Maximizing Sensitivity of the Psychomotor Vigilance Test (PVT) to Sleep Loss

Mathias Basner, MD, PhD, MSc; David F. Dinges, PhD

Division of Sleep and Chronobiology, Department of Psychiatry, University of Pennsylvania School of Medicine, Philadelphia, PA

Study Objectives: The psychomotor vigilance test (PVT) is among the most widely used measures of behavioral alertness, but there is large variation among published studies in PVT performance outcomes and test durations. To promote standardization of the PVT and increase its sensitivity and specificity to sleep loss, we determined PVT metrics and task durations that optimally discriminated sleep deprived subjects from alert subjects.
Design: Repeated-measures experiments involving 10-min PVT assessments every 2 h across both acute total sleep deprivation (TSD) and 5 days of chronic partial sleep deprivation (PSD).
Setting: Controlled laboratory environment.
Participants: 74 healthy subjects (34 female), aged 22-45 years.
Interventions: TSD experiment involving 33 h awake (N = 31 subjects) and a PSD experiment involving 5 nights of 4 h time in bed (N = 43 subjects).
Measurements and Results: In a paired *t*-test paradigm and for both TSD and PSD, effect sizes of 10 different PVT performance outcomes were calculated. Effect sizes were high for both TSD (1.59–1.94) and PSD (0.88–1.21) for PVT metrics related to lapses and to measures of psychomotor speed, i.e., mean 1/RT (response time) and mean slowest 10% 1/RT. In contrast, PVT mean and median RT outcomes scored low to moderate effect sizes influenced by extreme values. Analyses facilitating only portions of the full 10-min PVT indicated that for some outcomes, high effect sizes could be achieved with PVT durations considerably shorter than 10 min, although metrics involving lapses seemed to profit from longer test durations in TSD.
Conclusions: Due to their superior conceptual and statistical properties and high sensitivity to sleep deprivation, metrics involving response speed and lapses should be considered primary outcomes for the 10-min PVT. In contrast, PVT mean and median metrics, which are among the most widely used outcomes, should be avoided as primary measures of alertness. Our analyses also suggest that some shorter-duration PVT versions may be sensitive to sleep loss, depending on the outcome variable selected, although this will need to be confirmed in comparative analyses of separate duration versions of the PVT. Using both sensitive PVT metrics and optimal test durations maximizes the sensitivity of the PVT to sleep loss and therefore potentially decreases the sample size needed to detect the same neurobehavioral deficit. We propose criteria to better standardize the 10-min PVT and facilitate between-study comparisons and meta-analyses.
Keywords: PVT, psychomotor vigilance, sleep deprivation, alertness, attention, lapse, response speed, response time, sensitivity, power
Citation: Basner M; Dinges DF. Maximizing sensitivity of the psychomotor vigilance test (PVT) to sleep loss. *SLEEP* 2011;34(5):581-591.

INTRODUCTION

There is extensive evidence that the neurobehavioral consequences of sleep loss can be measured in certain aspects of cognitive functioning.[1-3] Among the most reliable effects of sleep deprivation is degradation of attention,[2,4] especially vigilant attention as measured by the 10-min psychomotor vigilance test (PVT).[5,6] The effects of sleep loss on PVT performance appear to be due to variability in maintenance of the alert state (i.e., alerting network),[5] and can include deficits in endogenous selective attention,[7,8] but they may also occur in attention involved in orienting to sensory events (i.e., orienting network), and attention central to regulating thoughts and behaviors (i.e., executive network).[9-11] These multidimensional features of attention suggest it has a fundamental role in a wide range of cognitive functions, which may be the mechanisms by which sleep loss affects a range of performances, although it remains controversial whether impairment due to sleep deprivation is generic to all cognitive processes subserved by attentional processes.[12]

Submitted for publication August, 2010
Submitted in final revised form February, 2011
Accepted for publication February, 2011
Address correspondence to: Mathias Basner, MD, PhD, MSc, Division of Sleep and Chronobiology, Department of Psychiatry, University of Pennsylvania School of Medicine, 1013 Blockley Hall, 423 Guardian Drive, Philadelphia, PA 19104-6021; Tel: (215) 573-5866; Fax: (215) 573-6410; E-mail basner@mail.med.upenn.edu

The PVT[5,13-15] has become arguably the most widely used measure of behavioral alertness owing in large part to the combination of its high sensitivity to sleep deprivation[5,6] and its psychometric advantages over other cognitive tests. The standard 10-min PVT measures sustained or vigilant attention by recording response times (RT) to visual (or auditory) stimuli that occur at random inter-stimulus intervals (ISI).[6,13,16,17] It is not entirely accurate to describe the PVT as merely simple RT. The latter is a generic phrase historically used to refer to the measurement of the time it takes to respond to a stimulus with one type of response (in contrast a complex RT task can require different responses to different stimuli). A simple RT test assumes no specific number of RTs—in fact, it can be based on a single RT. Similar to simple RT, the PVT relies on a stimulus (typically visual) and an RT (typically a button press), but it also relies on sampling many responses to stimuli that appear at a random ISI within a pre-specified ISI range, and that therefore occur over a period of time (i.e., 10 minutes in terms of the most commonly used PVT). Therefore, time on task and ISI parameterization instantiate the "vigilance" aspect of the PVT. Response time to stimuli attended to has been used since the late 19th century in sleep deprivation research[16,18,19] because it offers a simple way to track changes in behavioral alertness caused by inadequate sleep, without the confounding effects of aptitude and learning.[5,6,15] Moreover, the 10-min PVT[13] has been shown to be highly reliable, with intra-class correlations for key metrics such as lapses measuring test-retest reliability above 0.8.[6]

PVT performance also has ecological validity in that it can reflect real-world risks, because deficits in sustained atten-

Table 1—Frequency of PVT outcome metrics reported in 141 journal publications published since 1986[1]

Outcome	Frequency
Number of Lapses[2]	66.7%
Mean RT[3]	40.4%
Mean 1/RT	30.5%
Fastest 10% RT	29.8%
Median RT	28.4%
Slowest 10% RT	19.9%
Slowest 10% 1/RT	12.8%
Number of False Starts	9.2%
Fastest 10% 1/RT	5.0%
Lapse Probability[4]	4.3%
Other	23.4%

[1]The 141 articles are the result of a Thomson ISI search on "psychomotor vigilance" in title, abstract, or keywords of peer-reviewed articles published since 1986 performed on 30 April 2010.
[2]Lapses were most commonly defined as response times > 500 ms or ≥ 500 ms, although individual studies used different definitions.
[3]RT = response time
[4]Lapse probability is usually calculated as the number of lapses divided by the number of valid stimuli.

tion and timely reactions adversely affect many applied tasks, especially those in which work-paced or timely responses are essential (e.g., stable vigilant attention is critical for safe performance in all transportation modes, many security-related tasks, and a wide range of industrial tasks). Lapses in attention as measured by the PVT can occur when fatigue is caused by either sleep loss or time on task,[16,20,21] which are the two factors that make up virtually all theoretical models of fatigue in real-world performance. There is a large body of literature on attentional deficits having serious consequences in applied settings.[22-25]

Sleep deprivation induces reliable changes in PVT performance, causing an overall slowing of response times, a steady increase in the number of errors of omission (i.e., lapses of attention, historically defined as RTs ≥ twice the mean RT or 500 ms), and a more modest increase in errors of commission (i.e., responses without a stimulus, or false starts).[26,27] These effects can increase as task duration increases,[28] and they form the basis of the state instability theory.[5,6,13-15,18] According to this theory, several competing systems influence behavior during periods of sleep loss, 2 of the most important being the involuntary drive to fall asleep and a counteracting top-down drive to sustain alertness.[5] The interaction of these sleep-initiating and wake-maintaining systems leads to unstable sustained attention as manifested in longer RTs occurring stochastically throughout each PVT performance bout.[5,15] Neuroimaging studies reveal that slowed responses on visual attention tasks—including the PVT—during sleep deprivation are associated with changes in neural activity in distributed brain regions that can include frontal and parietal control regions, visual and insular cortices, cingulate gyrus, and the thalamus.[8,29-31]

The 10-min PVT[5,6,13] has been shown to be sensitive to both acute total sleep deprivation (TSD)[15,27,32] and chronic partial sleep deprivation (PSD)[14,27,32-35]; to be affected both by sleep homeostatic and circadian drives[36,37]; to reveal large inter-subject variability in the response to sleep loss[38-40]; to demonstrate the effects of jet lag and shift work[41]; and to reveal improvements in alertness after wake-promoting interventions[42-44] and recovery from sleep loss,[45,46] and after initiation of CPAP treatment in patients with obstructive sleep apnea (OSA).[47]

There are many published reports on PVT performance relative to sleep quantity, quality and circadian placement, but there is considerable variability among these reports in (1) the specific PVT performance metrics used as outcomes, (2) the duration of the PVT task,[48] and (3) the platform on which the PVT is administered and timed (i.e., various computer-based PVTs,[13] and handheld devices such as the PVT-192[48-51] and the Palm PVT[52]). An analysis of 141 journal manuscripts reporting PVT results, published in the past 25 years (since 1986), shows great variability in the use of PVT outcome metrics—the most commonly reported PVT outcomes from these studies are listed in Table 1.

PVT performance metrics likely differ widely in their statistical properties and therefore in their capability to differentiate sleep deprived from alert subjects. However, we are not aware of any systematic comparison of the sensitivity of different PVT metrics to sleep loss. We therefore determined the extent to which 10 PVT performance metrics were sensitive to both acute TSD and chronic PSD. Our goal was to recommend a small set of PVT outcome metrics with superior statistical properties that should be routinely used and reported in studies. We also evaluated the sensitivity of each outcome for different portions of the full 10-min PVT to determine feasibility of shorter duration PVT versions for detecting deficits in alertness relative to sleep deprivation. Performance on the PVT deteriorates faster in sleep deprived than in alert subjects with time on task (see Figure 1). However, due to marked inter-individual differences in the vulnerability to sleep loss,[38] between-subject variability in PVT performance increases with time on task at the same time. As the power of the PVT to detect a statistically significant difference between alert and sleep deprived states increases with larger performance differences between states but decreases with increasing variability of these differences, it is unclear whether PVT sensitivity will always increase with increasing PVT duration. More likely, there will be a test duration with an optimal ratio of size and variability of the differences. Maximizing the power of the PVT to differentiate sleep deprived and alert subjects by choosing optimal test durations and outcome metrics would allow detecting the same difference in cognitive performance degradation with smaller sample sizes.

Finally, many factors other than test duration and outcome metric (e.g., hardware, programming of the PVT, and definitions for calculating PVT outcome metrics) influence PVT performance.[52] These factors differ widely between different versions of the PVT. Hence, the PVT is not very well standardized, which complicates meta-analyses of studies using the PVT. Therefore, we provide definitions and list the criteria we have been using for many years for our 10-min PVT (see Table 3). We hope that researchers will adopt these criteria in future research and therefore enhance standardization of the 10-min PVT.

METHODS

Subjects and Protocol

Acute total sleep deprivation (TSD) protocol

TSD data were gathered in a study on the effects of night work and sleep loss on threat detection performance on a simulated luggage screening task (SLST). A detailed description of the study is published elsewhere.[33] This analysis is based on data gathered on N = 31 subjects (mean age ± SD = 31.1 ± 7.3 y, 18 female). Study participants stayed in the research lab for 5 consecutive days, which included a 33-h period of TSD. The study started at 08:00 on day 1 and ended at 08:00 on day 5. A 33-h period of total sleep deprivation started either on day 2 (N = 22) or on day 3 (N = 9) of the study. Except for the sleep deprivation period, subjects had 8-h sleep opportunities between 00:00 and 08:00. The first sleep period was monitored polysomnographically to exclude possible sleep disorders.

Chronic partial sleep deprivation (PSD) protocol

PSD data were obtained from N = 43 healthy adults (16 females) who averaged 30.5 ± 7.3 y (mean ± SD) and were part of a laboratory protocol involving 5 consecutive nights of sleep restricted to 4 h per night (04:00 to 08:00). A detailed description of the experimental procedures is published elsewhere.[46]

In both TSD and PSD experiments, subjects were free of acute and chronic medical and psychological conditions, as established by interviews, clinical history, questionnaires, physical exams, and blood and urine tests. They were studied in small groups (4-5) while they remained for days in the Sleep and Chronobiology Laboratory at the Hospital of the University of Pennsylvania. Throughout both experiments subjects were continuously monitored by trained staff to ensure adherence to each experimental protocol. They wore wrist actigraphs throughout each protocol. Meals were provided at regular times throughout the protocol, caffeinated foods and drinks were not allowed, and light levels in the laboratory were held constant during scheduled wakefulness (< 50 lux) and sleep periods (< 1 lux). Ambient temperature was maintained between 22°and 24°C.

In both TSD and PSD experiments subjects completed 30-min bouts of a neurobehavioral test battery (NTB) that included a 10-min PVT every 2 h during scheduled wakefulness. Between neurobehavioral test bouts, subjects were permitted to read, watch movies and television, play card/board games and interact with laboratory staff to help them stay awake, but no naps/sleep or vigorous activities (e.g., exercise) were allowed.

All participants were informed about potential risks of the study, and a written informed consent and IRB approval were obtained prior to the start of the study. They were compensated for their participation, and monitored at home with actigraphy, sleep-wake diaries, and time-stamped phone records for time to bed and time awake during the week immediately before the PSD study.

PVT

To avoid problems of uncertainty regarding the accuracy of timing of the test platform, we utilized a precise computer-based version of the 10-min PVT. Subjects were instructed to monitor a red rectangular box on the computer screen, and

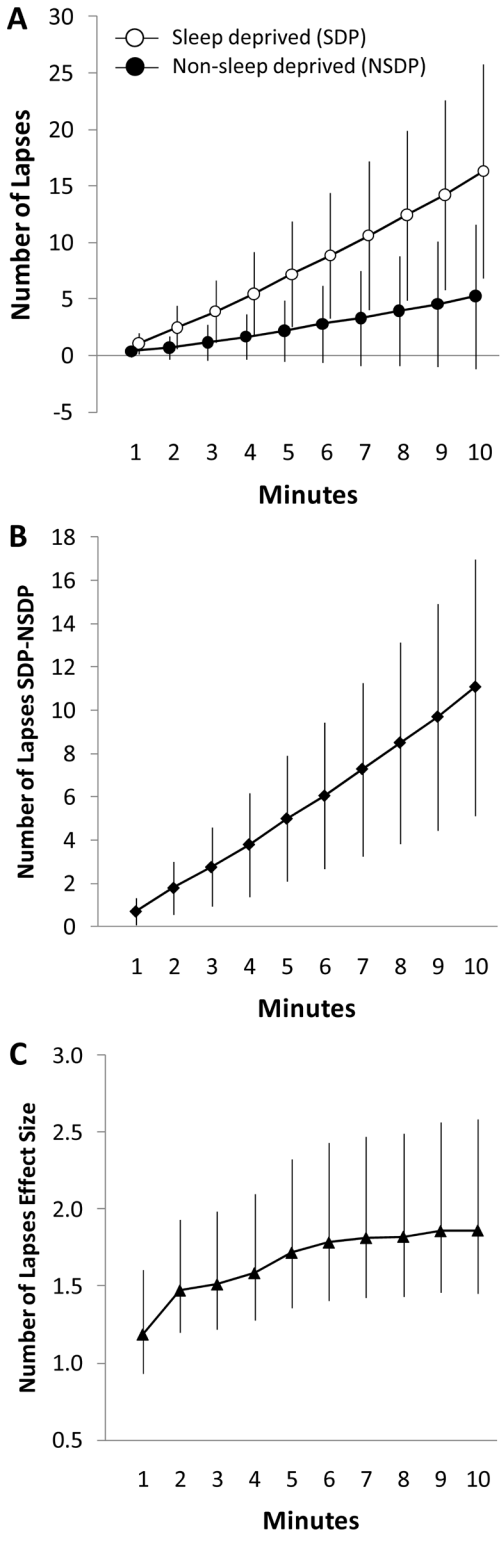

Figure 1—The analyses shown in A, B, and C are based on the TSD study and were restricted to the first 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10 min of the 10-min PVT (abscissa) **(A)** The number of lapses and their standard deviation are shown for the sleep deprived and the non–sleep deprived state. **(B)** The within-subject differences between sleep deprived and non–sleep deprived states of the number of lapses and their standard deviations are shown. **(C)** Effect sizes calculated as the within-subject differences between sleep deprived and non–sleep deprived states divided by their standard deviation are shown for lapses including 95% nonparametric bootstrap confidence intervals.

press a response button as soon as a yellow stimulus counter appeared on the CRT screen, which stopped the counter and displayed the RT in milliseconds for a 1-s period (see Table 3). The inter-stimulus interval, defined as the period between the last response and the appearance of the next stimulus, varied randomly from 2-10 s. The subject was instructed to press the button as soon as each stimulus appeared, in order to keep the RT as low as possible but not to press the button too soon (which yielded a false start warning on the display).

Outcome metrics

A PVT response was regarded valid if RT was ≥ 100 ms. Responses without a stimulus or RTs < 100 ms were counted as false starts (errors of commission). Pressing the wrong button or failing to release the button for 3 s or longer were counted as errors and excluded from the analysis. Lapses (errors of omission) were defined as RTs ≥ 500 ms. The millisecond counter timed out after 30 s without a response, and a sound was played back to alert the subject. The following PVT outcome metrics were assessed and included in our analyses: (1) median RT, (2) mean RT, (3) fastest 10% RT, (4) mean 1/RT (also called reciprocal response time or response speed), (5) slowest 10% 1/RT, (6) number of lapses, (7) lapse probability (i.e., number of lapses divided by the number of valid stimuli, excluding false starts), (8) number of false starts, (9) number of lapses and false starts, and (10) performance score, defined as 1 minus the number of lapses and false starts divided by the number of valid stimuli (including false starts).

The reciprocal transform (1/RT) was one of the first PVT outcomes found to be sensitive to total and partial sleep loss.[42] It emphasizes slowing in the optimum and intermediate response domain and it substantially decreases the contribution of long lapses, which is why the slowest 10% of RTs are usually reciprocally transformed. For calculating mean 1/RT and slowest 10% 1/RT, each RT (ms) was divided by 1,000 and then reciprocally transformed. The transformed values were then averaged. False starts were investigated as they were shown to increase in number during sleep deprivation and show the same pattern of circadian modulation as lapses.[15] As the duration of the PVT is fixed, long lapses will reduce the number of stimuli and therefore also the maximum possible number of lapses. Calculating lapse probability (i.e., expressing lapses relative to the number of stimuli) potentially corrects for this reduction in the number of stimuli.

Data Analyses and Statistical Procedures

To investigate the power of different PVT outcome metrics and test durations to differentiate sleep deprived from alert subjects, in the TSD study test bouts 1 to 7 (09:00 to 21:00) were averaged within subjects to reflect the non–sleep deprived state and test bouts 8 to 17 (23:00 to 17:00 on the next day) were averaged within subjects to reflect the sleep deprived state. This decision was based on visual inspection of the data and on reports that PVT performance decreases after 16 h of wakefulness.[27] After excluding one subject that dropped out after 26 h awake, the data for the remaining 31 subjects were complete. For the PSD study, daily averages of outcome variables were computed within subjects over the test bouts administered at 12:00, 16:00, and 20:00. The 08:00 test bout was not used because of possible sleep inertia effects. Average performance on baseline day 2 (BL2) reflected the non–sleep deprived state, while average performance on the day after sleep restriction night (R5) reflected the sleep deprived state. Only test bouts that existed in both conditions (non-sleep deprived and sleep deprived) were used for averaging to exclude differential effects of circadian modulation on PVT performance. For example, if the 16:00 test bout was missing for a subject in R5, the 16:00 test bout was also not used for averaging in BL2 for that subject, even if it existed. Overall, 23 out of 903 scheduled test bouts (2.5%) were missing.

Within subject comparisons are very common in sleep (deprivation) research. They control for some of the variance associated with inter-individual differences, and therefore the same effect can usually be found with smaller sample sizes. A paired *t*-test would be a valid method to investigate whether there is a statistically significant difference between non-sleep deprived and sleep deprived conditions. In the paired *t*-test, differences of outcome values between non-sleep deprived and sleep deprived conditions are calculated within subjects, and these differences are tested with a one-sample t-test against zero. With a given type I error rate α and a fixed number of subjects, the power of the paired *t*-test (i.e., the probability to detect a difference between conditions if there is a difference) depends only on the effect size. Effect size is calculated as the average of within-subject differences divided by the standard deviation of within-subject differences (i.e., the average of within-subject difference is expressed in standard deviation units). Effect size therefore increases with the magnitude of within-subject differences and decreases with increasing variability (i.e., the standard deviation) of the differences.

The one-sample *t*-test is the most powerful test available (i.e., it outperforms nonparametric tests that could be used alternatively) when its test assumptions are met. It requires (a) random sampling from a defined population, (b) interval or ratio scale of measurement, and (c) normally distributed population data (note that differences of two samples may be normally distributed even if the original samples are not). However, the one-sample *t*-test is relatively robust in terms of violations of the above assumptions. Also, it requires the distribution of sample means to be normal, not the sample itself. According to the Central Limit Theorem, the distribution of sample means will be normal even if the sample itself is not if sample size is large (usually N > 30). The samples of both the TSD (N = 31) and the PSD (N = 43) study were large enough for the Central Limit Theorem to apply.

Based on the above definitions of sleep deprived and non–sleep deprived states, we calculated the unit-less effect size for the 10 PVT outcome metrics both for the TSD and PSD studies, for the full 10-min PVT, and restricting the analyses to the first 1, 2, 3, 4, 5, 6, 7, 8, and 9 min of the 10 min PVT. As a measure of effect size precision, we calculated 95% nonparametric bootstrap confidence intervals based on 1,000,000 samples according to Efron and Tibshirani.[53] In contrast to standard confidence intervals, bootstrap confidence intervals have the advantage that they are range preserving (i.e. intervals always fall within the allowable range of the investigated variable) and do not enforce symmetry. Effect sizes for slowest 10% 1/RT and fastest 10% RT were not calculated for the 1-min test duration, be-

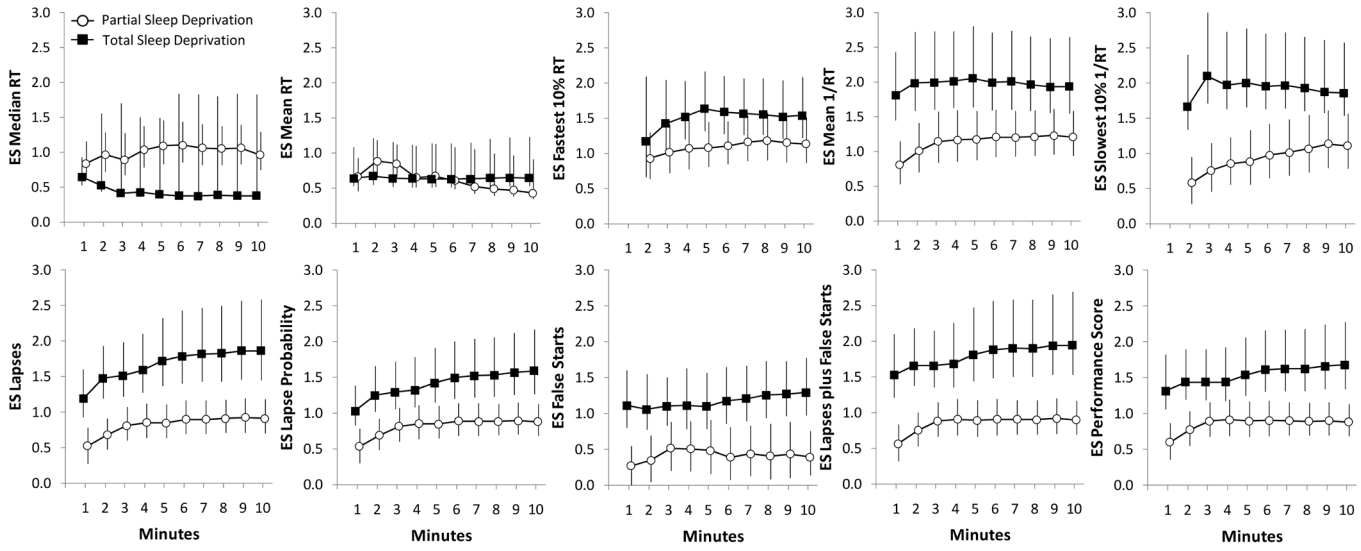

Figure 2—Effect sizes (ES) and 95% nonparametric bootstrap confidence intervals are shown for 10 outcome metrics of the PVT for both partial (open circles) and total (black squares) sleep deprivation depending on the analyzed portion of the 10-min PVT. Effect sizes of Mean 1/RT, Slowest 10% 1/RT, and the Performance Score were multiplied by -1 to facilitate comparisons.

Table 2—Rank order of effect sizes of 10 PVT outcome metrics for the 10-min PVT in both acute total and chronic partial sleep deprivation

	Acute Total Sleep Deprivation		Chronic Partial Sleep Deprivation	
Rank	**Outcome Metric**	**Effect Size (95% CI)**	**Outcome Metric**	**Effect Size (95% CI)**
1	Number of Lapses and False Starts	1.94 (1.53; 2.69)	Mean 1/RT	1.21 (0.94; 1.59)
2	Mean 1/RT	1.93 (1.55; 2.65)	Fastest 10% RT	1.13 (0.87; 1.51)
3	Number of Lapses	1.86 (1.45; 2.59)	Slowest 10% 1/RT	1.11 (0.78; 1.57)
4	Slowest 10% 1/RT	1.86 (1.54; 2.58)	Median RT	0.96 (0.74; 1.29)
5	Performance Score	1.68 (1.34; 2.28)	Number of Lapses	0.91 (0.71; 1.18)
6	Lapse Probability	1.59 (1.27; 2.16)	Number of Lapses and False Starts	0.90 (0.69; 1.17)
7	Fastest 10% RT	1.54 (1.22; 2.08)	Performance Score	0.88 (0.69; 1.13)
8	Number of False Starts	1.29 (0.98; 1.78)	Lapse Probability	0.88 (0.69; 1.13)
9	Mean RT	0.64 (0.53; 1.23)	Mean RT	0.43 (0.35; 0.91)
10	Median RT	0.38 (0.33; 1.83)	Number of False Starts	0.39 (0.14; 0.76)

RT, response time; CI, confidence interval; Effect sizes of Mean 1/RT, Slowest 10% 1/RT, and the Performance Score were multiplied by -1 to facilitate comparisons. Nonparametric 95% bootstrap confidence intervals are shown in parenthesis.

cause the first min contained only 9 stimuli on average. Effect sizes for mean 1/RT, slowest 10% 1/RT, and the performance score were multiplied by -1 to facilitate comparisons between outcome metrics.

Graphs showing the evolution of the effect size for each of the 10 outcome metrics during 33 h of TSD (relative to the average of bouts 1-7, i.e., the non–sleep deprived state) and across the 7 days (BL1 to R5) of the PSD protocol (relative to BL2) were generated for the full 10-min PVT only. Finally, we performed retrospective power calculations for the outcome metrics mean 1/RT and number of lapses, for the full 10-min PVT, and for both total and partial sleep deprivation.

RESULTS

The evolution of the number of lapses and their standard deviation is shown as an example for the sleep deprived and the non–sleep deprived state in Figure 1A. Both the mean difference between conditions and the standard deviation of these differences increased with time on task (see Figure 1B). For lapses, the increase in the average differences outweighed the increase in the standard deviation of the differences, and hence the effect size increased with increasing PVT duration (see Figure 1C).

However, this is not true for all outcome metrics. Effect sizes of the 10 outcome metrics depending on PVT duration are shown in Figure 2 for both the TSD and the PSD study. Except for median RT and mean RT, TSD effect sizes exceeded those observed during PSD. There were substantial differences in the ability of the different outcome metrics to differentiate sleep deprived from alert subjects. In general, the reciprocal metrics mean 1/RT and slowest 10% 1/RT scored the highest effect sizes, matched or closely followed by outcomes involving number of lapses and the fastest 10% RT (see Table 2 for a comparison of effect sizes for the full 10-min PVT). Lapse probability and the performance score performed somewhat worse than the

Table 3—Criteria for the analysis of the 10-min PVT

Standard outcomes	Mean 1/RT[1] and number of lapses
Stimulus	Visual millisecond counter in rectangular box
Test duration	10 min (the tests stops with the last response after an elapsed total time of 10 min)
Inter-stimulus interval	2-10 s (defined as the period between the last response and the appearance of the next stimulus)[2]
Feedback	The response time is displayed for 1 s. This period is part of the next inter-stimulus interval.
Errors of commission (false starts)[4]	Responses without a stimulus or response times < 100 ms.[3] „FS“ is displayed for 1 second (this period is part of the next inter-stimulus interval).
Errors of omission (lapses)	Response times ≥ 500 ms[3]
Time out	The millisecond counter times out after 30,000 ms without a response. “OVERRUN” is displayed for 1 second (this period is part of the next inter-stimulus interval) and a sound is played back to alert the subject. The stimulus is counted as valid, i.e., as a lapse with a response time of 30,000 ms.
Button fail-to-release[4]	“BUTTON” is displayed after the response button has not been released for 3 s and a signal is continuously played back until the button is released. The new inter-stimulus interval starts once the button is released.
Wrong key press[4]	“ERR” is displayed for 1 s if the wrong response key is pressed (this period is part of the next inter-stimulus interval). If the wrong key was pressed prematurely “FS/ERR” is displayed instead of “ERR.”

RT, response time.
[1]Individual raw RTs [in ms] are first transferred to response speed by calculating 1000/RT. These response speeds are then averaged.
[2]In the PVT used in the total and partial sleep deprivation studies, we randomly drew full second inter-stimulus intervals (ISI), i.e., 2, 3, 4, 5, 6, 7, 8, 9, or 10 s. This completely random process may introduce variance in the number of stimuli between test bouts, where bouts with a low number of stimuli will be biased towards longer ISIs. In a revised version of the test, a block randomization technique is used that guarantees that the number of stimuli is similar between bouts without losing the random component of ISIs.
[3]For auditory stimuli, this threshold would have to be lower.
[4]These events do not count as valid stimuli, and thus they do not contribute to the calculation of mean 1/RT.

generic metrics they are based on (number of lapses and number of lapses plus false starts). Mean RT, median RT, and the number of false starts scored the lowest effect sizes, especially during TSD (although they would still be classified as small to medium according to Cohen‘s criteria[54]).

Maximum effect sizes were not always observed after the full 10 min PVT duration, although outcomes involving lapses and/or false starts seemed to profit from longer test durations, especially during TSD. For all other outcome metrics, the maximum effect size was observed with shorter than 10-min test durations during both TSD and PSD. For PSD, maximum effect size was reached with PVT duration < 10 min for all outcome variables. Also, the effect size curves depicted in Figure 2 generally involved saturating profiles (i.e., effect size did not substantially increase further after a certain PVT duration was reached). This especially applies to the PSD study, where effect sizes did not increase substantially after the third minute of the test for the fastest 10% RT, mean 1/RT, and all outcome measures based on lapses.

For the full 10-min PVT and TSD, the highest overall effect size was observed for number of lapses and false starts (1.94), followed by mean 1/RT (1.93), and number of lapses and slowest 10% 1/RT (both 1.86, see Table 2). Mean RT and median RT, 2 commonly reported PVT performance outcomes (see Table 1), had the lowest effect sizes in the TSD experiment. On closer inspection of the raw data, this appeared to be caused by 2 extreme observations: median RT and mean RT differed by 2217 ms and 3066 ms, respectively, between sleep deprived and non-sleep deprived states in one subject, and by 408 ms and 2307 ms in another subject. The average difference in the remaining subjects was 67 ms for median RT and 264 ms for mean RT. The effect size calculations were therefore repeated for all outcome metrics without these 2 extreme observations (for the TSD condition and for the full 10-min PVT only). Effect sizes for median RT (1.62 vs. 0.38) and mean RT (1.09 vs. 0.64) increased markedly compared to the initial analyses, although they still ranked in position 8 (median RT) and 10 (mean RT) relative to the other PVT outcome metrics. At the same time, excluding the two extreme observations reduced effect sizes of 5 of the other 8 outcome metrics (mean 1/RT, slowest 10% 1/RT, number of lapses, number of false starts, and number of lapses and false starts). Additionally, we repeated the analysis (including the extreme observations) with a base e log-transformation of median RT and mean RT. Again, effect sizes increased for both median RT (0.82 vs. 0.38) and mean RT (1.33 vs. 0.64), but they still ranked in position 8 (mean RT) and 10 (median RT) relative to the other PVT outcome metrics.

Figures 3 and 4 show the evolution of effect sizes of the different outcome metrics during 33 h of TSD and across the 7 days (BL1 to R5) of the PSD protocol for the full 10 min PVT. In the TSD study (Figure 3), there was a steep increase of effect sizes after 15 h to 17 h awake (corresponding to 23:00 and 01:00 on the first day of sleep restriction). The highest effect sizes were observed either after 25 h or after 27 h awake (corresponding to 09:00 or 11:00 on the second day of sleep restriction). The lowest effect sizes (indicating high alertness) were observed after 11 h awake in 8 out of 10 outcome measures (corresponding to 19:00 on the first day of restriction). In the PSD study (Figure 4), effect sizes generally increased across the days of restriction, less clear for median RT, mean RT, and the number of false starts, with the highest effect sizes after restriction night R5.

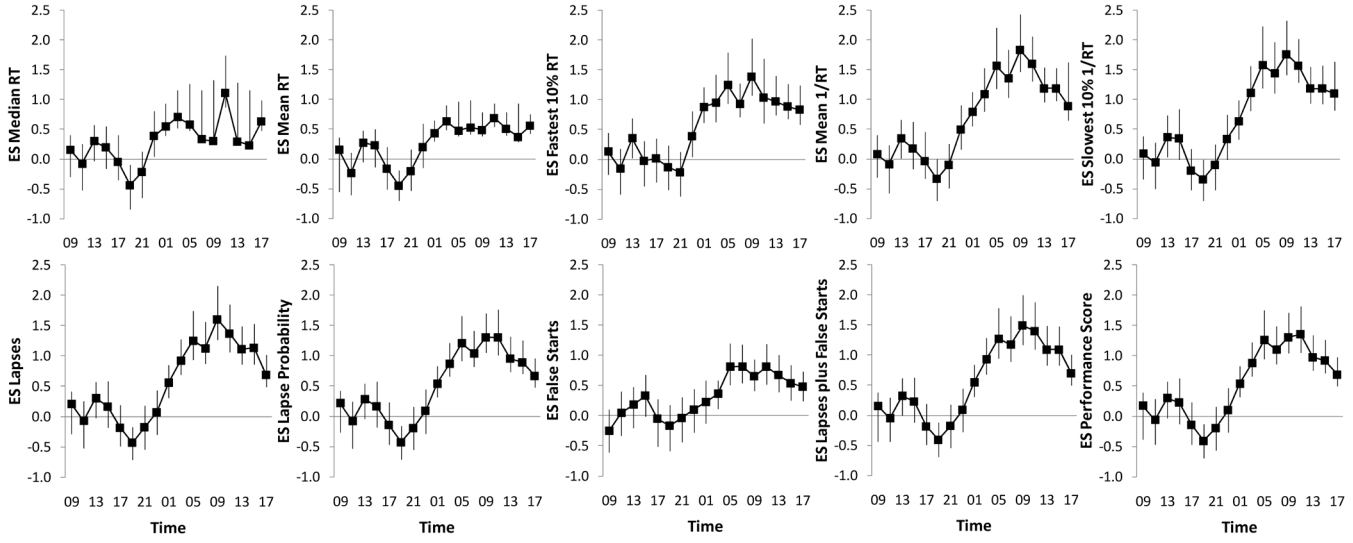

Figure 3—Effect sizes (ES) and 95% nonparametric bootstrap confidence intervals are shown for each of the 10 outcome metrics of the 10-min PVT and for each of the 17 tests performed during a 33-h period of total sleep deprivation (TSD). The average of the first 7 tests performed between 1 to 13 hours awake (corresponding to 09:00 and 21:00) served as the non–sleep deprived state. Effect sizes of Mean 1/RT, Slowest 10% 1/RT, and the Performance Score were multiplied by -1 to facilitate comparisons.

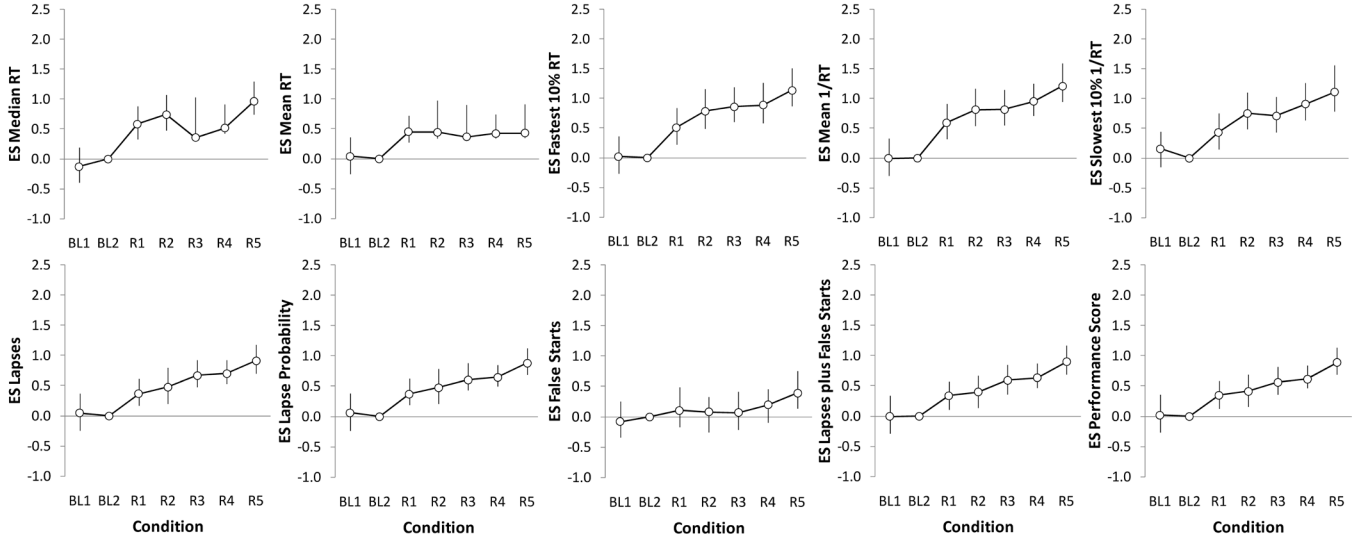

Figure 4—Effect sizes (ES) and 95% nonparametric bootstrap confidence intervals are shown for each of the 10 outcome metrics of the 10-min PVT and for each of the 7 conditions of the partial sleep deprivation (PSD) study (BL = baseline, R = restriction). Baseline day 2 (BL2) served as the non–sleep deprived state. Effect sizes of Mean 1/RT, Slowest 10% 1/RT, and the Performance Score were multiplied by -1 to facilitate comparisons.

Retrospective power calculations for the outcome metrics mean 1/RT and number of lapses for the 10-min PVT and for both total and partial sleep deprivation are shown in Table 4. These can be used by other researchers for prospective power calculations.

DISCUSSION

To our knowledge, this is the first study systematically investigating the effects of PVT duration and performance outcome metrics on the power to discriminate sleep deprived and alert subjects in both total and chronic partial sleep deprivation experiments. With fixed sample size and type I error rate α, power only depends on effect size, which differed considerably among PVT outcome metrics. High effect sizes were observed for the reciprocal metrics mean 1/RT and slowest 10% 1/RT in both the total and the partial sleep loss experiments. This can most likely be attributed to the superior statistical properties of these metrics. The reciprocal transform emphasizes even small changes in the optimal and intermediate response domain (fast and very fast RTs). At the same time, it de-emphasizes the influence of long lapses, which still markedly affect these outcome metrics without operating as extreme values. In contrast, median RT and especially mean RT performance are very prone to being affected by single extreme values, which diminishes their ability to detect differences between sleep deprived and alert states. Further evidence stems from the highly asymmetrical bootstrap

Table 4—Retrospective power calculation for the outcomes mean 1/RT and number of lapses for differences between sleep deprived (SDP) and non–sleep deprived (NSDP) states in total and partial sleep deprivation for the 10-min PVT

	Mean 1/RT				Number of Lapses			
	Difference SDP - NSDP [1/s]	Standard Deviation SDP - NSDP [1/s]	Effect Size[1]	Required Sample Size[2] [N]	Difference SDP - NSDP [N]	Standard Deviation SDP - NSDP [N]	Effect Size	Required Sample Size[2] [N]
Total Sleep Deprivation								
23 h TSD[3]	-0.627 (-0.775; -0.475)	0.434 (0.331; 0.516)	1.44 (1.03; 2.13)	6 (4; 12)	10.73 (7.79; 13.70)	8.54 (6.45; 10.20)	1.26 (0.90; 1.80)	8 (5; 12)
33 h TSD[4]	-0.659 (-0.779; -0.543)	0.341 (0.250; 0.414)	1.93 (1.55; 2.65)	5 (4; 6)	11.07 (9.00; 13.11)	5.94 (4.58; 6.99)	1.86 (1.45; 2.59)	5 (4; 6)
Partial Sleep Deprivation								
1 Night of 4 h TIB[5]	-0.125 (-0.188; -0.063)	0.212 (0.162; 0.254)	0.59 (0.32; 0.92)	25 (12; 81)	1.21 (0.36; 2.29)	3.30 (1.65; 4.93)	0.37 (0.18; 0.62)	61[6] (23; 262)
2 Nights of 4 h TIB[5]	-0.248 (-0.339; -0.158)	0.306 (0.239; 0.363)	0.81 (0.54; 1.17)	14 (8; 29)	2.38 (0.93; 3.90)	5.01 (3.42; 6.35)	0.47 (0.20; 0.79)	37[6] (15; 205)
3 Nights of 4 h TIB[5]	-0.314 (-0.430; -0.202)	0.385 (0.304; 0.452)	0.82 (0.55; 1.15)	14 (9; 28)	3.47 (2.02; 5.09)	5.18 (3.52; 6.78)	0.67 (0.48; 0.92)	20[6] (12; 37)
4 Nights of 4 h TIB[5]	-0.440 (-0.580; -0.306)	0.464 (0.363; 0.544)	0.95 (0.71; 1.25)	11 (8; 18)	5.57 (3.36; 8.05)	7.95 (5.22; 10.15)	0.70 (0.53; 0.92)	19[6] (12; 30)
5 Nights of 4 h TIB[5]	-0.545 (-0.680; -0.413)	0.452 (0.373; 0.515)	1.21 (0.94; 1.59)	8 (6; 11)	7.82 (5.36; 10.43)	8.58 (6.82; 9.90)	0.91 (0.71; 1.18)	12 (8; 18)

Nonparametric 95% bootstrap confidence intervals are given in parenthesis. [1]These values were multiplied by -1 to facilitate comparisons between effect sizes of both outcome metrics. [2]Sample size was calculated with Proc Power (SAS, Version 9.2) for a two-sided one-sample *t*-test with α set to 0.05 and 1-ß set to 0.8. For the calculation of sample size confidence intervals, corresponding values for Difference and Standard Deviation of SDP-NSDP were extracted from the 2.5th and 97.5th percentile of the empirical bootstrap distribution of the effect size. Sample sizes were rounded to the next highest integer. [3]Test bouts performed at 15-23 h awake represented the SDP state and those performed at ≤ 13 h awake represented the NSDP state. [4]Test bouts performed at 15-33 h awake represented the SDP state and those performed at ≤ 13 h awake represented the NSDP state. [5]Daytime PVT performance after the respective restriction night(s) was contrasted to daytime PVT performance after baseline night 2. [6]Visual inspection and statistical tests indicated that the distribution of this sample was non-normal. RT, response time; TSD, total sleep deprivation.

confidence intervals for median and mean RT presented in Figures 2, 3, and 4. They suggest there is a high probability that the population effect sizes for median and mean RT are higher than our study specific point estimates. However, effect sizes would still rank low relative to the other outcomes, which we demonstrated with sensitivity analyses log-transforming the data (which changes the measurement scale from ratio to ordinal) or excluding the extreme values. We intentionally avoid the expression "outliers" here, as the observations that were excluded in the sensitivity analysis likely represent two subjects very vulnerable to the effects of sleep deprivation that by no means should be excluded from the analysis only because of the statistical properties of the outcome variable.

The number of PVT performance lapses also scored high effect sizes both during total and chronic partial sleep loss. Although the number of false starts was not a very sensitive outcome to partial sleep deprivation, the combination of number of lapses and number of false starts was among the more discriminating measures, and scored the highest effects size in TSD. Taking false starts into account may also help to identify noncompliant subjects and those who try to prevent lapses (i.e., longer RTs as errors of omission) by biasing toward false starts (i.e., premature responses as errors of commission). Although the PVT performance score performed somewhat worse compared to both the number of lapses and the number of lapses and false starts in the TSD study, it still scored high effect sizes. The appeal of the performance score is its easy interpretability, with 100% indicating perfect performance and 0% indicating worst possible performance. The performance score also takes false starts into account.

PVT performance is affected by many factors other than the characteristics of the subject performing the test.[52] These include, among others, hardware, programming of the PVT, and definitions for calculating PVT outcome metrics. In this respect, the PVT is not very well standardized, which complicates comparisons between different studies and systematic meta-analyses. In Table 3, we list the criteria we used to program the 10-min PVT and to calculate outcome metrics. We encourage other researchers to adopt these criteria in future research to increase standardization of the 10-min PVT.

Based on the findings of these analyses, we suggest that mean 1/RT (i.e., response speed) and the number of lapses serve as PVT primary outcomes. Due to its superior statistical properties and robustness to extreme values, mean 1/RT scored the highest effect sizes (rank #2 in TSD and rank #1 in PSD), indicating its peak sensitivity to sleep loss among PVT measures. Lapses are an appropriate primary outcome as well, because they (a) reflect state instability[5,6,15]; (b) have high ecological valid-

ity that relates to risk in attention-demanding, real-world tasks (e.g., driving); and (c) are the most common outcome metric in the peer-reviewed literature (see Table 1). However, PVT users should still explore ways to improve sensitivity and specificity of the PVT, especially in experimental designs or interventions that differ from the ones investigated here, very much like we did in this manuscript (i.e., the standards proposed here are not intended to restrict researchers in any way).

Another important result of the study is that it evaluated the growing practice of using increasingly briefer versions of the PVT (i.e., PVT durations < 10 min). We found that with the exception of performance metrics that relied on lapse frequency in TSD, many of the PVT performance metrics reached maximum effect sizes before the full 10 min on the test, or showed saturating profiles. Importantly, this observation was not restricted to the sensitivity of the outcome metrics, but it applied to the ability of these metrics to discriminate sleep deprived from alert subjects in the two experiments. We believe this reveals the influence of time on task on size and variability of within-subject differences (between alert and sleep deprived states); and for some outcome metrics their ratio (i.e., effect size) seems to be optimal for shorter than 10-min PVT durations. This finding also demonstrates the potential feasibility of implementing briefer versions of the PVT in clinical and operational contexts. The fact that the maximum effect size was observed in TSD with the slowest 10% 1/RT using data of the first 3 minutes of the 10-min PVT suggests that versions shorter than the 5-min version that is already in use could probably be applied successfully.[48-51] This would likely increase practicality and user acceptance of the PVT compared to the current standard 10-min PVT.

Limitations

Several limitations have to be taken into account when interpreting the results of these analyses. First, instead of separate tests of different duration, we analyzed portions of the same 10-min PVT. However, knowledge of test duration may affect test performance. It is, for example, likely that subjects ration their effort over the test period, and that they would therefore perform differently if they knew a test is going to last for, e.g., only 3 minutes. The results on PVT sensitivity and test duration are therefore preliminary—that is, they have to be confirmed by comparisons of shorter versions of the PVT with the standard 10-min PVT, preferably in a controlled crossover design. For the same reason, our suggestions regarding outcome metrics primarily relate to the 10-min PVT. Second, our results may be somewhat unique to the specific PVT hardware and software used in our experiments, which had an exceptionally high measurement accuracy. Different hard- and software with different stimulus presentation and response characteristics may yield different results. Unless there is a way to effectively calibrate different systems, our analysis would need to be repeated for other setups, although certainly the basic findings may still be valid for other PVT platforms (e.g., the restricted value of median and mean RT outcome metrics). Third, we did not investigate and the findings may therefore not extend to PVT versions longer than 10 minutes. The same is true for other modalities (e.g., auditory stimuli). Fourth, it is unclear whether the results are valid for experimental designs or interventions other than the two investigated here (e.g., longer than 33 h of TSD or chronic PSD with a different number of restricted nights and/or with different amounts of sleep per night). Fifth, the findings in a strict sense only apply to the paired *t*-test investigated here, not to the independent sample *t*-test, nonparametric tests (which should be used instead of the paired *t*-test in small samples that are not normally distributed) or regression models. However, it is likely that, because of the basic statistical properties of the different outcome metrics, some of the findings extend to other forms of statistical analysis. Also, paired *t*-tests are very common in sleep research in general and sleep deprivation research specifically, and they are often also performed post hoc in ANOVA contexts. Therefore, we believe that our results will be relevant for a wide body of sleep research using the PVT. On the same note, we acknowledge that there are other ways to compare different PVT outcome metrics that may be more appropriate from a statistical point of view (e.g., the nonparametric bootstrap approach mentioned in Balkin et al.),[34] but less appropriate from an applied point of view (i.e., these forms of analysis are not frequently used). Sixth, this paper deals specifically with the PVT that, according to a comparative study by Balkin et al.,[34] "was among the most sensitive to sleep restriction, was among the most reliable with no evidence of learning over repeated administrations, and possesses characteristics that make it among the most practical for use in the operational environment." However, we acknowledge that there are other neurobehavioral tests that, in specific contexts, may outperform the PVT or deliver additional relevant information. Naturally, we were not able to assess these tests in this study. Finally, the investigated subjects were healthy and had a restricted age range. The results may therefore not generalize to non-healthy, older, or younger groups of subjects.

CONCLUSIONS

This study investigated the effect of PVT task duration and performance outcome metric on the power of the paired *t*-test to discriminate sleep deprived and alert subjects. Due to its superior statistical properties, the reciprocal metric mean 1/RT (response speed) scored very high effect sizes and should therefore preferably be used as a primary outcome metric. The number of lapses also scored high effect sizes that in total sleep deprivation were further improved by including the number of false starts in the outcome metric. Because of the high operational validity of lapses, they should also be considered a primary outcome metric. Our results suggest that using these PVT performance metrics increases the likelihood of finding differences between sleep deprived and alert states with smaller sample sizes. In contrast, the widespread practice of using PVT mean RT and median RT performance metrics should be avoided, as these outcomes were prone to bias from extreme observations and, even without extreme observations, scored only moderate effect sizes.

Analyses facilitating only portions of the full 10-min of the PVT showed that, for some outcomes, high effect sizes were achieved with PVT task durations considerably shorter than 10 min. Shorter PVT versions would increase acceptance of the test in settings where the 10-min PVT is considered impractical. However, these results need to be confirmed with direct comparisons of separate versions of the PVT with different test durations.

ACKNOWLEDGMENTS

This investigation was sponsored by the Human Factors Program of the Transportation Security Laboratory, Science and Technology Directorate, U.S. Department of Homeland Security (FAA #04-G-010), by NIH grants R01 NR004281 and UL1 RR024134, and in part by the National Space Biomedical Research Institute through NASA NCC 9-58. We thank the subjects participating in the experiments and the faculty and staff who helped acquire the data.

DISCLOSURE STATEMENT

This was not an industry supported study. Dr. Basner is Associate Editor of *SLEEP*. Dr. Basner has consulted for Purdue University for the FAA PARTNER Center of Excellence Project 25B. Dr. Basner has made paid presentations to the World Health Organization (WHO, German office). Dr. Dinges has received compensation for consulting to Eli Lilly and for serving on a scientific advisory council for Mars, Inc. He has received research support from Merck & Co., and he is compensated by the APSS for serving as Editor-in-Chief of *SLEEP*. He recuses himself from all decisions related to manuscripts on which he has a conflict of interest.

REFERENCES

1. Banks S, Dinges DF. Behavioral and physiological consequences of sleep restriction. J Clin Sleep Med 2007;3:519-28.
2. Goel N, Rao H, Durmer JS, Dinges DF. Neurocognitive consequences of sleep deprivation. Semin Neurol 2009;29:320-39.
3. Van Dongen HP, Vitellaro KM, Dinges DF. Individual differences in adult human sleep and wakefulness: Leitmotif for a research agenda. Sleep 2005;28:479-96.
4. Lim J, Dinges DF. A meta-analysis of the impact of short-term sleep deprivation on cognitive variables. Psychol Bull 2010;136:375-89.
5. Lim J, Dinges DF. Sleep deprivation and vigilant attention. Molecular and biophysical mechanisms of arousal, alertness, and attention. Ann New York Acad Sci 2008:305-22.
6. Dorrian J, Rogers NL, Dinges DF, Kushida CA. Psychomotor vigilance performance: Neurocognitive assay sensitive to sleep loss. Sleep deprivation: clinical issues, pharmacology and sleep loss effects. New York, NY: Marcel Dekker, Inc., 2005:39-70.
7. Cordova CA, Said BO, McCarley RW, Baxter MG, Chiba AA, Strecker RE. Sleep deprivation in rats produces attentional impairments on a 5-choice serial reaction time task. Sleep 2006;29:69-76.
8. Lim J, Tan JC, Parimal S, Dinges DF, Chee MW. Sleep deprivation impairs object-selective attention: a view from the ventral visual cortex. PLoS One 2010;5:e9087.
9. Posner MI. Measuring alertness. Ann N Y Acad Sci 2008;1129:193-9.
10. Trujillo LT, Kornguth S, Schnyer DM. An ERP examination of the different effects of sleep deprivation on exogenously cued and endogenously cued attention. Sleep 2009;32:1285-97.
11. Anderson C, Horne JA. Sleepiness enhances distraction during a monotonous task. Sleep 2006;29:573-6.
12. Tucker AM, Whitney P, Belenky G, Hinson JM, Van Dongen HPA. Effects of sleep deprivation on dissociated components of executive functioning. Sleep 2010;33:47-57.
13. Dinges DF, Powell JW. Microcomputer analysis of performance on a portable, simple visual RT task during sustained operations. Behav Res Methods Instrum Comput 1985;6:652-5.
14. Dinges DF, Pack F, Williams K, et al. Cumulative sleepiness, mood disturbance, and psychomotor vigilance performance decrements during a week of sleep restricted to 4-5 hours per night. Sleep 1997;20:267-77.
15. Doran SM, Van Dongen HP, Dinges DF. Sustained attention performance during sleep deprivation: Evidence of state instability. Archives Italiennes de Biologie: A Journal of Neuroscience 2001;139:1-15.
16. Dinges DF, Kribbs NB. Performing while sleepy: Effects of experimentally-induced sleepiness. In: Monk TH, ed. Sleep, sleepiness and performance. Chichester, United Kingdom: John Wiley and Sons, Ltd., 1991:97-128.
17. Warm JS, Parasuraman R, Matthews G. Vigilance requires hard mental work and is stressful. Hum Factors 2008;50:433-41.
18. Patrick GTW, Gilbert JA. On the effects of sleep loss. Psychol Rev 1896;3:469-83.
19. Dinges DF. Probing the limits of funcional capability: The effects of sleep loss on short-duration task. In: Broughton RJ, Ogilvie RD, eds. Sleep, arousal, and performance. Boston, MA: Birkhäuser, 1992:177-88.
20. Lim J, Wu WC, Wang J, Detre JA, Dinges DF, Rao H. Imaging brain fatigue from sustained mental workload: an ASL perfusion study of the time-on-task effect. Neuroimage 2010;49:3426-35.
21. Davies DR, Parasuraman R. The psychology of vigilance. New York, NY: Academic Press, 1982.
22. Philip P, Akerstedt T. Transport and industrial safety, how are they affected by sleepiness and sleep restriction? Sleep Med Rev 2006;10:347-56.
23. Dinges DF. An overview of sleepiness and accidents. J Sleep Res 1995;4:4-14.
24. Van Dongen HP, Dinges DF. Sleep, circadian rhythms, and psychomotor vigilance. Clin Sports Med 2005;24:237-49, vii-viii.
25. Gunzelmann G, Moore LR, Gluck KA, Van Dongen HP, Dinges DF. Individual differences in sustained vigilant attention: Insights from computational cognitive modeling. In: Love BC, McRae K, Sloutsky VM, eds. 30th Annual Meeting of the Cognitive Science Society; 2008; Austin, TX: Cognitive Science Society, 2008: 2017-22.
26. Dinges DF, Mallis M. Managing fatigue by drowsiness detection: can technological promises be realized? In: Hartley L, ed.; Managing fatigue in transportation. Pergamon, 1998: 209-29.
27. Van Dongen HP, Maislin G, Mullington JM, Dinges DF. The cumulative cost of additional wakefulness: dose-response effects on neurobehavioral functions and sleep physiology from chronic sleep restriction and total sleep deprivation. Sleep 2003;26:117-26.
28. Gunzelmann G, Moore LR, Gluck KA, Van Dongen HP, Dinges DF. Fatigue in sustained attention: Generalizing mechanisms for time awake to time on task. In: Ackerman PL, ed. Cognitive fatigue: Multidisciplinary perspectives on current research and future applications. Washington, D.C.: American Psychological Association, 2010:83-101.
29. Chee MW, Tan JC, Zheng H, et al. Lapsing during sleep deprivation is associated with distributed changes in brain activation. J Neurosci 2008;28:5519-28.
30. Drummond SP, Bischoff-Grethe A, Dinges DF, Ayalon L, Mednick SC, Meloy MJ. The neural basis of the psychomotor vigilance task. Sleep 2005;28:1059-68.
31. Tomasi D, Wang RL, Telang F, et al. Impairment of attentional networks after 1 night of sleep deprivation. Cereb Cortex 2009;19:233-40.
32. Jewett ME, Dijk DJ, Kronauer RE, Dinges DF. Dose-response relationship between sleep duration and human psychomotor vigilance and subjective alertness. Sleep 1999;22:171-9.
33. Belenky G, Wesensten NJ, Thorne DR, et al. Patterns of performance degradation and restoration during sleep restriction and subsequent recovery: a sleep dose-response study. J Sleep Res 2003;12:1-12.
34. Balkin TJ, Bliese PD, Belenky G, et al. Comparative utility of instruments for monitoring sleepiness-related performance decrements in the operational environment. J Sleep Res 2004;13:219-27.
35. Mollicone DJ, van Dongen HPA, Rogers NL, Dinges DF. Response surface mapping of neurobehavioral performance: Testing the feasibility of split sleep schedules for space operations. Acta Astronaut 2008;63:833-40.
36. Wyatt JK, Ritz-De Cecco A, Czeisler CA, Dijk DJ. Circadian temperature and melatonin rhythms, sleep, and neurobehavioral function in humans living on a 20-h day. Am J Physiol 1999;277:R1152-63.
37. Graw P, Krauchi K, Knoblauch V, Wirz-Justice A, Cajochen C. Circadian and wake-dependent modulation of fastest and slowest reaction times during the psychomotor vigilance task. Physiol Behav 2004;80:695-701.
38. Van Dongen HP, Baynard MD, Maislin G, Dinges DF. Systematic interindividual differences in neurobehavioral impairment from sleep loss: evidence of trait-like differential vulnerability. Sleep 2004;27:423-33.
39. Killgore WD, Grugle NL, Reichardt RM, Killgore DB, Balkin TJ. Executive functions and the ability to sustain vigilance during sleep loss. Aviat Space Environ Med 2009;80:81-7.

40. Goel N, Banks S, Mignot E, Dinges DF. PER3 polymorphism predicts cumulative sleep homeostatic but not neurobehavioral changes to chronic partial sleep deprivation. PLoS One 2009;4:e5874.
41. Neri DF, Oyung RL, Colletti LM, Mallis MM, Tam PY, Dinges DF. Controlled breaks as a fatigue countermeasure on the flight deck. Aviat Space Environ Med 2002;73:654-64.
42. Dinges DF, Orne MT, Whitehouse WG, Orne EC. Temporal placement of a nap for alertness: contributions of circadian phase and prior wakefulness. Sleep 1987;10:313-29.
43. Van Dongen HP, Price NJ, Mullington JM, Szuba MP, Kapoor SC, Dinges DF. Caffeine eliminates psychomotor vigilance deficits from sleep inertia. Sleep 2001;24:813-9.
44. Czeisler CA, Walsh JK, Roth T, et al. Modafinil for excessive sleepiness associated with shift-work sleep disorder. N Engl J Med 2005;353:476-86.
45. Rupp TL, Wesensten NJ, Bliese PD, Balkin TJ. Banking sleep: realization of benefits during subsequent sleep restriction and recovery. Sleep 2009;32:311-21.
46. Banks S, Van Dongen HP, Maislin G, Dinges DF. Neurobehavioral dynamics following chronic sleep restriction: Dose-response effects of one night of recovery. Sleep 2010;33:1013-26.
47. Kribbs NB, Pack AI, Kline LR, et al. Effects of one night without nasal CPAP treatment on sleep and sleepiness in patients with obstructive sleep apnea. Am Rev Respir Dis 1993;147:1162-8.
48. Loh S, Lamond N, Dorrian J, Roach G, Dawson D. The validity of psychomotor vigilance tasks of less than 10-minute duration. Behav Res Methods Instrum Comput 2004;36:339-46.
49. Lamond N, Jay SM, Dorrian J, Ferguson SA, Roach GD, Dawson D. The sensitivity of a palm-based psychomotor vigilance task to severe sleep loss. Behav Res Methods 2008;40:347-52.
50. Roach GD, Dawson D, Lamond N. Can a shorter psychomotor vigilance task be used as a reasonable substitute for the ten-minute psychomotor vigilance task? Chronobiol Int 2006;23:1379-87.
51. Lamond N, Dawson D, Roach GD. Fatigue assessment in the field: validation of a hand-held electronic psychomotor vigilance task. Aviat Space Environ Med 2005;76:486-9.
52. Thorne DR, Johnson DE, Redmond DP, Sing HC, Belenky G, Shapiro JM. The Walter Reed palm-held psychomotor vigilance test. Behav Res Methods 2005;37:111-8.
53. Efron B, Tibshirani RJ. An introduction to the bootstrap. first ed. New York, NY: Chapman & Hall, 1993.
54. Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum, 1988.