). Subjects were sleep-deprived for 35 h, during which they completed s2P along with the gold-standard PVT on every even hour, and with different measures of subjective sleepiness. The results indicated that s2P is a valid tool for differentiating alert from sleepy states in the same individual, and that this tool is just as sensitive as the classic PVT to track fatigue-related changes during extended wakefulness and sleep loss conditions.

One concern about the s2P tool was its shorter duration (3 min in this study), as compared to the standard, 10-min RT tests that have usually been reported in previous research (Basner et al., 2011; Basner & Rubinstein, 2011; Lamond

et al., 2005, 2008; Roach et al., 2006; Thorne et al., 2005). This choice of a shorter duration was deliberate, to address many critiques that the 10-min PVT is impractically long for many contexts (Basner et al., 2011; Basner & Rubinstein, 2011; Lamond et al., 2005, 2008; Roach et al., 2006; Thorne et al., 2005). The duration of an alertness test is crucial, since even severely sleep-deprived individuals will be able to compensate by increasing effort, which will result in adequate performance. Thus, a valid and sensitive alertness test will capture subtle changes in fatigue-related behavior, even in a very brief period of time on task.

Not only did the results from this study show that the 3-min version of s2P can distinguish a sleep-deprived from a non-sleep-deprived state, but they also demonstrated that s2P was as good as the 3-min version of the PVT at distinguishing those two states and measuring variations in performance. The effect sizes obtained for the differentiation between the alert and sleepy states were slightly higher for the PVT than for s2P, but these differences were not significant. Basner et al. (2011) found large effect sizes (over 1.5 for all outcomes) for the 10-min version of the PVT. Since we found the effect sizes for both 3-min versions of the RT tests to be moderate to large, this still suggests a good sensitivity for the shorter versions. Moreover, taking into account that the effect sizes of the 3-min PVT and s2P were not significantly different, we could assume that a 10-min version of s2P would present similarly large effect sizes, but this remains to be tested.

The validity of s2P was also corroborated by a strong relationship found between this test and subjective measures of sleepiness, as was previously found with the PVT (Kaida et al., 2006; Van Dongen et al., 2003), as well as with the PVT in the present study. Interestingly, s2P also seems to be more highly correlated with subjective measures of sleepiness, although after testing, this result was found to be nonsignificant.

Table 3 Repeated measures ANOVAs for Device (s2P vs. PVT) × Test Time (test bouts from 8 AM to 6 PM the next day) on all outcome variables

Outcomes	Device (DF = 1) <i>p</i> Values	Time Test (DF = 17) <i>p</i> Values	Device * Time Test (DF = 17) <i>p</i> Values
Number of lapses	.273	<.001	.736
False starts	.090	.067	.935
Mean RTs	<.001	<.001	.987
RRT	<.001	<.001	.896
10 % fastest RT	<.001	<.001	.999
10 % slowest RRT	.057	<.001	.757

Fig. 3 For each outcome variable, between-subjects averages are shown for each test bout from 8 AM to 6 PM the next day (over the 35 h of total sleep deprivation) for both the PVT and s2P. Errors bars represent 95 % nonparametric bootstrap confidence intervals based on 1,000,000 samples

Moreover, changes were measured by both devices on all outcome variables (with the exception of false starts on both devices), suggesting that both tools were sensitive to the same behavioral changes over the course of extended wakefulness and sleep loss conditions. As we expected on the basis of previous research (Basner & Dinges, 2011; Basner et al., 2011; Lamond et al., 2008), lower mean RTs were observed on both devices when subjects were more alert (during the day), and a slowdown in mean RTs was found over the course of the sleep deprivation protocol. Basner and Dinges (2011) stated that in the 10-min version of the PVT, the RRT is a more sensitive and robust measure of alertness fluctuations than is the RT, but when testing for shorter periods of time, the mean RT seems to be an appropriate measure in general (Loh et al., 2004). Interestingly, in our study, both the RRT and mean RT on the 3-min version of s2P were highly sensitive to changes of alertness.

Despite the close similarities between s2P and the PVT, some discrepancies were observed between the tests. Although both devices behaved similarly on the numbers of lapses and false starts, the results for these outcomes were not as clear or as strong as those for the

mean RTs and the RRT. This was not unexpected, since other studies on the validity of shorter-duration PVTs have suggested that by decreasing time on task, the numbers of lapses and false starts at each session decrease considerably, and thus the test loses its sensitivity to these variables (Basner & Dinges, 2011; Basner et al., 2011). Indeed, lapses and false starts rarely occur in a 3-min version of the PVT. Interestingly, on s2P, false starts are more often made in the sleepy-state condition, possibly because of an increased chance of “brushing” inadvertently on the touch screen.

A systematic difference in the performance on the devices was also found in which the mean RTs obtained on the PVT were systematically lower than the mean RTs obtained on s2P. This effect could be explained by technical and physical differences between the tools. First, the size and weight of each device are different. Also, despite the fact that subjects were asked to hold both devices the same way (with their nondominant hand and the same posture), the use of a touch screen by s2P instead of a button to press on the PVT could also have an impact. The type of programming/software on each device (i.e., the refresh rate

Fig. 4 For each outcome variable, the between-subjects averages were centered around alert performance (average of the test bouts from 8 AM to 10 PM) and are shown for each test bout from 8 AM to 6 PM the next day (over the 35 h of total sleep deprivation) for both the PVT and s2P. Errors bars represent 95 % nonparametric bootstrap confidence intervals

based on 1,000,000 samples. Paired *t* tests were performed on each test bout from 12 AM to 6 PM to test whether the s2P and PVT differed statistically significantly (adjusted for multiple testing). No significant differences between the devices were found

of the touch screen), the size or type of stimulus (sun vs. digits), and the ISI (2 to 10 s for the PVT vs. 4 to 15 s for s2P) are different from one device to the other. Of these factors, the ISI could have some importance. However, the almost perfect convergence of the mean RTs of both

devices when results were centered on the average performance in the alert state strongly suggests that technical, physical, and/or programming factors are probably responsible for discrepancies in RTs (Basner et al., 2011; Lamond et al., 2008; Roach et al., 2006).

Table 4 Pearson product-moment correlations between each outcome variable on s2P and the PVT and the Stanford Sleepiness Scale (SSS) total score, Karolinska Sleepiness Scale (KSS) total score, and visual analogue scale (VAS) total score

Outcomes	SSS		KSS		VAS	
	s2P	PVT	s2P	PVT	s2P	PVT
Number of lapses	.355**	.379**	.347**	.358**	.366**	.373**
False starts	.256**	.100	.257**	.103	.246**	.087
Mean RTs	.447**	.336**	.390**	.325**	.462**	.340**
RRT	-.432**	-.321**	-.347**	-.303**	-.454**	-.332**
10 % fastest RT	.331**	.216**	.290**	.208**	.351**	.212**
10 % slowest RRT	-.393**	-.344**	-.332**	-.314**	-.424**	-.372**

** $p < .001$

One may argue that even when centered on the averaged performance on s2P, RTs seem to be slightly slower than the performance on the PVT. Among the possible factors, one distinctive characteristic of both tests could be a potential explanation for a trend toward slightly lower RTs on the PVT than on s2P. Indeed, the presence of immediate feedback (instant display of the RT each time the button is pressed on the PVT-192) could have an impact on the performance of an individual, by adding motivational and psychological components (Thorne et al., 2005). In fact, Thorne et al. observed during the validation of a PDA-PVT the same phenomenon, in which RTs were lower on the PVT than on their PDA-PVT, but this was inconsistent across the whole experiment. The authors hypothesized that the immediate feedback and possibly other characteristics (e.g., programming, type of stimulus) were responsible for this difference. Eckner, Chandran, and Richardson (2011) looked specifically at the roles of feedback and motivation for RTs. They administered three different RT tests to 31 subjects: (1) a “clinical” RT test, in which the subjects had to catch an 80-cm rod falling through a disk as quickly as possible and the feedback consisted of the visual length of the stem having fallen through the disk once the rod was caught by the subjects, (2) an RT test on a computer with immediate feedback, and (3) an RT test on a computer without any feedback. They also measured subjects’ level of motivation to do each test. This study showed that the clinical RT test and the RT test with feedback had lower RTs, correlated strongly together, and correlated less with the RT test without feedback. In addition, the authors reported greater motivation to do both tests with feedback than to do the RT test without feedback. These authors concluded that feedback is a positive motivational factor that promotes lower RTs and overall helps maintain better performance (Eckner et al., 2011).

Figure 5 shows the relative frequency distributions of the RTs on the PVT and s2P for the alert state (test bouts from 8 AM to 10 PM) and the sleepy state (test bouts from 12 AM to 6 PM). The validity of s2P is supported by this visual representation, which shows very similar distributions of RTs for both the PVT and s2P in the alert and sleepy states. The ability of each test to distinguish between alert and sleepy states in the same individual can be derived by how much the alert-state distribution is separated from the sleepy-state distribution. For the PVT, even though the RT peak frequency in the alert state was greatly different from that in the sleepy state, the RTs with the highest frequencies were 204 ms in the alert time period and 205 ms in the sleepy time period, representing an overall slowing of 1 ms. For the s2P, the RT peak frequency in the alert time period was also greatly different from the RT peak frequency in the sleepy time period. However, the RTs with the highest frequency were 224 ms in the alert time period and 263 ms in the sleepy time period, representing a slowing of 39 ms.

Fig. 5 Relative frequency distributions are shown for the PVT and s2P reaction times for the alert state (test bouts from 8 AM to 10 PM) and sleepy state (test bouts from 12 AM to 6 PM)

These results suggest that despite great similarities in the 3-min PVT and s2P, there were differences in the abilities of the tests to track fatigue-related changes that can occur with extended wakefulness. We suggest that at least one explanation for this finding is that the feedback given during performance on the PVT increases motivation, which increases alertness and thus contributes to lower RTs on the PVT than those obtained on s2P (immediate feedback is displayed on the PVT but not on s2P; Eckner et al., 2011). The performance measured by the PVT may be influenced by motivation, but s2P could be less impacted by this motivation component, giving s2P an added value in certain research and applied settings.

Conclusion

This validation study showed that a 3-min version of s2P, a PVT-type test designed for smartphones, is a valid tool for differentiating alert from sleepy states in the same individual and is as sensitive as the gold-standard PVT for tracking fatigue-related changes during an extended wakefulness and sleep loss condition. We propose that, contrary to the PVT, the performance measured by s2P is less impacted by the confounding effects of motivation, giving this app an added value in certain research and applied settings.

Our validation study used a fourth-generation iPod Touch. Since the stimulus size and location is always the same, independently of the mobile device or screen size, it is highly plausible that every mobile device that may physically resemble the fourth-generation iPod Touch will present results with similar sensitivities to the component of sleep, because the task is performed the same way within a given form factor. Future research should look at validating the use of s2P with

other technologies, such as tablets (iPad) or smart watches, since s2P can be adapted to any form of mobile device that may be developed in the near future.

Author note J.-F.B. and G.F. developed the study concept and design. J.-F.B., D.D., and G.F. were involved in the data collection. All authors were involved in the data analysis and interpretation. J.-F.B. and G.F. drafted and revised the manuscript, and all authors approved the final version for submission.

References

- Åkerstedt, T., & Gillberg, M. (1990). Subjective and objective sleepiness in the active individual. *International Journal of Neuroscience*, 52, 29–37.
- Balkin, T. J., Bliese, P. D., Belenky, G., Sing, H., Thome, D. R., Thomas, M., ... Wesensten, N. J. (2004). Comparative utility of instruments for monitoring sleepiness-related performance decrements in the operational environment. *Journal of Sleep Research*, 13, 219–227.
- Basner, M., & Dinges, D. F. (2011). Maximizing sensitivity of the psychomotor vigilance test (PVT) to sleep loss. *Sleep*, 34, 581–591.
- Basner, M., & Dinges, D. F. (2012). An adaptive-duration version of the PVT accurately tracks changes in psychomotor vigilance induced by sleep restriction. *Sleep*, 35, 193–202. doi:10.5665/sleep.1620
- Basner, M., Mollicone, D., & Dinges, F. D. (2011). Validity and sensitivity of a brief psychomotor vigilance test (PVT-B) to total and partial sleep deprivation. *Journal of Sleep Research*, 69, 949–959. doi:10.1016/j.actaastro.2011.07.015
- Basner, M., & Rubinstein, J. (2011). Fitness for duty: A 3-minute version of the psychomotor vigilance test predicts fatigue-related declines in luggage-screening performance. *Journal of Occupational and Environmental Medicine*, 53, 1146–1154. doi:10.1097/JOM.0b013e31822b8356
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale: Erlbaum.
- Curran-Everett, D. (2000). Multiple comparisons: Philosophies and illustrations. *American Journal of Physiology—Regulatory, Integrative and Comparative Physiology*, 279, R1–R8.
- Dinges, D. F. (1995). An overview of sleepiness and accidents. *Journal of Sleep Research*, 4, 4–14.
- Dinges, D. F., & Powell, J. W. (1985). Microcomputer analyses of performance on a portable, simple visual RT task during sustained operations. *Behavior Research Methods, Instruments, & Computers*, 17, 652–655. doi:10.3758/BF03200977
- Dorrian, J., Rogers, N. L., & Dinges, D. F. (2005). Psychomotor vigilance performance: A neurocognitive assay sensitive to sleep loss. In C. Kushida (Ed.), *Sleep deprivation: Clinical issues, pharmacology and sleep loss effects* (pp. 39–70). New York: Marcel Dekker.
- Eckner, J. T., Chandran, S., & Richardson, J. K. (2011). Investigating the role of feedback and motivation in clinical reaction time assessment. *PM R*, 3, 1092–1097. doi:10.1016/j.pmrj.2011.04.022
- Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.
- Gunzelmann, G., Moore, L. R., Gluck, K. A., Van Dongen, H. P. A., & Dinges, D. F. (2011). Fatigue in sustained attention: Generalizing mechanisms for time awake to time on task. In P. L. Ackerman (Ed.), *Cognitive fatigue: Multidisciplinary perspectives on current research and future application* (pp. 83–101). Washington, DC: American Psychological Association.
- Herbert, M., Johns, M. W., & Doré, C. (1976). Factor analysis of analogue scales measuring subjective feelings before and after sleep. *British Journal of Medical Psychology*, 49, 373–379.
- Hoddes, E., Zarcone, V., Smythe, H., Phillips, R., & Dement, W. C. (1973). Quantification of sleepiness: A new approach. *Psychophysiology*, 10, 431–436.
- Ingraham, N. (2015). Apple's app store has passed 100 billion app downloads. Retrieved from www.theverge.com/2015/6/8/8739611/apple-wwdc-2015-stats-update
- Kaida, K., Takahashi, M., Åkerstedt, T., Nakata, A., Otsuka, Y., Haratani, T., & Fukasawa, K. (2006). Validation of the Karolinska sleepiness scale against performance and EEG variables. *Clinical Neurophysiology*, 117, 1574–1581.
- Khitrov, M. Y., Laxminarayan, S., Thorsley, D., Ramakrishnan, S., Rajaraman, S., Wesensten, N. J., & Reifman, J. (2014). PC-PVT: A platform for psychomotor vigilance task testing, analysis, and prediction. *Behavior Research Methods*, 46, 140–147. doi:10.3758/s13428-013-0339-9
- Lamond, N., Dawson, D., & Roach, G. D. (2005). Fatigue assessment in field: Validation of a hand-held electronic psychomotor vigilance task. *Aviation, Space, and Environmental Medicine*, 76, 486–489.
- Lamond, N., Jay, S. M., Dorrian, J., Ferguson, S. A., Roach, G. D., & Dawson, D. (2008). The sensitivity of a palm-based psychomotor vigilance task to severe sleep loss. *Behavior Research Methods*, 40, 347–352. doi:10.3758/BRM.40.1.347
- Loh, S., Lamond, N., Dorrian, J., Roach, G., & Dawson, D. (2004). The validity of psychomotor vigilance tasks of less than 10-minute duration. *Behavior Research Methods, Instrument, & Computers*, 36, 339–346. doi:10.3758/BF03195580
- MacLean, A. W., Fekken, G. C., Saskin, P., & Knowles, J. B. (1992). Psychometric evaluation of the Stanford sleepiness scale. *Journal of Sleep Research*, 1, 35–39.
- Philip, P., & Åkerstedt, T. (2006). Transport and Industrial safety, how are they affected by sleepiness and sleep restriction? *Sleep Medicine*, 10, 347–356.
- Proactive Life LLC. (2012). Sleep-2-Peak. Retrieved May 20, 2016, from <https://itunes.apple.com/us/app/sleep-2-peak/id557632863?mt=8>
- Roach, G. D., Dawson, D., & Lamond, N. (2006). Can a shorter psychomotor vigilance task be used as a reasonable substitute for the ten-minute psychomotor vigilance task? *Chronobiology International*, 23, 1379–1387. doi:10.1080/07420520601067931
- Thome, D. R., Johnson, D. E., Redmond, D. P., Sing, H. C., Belenky, G., & Shapiro, J. M. (2005). The Walter Reed palm-held psychomotor vigilance test. *Behavior Research Methods*, 37, 111–118. doi:10.3758/BF03206404
- Van Dongen, H. P. A., Maislin, G., Mullington, J. M., & Dinges, D. F. (2003). The cumulative cost of additional wakefulness: Dose-response effects on neurobehavioral functions and sleep physiology from chronic sleep restriction and total sleep deprivation. *Sleep*, 26, 117–129.
- ZDNet. (2015). *Chiffres clés: Les ventes de mobiles et de smartphones*. Retrieved from www.zdnet.fr/actualites/chiffres-cles-les-ventes-de-mobiles-et-de-smartphones-39789928.htm