Published in final edited form as:
Acta Astronaut. 2011 December 1; 69(11-12): 949–959. doi:10.1016/j.actaastro.2011.07.015.

Validity and Sensitivity of a Brief Psychomotor Vigilance Test (PVT-B) to Total and Partial Sleep Deprivation

Mathias Basner, MD, PhD, MSc[1], **Daniel Mollicone, PhD**[1,2], and **David F. Dinges, PhD**[1]

[1]Unit of Experimental Psychiatry, Division of Sleep and Chronobiology, Department of Psychiatry, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania

[2]Pulsar Informatics, Inc., Philadelphia, Pennsylvania

Abstract

The Psychomotor Vigilance Test (PVT) objectively assesses fatigue-related changes in alertness associated with sleep loss, extended wakefulness, circadian misalignment, and time on task. The standard 10-min PVT is often considered impractical in applied contexts. To address this limitation, we developed a modified brief 3-min version of the PVT (PVT-B). The PVT-B was validated in controlled laboratory studies with 74 healthy subjects (34 female, aged 22–45 years) that participated either in a total sleep deprivation (TSD) study involving 33 hours awake (N=31 subjects) or in a partial sleep deprivation (PSD) protocol involving 5 consecutive nights of 4 hours time in bed (N=43 subjects). PVT and PVT-B were performed regularly during wakefulness. Effect sizes of 5 key PVT outcomes were larger for TSD than PSD and larger for PVT than for PVT-B for all outcomes. Effect size was largest for response speed (reciprocal response time) for both the PVT-B and the PVT in both TSD and PSD. According to Cohen's criteria, effect sizes for the PVT-B were still large (TSD) or medium to large (PSD, except for fastest 10% RT). Compared to the 70% decrease in test duration the 22.7% (range 6.9%-67.8%) average decrease in effect size was deemed an acceptable trade-off between duration and sensitivity. Overall, PVT-B performance had faster response times, more false starts and fewer lapses than PVT performance (all $p<0.01$). After reducing the lapse threshold from 500 ms to 355 ms for PVT-B, mixed model ANOVAs indicated no differential sensitivity to sleep loss between PVT-B and PVT for all outcome variables (all $P>0.15$) but the fastest 10% response times during PSD ($P<0.001$), and effect sizes increased from 1.38 to 1.49 (TSD) and 0.65 to 0.76 (PSD), respectively. In conclusion, PVT-B tracked standard 10-min PVT performance throughout both TSD and PSD, and yielded medium to large effect sizes. PVT-B may be a useful tool for assessing behavioral alertness in settings where the duration of the 10-min PVT is considered impractical, although further validation in applied settings is needed.

Keywords

PVT; psychomotor vigilance; alertness; sleep deprivation; sleep restriction; attention; lapse; response time

Correspondence address: 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, USA, phone (215) 573-5866, fax (215) 573-6410, basner@mail.med.upenn.edu.

1. INTRODUCTION

Undisturbed sleep of sufficient length on a regular basis is of paramount importance for recuperation and the maintenance of behavioral alertness and cognitive performance [1, 2]. Nevertheless, large parts of the population engage in acute or chronic partial sleep loss, suggesting that sleep is perceived as a flexible commodity that can be exchanged for waking activities considered more essential or of greater value [3]. In a recent analysis of time use in the US [4], work time was the waking activity most strongly reciprocally related to sleep time. At the same time the prevalence of shift work, requiring employees to both work and sleep at adverse times relative to their circadian phase, has increased over the past years [5]. Therefore, sleep disorders, lifestyle and work related curtailments of sleep, and working during unfavorable circadian times all may reduce neurobehavioral alertness to levels that increase the risk of errors and accidents [6, 7]. Prevention of these outcomes through detection of fatigue (i.e., loss of alertness, sleepiness) remains a high priority in many safety-sensitive areas of human activity, and is also crucial for mission success in space flight.

Objective and quantitative assessments are necessary to evaluate the presence of fatigue-related deficits and to develop strategies for fatigue mitigation, especially as self-reports of sleepiness and self-assessments of performance capability have been shown to be unreliable [8, 9]. In this context, neurobehavioral tests for fatigue assessment not only need to be operationally and conceptually valid, reliable, sensitive, specific, generalizable, and easy to use [10, 11], but also brief enough to be acceptable for the target population and to allow for repeated administration in operational environments.

Many performance tests have been developed to objectively assess the degree of cognitive performance deterioration related to sleep loss. Among these, the Psychomotor Vigilance Test (PVT) is widely used [12, 13]. It is based on simple reaction time (RT) to stimuli that occur at random intervals and therefore measures vigilant attention [14]. Auditory and visual reaction time tests have been used since the late 19th century in sleep research [15], but the PVT in its current version (i.e., 10-min duration with random inter-stimulus intervals (ISI) between 2 and 10 sec) was proposed by Dinges and Powell in 1985 [16]. When appropriate PVT outcomes are used with precision timing of RT, the standard 10-min PVT has proven to be very sensitive to the dynamics of acute total sleep deprivation (TSD) and chronic partial sleep deprivation (PSD). [12]

Sleep deprivation causes both an overall slowing of PVT response times and an increase in the number of PVT errors of omission (i.e. lapses, usually defined as RTs ≥ 500 ms), as well as a smaller increase in errors of commission (responses without a stimulus) [14, 17]. These effects increase with time on task [18]. An advantage the PVT has over nearly all other cognitive tests is that it is virtually unaffected by either aptitude (inter-individual variability) or learning (intra-subject variability)—that is, PVT performance does not improve as a function of repeated administration [19]. The test has high reliability, with intra-class correlations measuring test-retest reliability above 0.8 [13].

The 10-min PVT has been shown to be a valid tool for assessing behavioral alertness and vigilant attention performance in a large number of experimental, clinical, and operational paradigms. In addition to being sensitive to both TSD [17, 20] and PSD [21, 22], the PVT has demonstrated sensitivity to other perturbations of sleep homeostatic and circadian drives; [23, 24] to inter- and intra-subject variability in the response to sleep loss; [9] to the effects of jet lag and shift work; [25] and to improvements in alertness following initiation of CPAP treatment in obstructive sleep apnea (OSA) patients; [26] administration of wake-promoting drugs; [27, 28] and following naps. [29] Balkin et al. [30] assessed the utility of a

variety of instruments for monitoring sleepiness-related performance decrements and concluded that the PVT "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."

The standard 10-min PVT with 2–10s ISI is most commonly used, although both longer [18, 31] and shorter [32] duration versions have been evaluated. Test duration is an important aspect of the PVT because even severely sleep deprived subjects may be able to perform normally for a short time by increasing compensatory effort. However, in a systematic analysis of PVT duration, we showed that the ability of the PVT to differentiate alert and sleepy subjects was, depending on the outcome variable, only marginally lower (and at times higher) for shorter than 10-min test durations [12]. Therefore, optimal PVT duration may be shorter than 10-min for some outcome variables, demonstrating feasibility of shorter versions of the PVT. Accordingly, a 5-min handheld version of the PVT already exists [32, 33, 34, 35, 36]. However, both 2-min [32] and 90 s [34] versions of the PVT were deemed to be too insensitive to be used as valid tools for the detection of neurobehavioral effects of fatigue, leaving open the question of whether a brief PVT that was sensitive to sleep loss could be developed.

We therefore set out to develop a brief PVT (PVT-B) that was as sensitive to TSD and PSD as the standard 10-min PVT. Based on our theory of how sleepiness manifests in performance, our large PVT databases, knowledge on the importance of outcome variable [12], ISI, and precision of timing for the ability of the PVT to differentiate sleep deprived and alert subjects, we shortened test duration from 10 min to 3 min and ISI from the standard 2 – 10 s to 1 – 4 s to create the PVT-B, while maintaining sufficient response sampling rates to detect wake state instability. [14] We hypothesized that PVT-B would retain its sensitivity and specificity to sleep loss, and therefore be a practical tool for fatigue assessment. A sensitive, specific, brief PVT-B would meet the criteria for fitness-for-duty testing not only prior to the start of a shift but also during repeated administrations while on the shift.

2. MATERIAL AND METHODS

2.1 Subjects and protocol

This investigation used data from a TSD and from a PSD protocol. The 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 [37]. This analysis is based on data gathered in a pilot study on N=12 subjects and in the main study on N=24 subjects. Four subjects were excluded from the analysis due to non-compliance or excessive fatigue during the first 16 hours of wakefulness. Another subject withdrew after 26 h awake. Therefore, a subset of N=31 subjects (mean age ± standard deviation = 31.1 ± 7.3 yr, 18 female) contributed to the analyses. Study participants stayed in the research lab for five consecutive days, which included a 33 h period of TSD. The study started at 8 am on day 1 and ended at 8 am on day 5. During 1 of every 2 hours awake, subjects performed a 30-min computerized neurobehavioral test battery (NTB) that included a 10-min PVT, followed by an SLST. The PVT-B was performed after the NTB and immediately prior to the SLST in 23 subjects. In an effort to investigate possible order effects of PVT administration, the PVT-B was administered immediately before the NTB in the remaining N=8 subjects. A 33 h period of total sleep deprivation started either on day 2 (N=22) or on day 3 (N=9) of the study (the latter condition was added to the final protocol due to a time-in-study effect in SLST performance that was found in the pilot study) [37]. Except for the sleep deprivation period,

subjects had 8 h sleep opportunities between 12 pm and 8 am. The first sleep period was monitored polysomnographically to exclude possible sleep disorders.

In the partial sleep deprivation study, a total of 159 healthy adults completed a 12-day laboratory protocol. A detailed description of the study is published elsewhere [38]. The PVT-B was added later to the protocol, and thus only a subset of 47 subjects performed both the PVT and the PVT-B. Three subjects were excluded from the analysis due to non-compliance or excessive fatigue during baseline data collection. One additional subject had no valid baseline data. Therefore, this analysis is based on N=43 subjects (mean age ± SD = 30.5 ± 7.3 years, 16 female) that were studied in small groups for 12 consecutive days. They had two initial baseline nights (BL1, BL2) of 10 h TIB per night (10 pm–8 am), followed by 5 nights (R1-R5) of sleep restricted to 4 h time in bed per night (4 am–8 am). The remaining 5 nights of the study involved other conditions not reported here. Subjects were continuously behaviorally monitored by trained staff to ensure adherence to the experimental protocol. They wore a wrist actigraph throughout the 12-day laboratory protocol. On BL1, BL2, R1, and R5 they wore ambulatory EEG and ECG recording equipment throughout the day and night. During the days without EEG, subjects were given shower opportunities between 2:30 pm and 4 pm. Meals were provided at regular times throughout the protocol (8:30 am-10 am; 12:30 am-2 pm; 6:30 pm-8 pm). Subjects completed 30min bouts of the NTB which included a 10-min PVT every 2 h during scheduled wakefulness beginning at 8 am each day. The PVT-B always immediately followed the NTB, but only every 4 h during scheduled wakefulness.

In both studies, participants were investigated in the Sleep and Chronobiology Laboratory at the Hospital of the University of Pennsylvania. They 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. All subjects were compensated for their participation. They were free of acute or chronic medical and psychological conditions, as established by interviews, clinical history, questionnaires, physical exams, and blood and urine tests. Subjects were 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 study. Subjects were not permitted to use caffeine, alcohol, tobacco and medications (except oral contraceptives) in the week before the laboratory experiment, as verified by blood and urine screens. 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. The light levels in the laboratory were held constant at less than 50 lux during scheduled wakefulness and less than 1 lux during scheduled sleep periods. Ambient temperature was maintained between 22°–24° C.

2.2 PVT

In both studies, the 10-min PVT was performed on a personal computer and performed and analyzed according to the criteria reported in Basner and Dinges. [12] Subjects were instructed to monitor a red rectangular box and 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. The inter-stimulus intervals varied randomly from 2–10 s (including a 1 s RT feedback interval). The PVT-B was performed on the PVT-192 (Ambulatory Monitoring Inc., Ardsley, NY), a handheld device measuring 21 × 11 × 6 cm and weighing ca. 650 g. The visual RT stimulus and performance feedback were presented on the device's 2.5 × 1 cm four-digit LED display. The inter-stimulus intervals varied randomly from 1–4 s (including a 1 s RT feedback interval). For both versions of the PVT, subjects were instructed to press the response button as soon as each stimulus appeared, in order to keep RT as low as possible, but not to press the button too soon (which yielded a

false start warning on the display). Both versions gave a signal after a 30 s period without response, which was counted as a lapse (see below) with 30 s response time.

2.3 Outcome measures

Based on our previous systematic analysis of different PVT outcome measures and on the publication frequency of PVT outcome measures in the literature, [12] we chose to include the following five variables in our analyses: (1) mean 1/RT (also called reciprocal response time or response speed), (2) slowest 10% 1/RT, (3) number of lapses, (4) fastest 10% of RT, and (5) a newly developed performance score. A 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. For calculating mean 1/RT and slowest 10% 1/RT, each RT was divided by 1,000 and then reciprocally transformed. The transformed values were then averaged. Lapses (errors of omission) were defined as RTs ≥ 500 ms. Initial analyses showed that RTs were shorter and lapse probability was lower on the PVT-B compared to the PVT (see Figure 1). Hence, a 6^{th} variable was generated for the PVT-B comprising the number of lapses based on a lapse definition of ≥ 355 ms. This threshold was chosen because, under comparable conditions (TSD, time on task < 3 min, ISI between 2 s and 4 s), it raised PVT-B lapse frequency to levels observed in the 10-min PVT with the standard 500 ms lapse definition. Both the number of 500 ms and 355 ms lapses on the PVT-B were compared to the number of standard 500 ms lapses on the 10-min PVT. The performance score is calculated as 100% minus the number of lapses and false starts relative to the number of valid stimuli and false starts. It ranges from 100% (optimal performance, no lapses or false starts) to 0% (worst possible performance, only lapses and false starts). In this analysis, the PVT-B performance score was always calculated based on the 355 ms lapse threshold.

2.4 Data analysis and statistical procedures

The following paragraphs are in part reproduced from Basner and Dinges. [12] A pair of PVT-B and PVT test bouts was excluded from the analysis if either the PVT-B or the PVT test bout was missing or incomplete. This way, 24 PVT pairs out of a total of 903 (2.7%) were excluded from the analysis in the PSD protocol. The data in the TSD study were complete.

To compare the utility of the PVT and the PVT-B to differentiate sleep deprived from alert subjects, in the TSD study test bouts 1 to 7 (9 am to 9 pm) were averaged within subjects to reflect the non sleep deprived state and test bouts 8 to 17 (11 pm to 5 pm on the following 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 begins to decrease only after 16 h of wakefulness [21]. For the PSD study, daily averages of outcome variables were computed within subjects over the test bouts administered at 12 am, 4 pm, and 8 pm. The 8 am test bout was not used because of possible sleep inertia effects. Average performance on BL2 reflected the non sleep deprived state, while average performance on R5 reflected the sleep deprived state. Only test bouts that existed in both conditions (non-SD and SD) were used for averaging. For example, if the 4 pm test bout was missing for a subject in R5, the 4 pm test bout was not used for averaging in BL2, even if it existed. [12]

In earlier validation studies of shorter than 10-min versions of the PVT, the authors' main concern was that the shorter version of the PVT retained its *sensitivity* to sleep loss. However, a good test should both be sensitive (detect those with relevant degrees of cognitive impairment) and specific (indicate no relevant impairment in alert subjects) [10]. In our view, the ability or, in statistical terms, the power of the PVT to discriminate between

sleep deprived and alert subjects is a better criterion for the validation of shorter duration PVTs.

A paired t-test would be a valid method to investigate whether there is a statistically significant difference between non-SD and SD conditions. In the paired t-test, differences of outcome values between non-SD and SD conditions are calculated within subjects, and these differences are tested with a one sample t-test against zero. With a given type-1 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. A powerful test will indicate high degrees of cognitive impairment in sleep deprived subjects, low degrees of cognitive impairment in alert subjects, and it will do so consistently.

The one-sample t-test is the most powerful test available (i.e., it outperforms non-parametric 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 unitless effect size for PVT-B and PVT, for the 6 outcome metrics, and for the TSD and the PSD study. As a measure of effect size precision, we calculated 95% non-parametric bootstrap confidence intervals based on 1,000,000 samples according to Efron and Tibshirani [39]. 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 mean 1/RT, slowest 10% 1/RT, and the performance score were multiplied by -1 to facilitate comparisons between outcome metrics.

Graphs comparing the evolution of the different outcome metrics between the PVT and the PVT-B during 33 h of TSD and across the 7 days (BL1 to R5) of the PSD protocol were generated. Bias-corrected and accelerated 95% bootstrap confidence intervals based on 1,000,000 bootstrap samples were calculated for each estimate. [39] A random subject effect mixed model ANOVA (SAS Version 9.2, SAS Institute) with two within-subject factors (test version and test time) and their interaction was calculated for the five outcome variables and both SD protocols. Denominator degrees of freedom (DF) were computed using Satterthwaite's approximation. We were especially interested in whether test outcomes differed significantly between versions of the test, and if so, whether this difference constituted merely a parallel shift of outcome values across test bouts (no significant interaction) or whether there was evidence of differential sensitivity to sleep loss between tests (significant interaction).

Both graphs showing the original data and graphs showing outcome measures centered around alert performance (test bouts 1 to 7 for TSD and BL2 for PSD) were created.

Centering around baseline performance was intended to remove systematic differences between the two PVT versions due to differences in hardware, knowledge of test duration, or order of test administration without removing differences between tests due to differential sensitivity to sleep deprivation. We felt this approach was more valid compared to centering data around their overall mean or even standardizing the data (as in Lamond et al. [33, 35]), which reduces both within- and between-subject variability.

It was then tested with a paired t-test for each given time point during sleep deprivation, i.e. test bouts 8 to 15 (TSD) and R1 to R5 (PSD), whether PVT and PVT-B differed significantly from each other. In order to account for multiple testing we adjusted p-values according to the false discovery rate method, which limits the expected fraction of null hypotheses rejected mistakenly to a certain probability [40]. In the graphs adjusted two-sided significance levels were marked as follows: * $p<0.05$, ** $p<0.01$, and *** $p<0.001$.

Test duration, hardware, and ISIs were changed simultaneously in the PVT-B relative to the PVT. In an effort to disentangle the contributions of each of these factors, we performed a random subject effect mixed model ANOVA for the TSD data set with the following variables: test version (PVT vs. PVT-B), ISI (10 levels), time on task (10 levels), test order (PVT-B first vs. PVT first), hours awake (17 levels), and start day of the TSD period (day 2 vs. day 3). Overall, 80,438 valid response times contributed to the analysis (false starts were excluded). Response times were base e log-transformed after subtracting a constant value of 99 ms from each RT (and thus anchoring the minimum value at 1). We only used the TSD data set for this analysis as the order of tests was constant in the PSD protocol.

3. RESULTS

Although PVT-B test duration was shortened by 70%, the average number of sampled RTs decreased by only 32.4% (62.3 versus 93.6 stimuli, $p<0.001$) due to the simultaneously decreased ISIs in the PVT-B. Subjects were faster and the false start rate was significantly higher on the PVT-B compared to the PVT both during SD and while alert (see Figure 1). Also, lapse frequency on the PVT-B (5.3%) was significantly lower than on the PVT (9.6%, $p<0.0001$).

Adjusting for all other variables in the model, response times were significantly influenced by hours awake, ISI, time on task, and the version of the test (all $P<0.0001$) in the TSD study (Table 1). There was no significant influence of the order of the tests ($P=0.0661$) or sleep deprivation start day ($P=0.4286$). Post-hoc analyses (not shown) indicated that RTs decreased continuously with ISIs increasing from 1 s to 6 s and leveled off with ISIs 7 s or longer. RTs increased continuously with time on task over the 10 min period. Finally and corroborating the findings shown in Figure 1, RTs were significantly shorter on the PVT-B relative to the PVT.

The results of the effect size analyses are shown in Figure 2. As expected, effect sizes were lower for PSD compared to TSD. The highest effect size was observed for 1/RT in TSD for both the PVT and the PVT-B. For all outcome variables and during both TSD and PSD, effect sizes for the PVT-B were lower compared to the PVT. They decreased on average by 22.7%. The smallest decrease (6.9%) was observed for the performance score during TSD and the highest decrease (67.8%) for the fastest 10% RT during PSD. Changing lapse definition from 500 ms to 355 ms increased the effect size for lapsing on the PVT-B during both TSD (1.38 versus 1.49) and PSD (0.65 versus 0.76).

Figure 3 compares all 17 test bouts performed during 33 h of TSD for the 6 outcome variables between the PVT-B and the PVT. The untransformed data shown in the 6 graphs on the left of Figure 3 indicate that subjects were faster and produced fewer 500 ms lapses

on the PVT-B. This is corroborated by significant main effects for test version (all $p<0.01$, see results of mixed model ANOVAs in Table 2). Otherwise, performance on the PVT-B tracked PVT performance closely, and, except for the outcome *number of lapses* based on the standard lapse definition (500 ms), mixed model ANOVAs did not indicate differential sensitivity to TSD between tests. Lowering the lapse threshold from 500 ms to 355 ms increased the sensitivity of the PVT-B to sleep loss, attenuated the differences between the two test versions and decreased differential sensitivity between tests (the p-value for the interaction test version * test time increased from $P<0.0001$ to $P=0.1554$).

These results were confirmed by post-hoc paired t-tests based on outcome variables centered around alert PVT performance (six graphs on the right of Figure 3). These tests indicate significant differences between PVT-B and PVT performance only for the outcome variable *number of lapses* if the standard 500 m lapse definition was applied to both tests. Otherwise, no statistically significant differences were found for any time point between both versions of the test. However, descriptively subjects were slightly faster and exhibited fewer lapses (based on the 355 ms lapse definition) on the PVT-B compared to the PVT during TSD.

Figure 4 compares the seven conditions of the PSD protocol (BL1, BL2, and R1 to R5) for the 6 outcome variables between the PVT-B and the PVT. The untransformed data shown in the 6 graphs on the left of Figure 4 indicate that subjects were faster and produced fewer lapses on the PVT-B. This is corroborated by significant main effects for test version (all $p<0.0001$, see Table 2). However, the mixed model ANOVAs indicate differential sensitivity to PSD between tests for the outcomes *fastest 10% RT* and *number of lapses* based on the standard lapse definition (p-values for the interaction both <0.0001). Lowering the lapse threshold from 500 ms to 355 ms increased the sensitivity of the PVT-B to sleep loss and attenuated the differences between the two test versions (p-value interaction 0.3531). Otherwise, performance on the PVT-B tracked PVT performance closely. Corroborating the findings of the mixed model ANOVA, post-hoc paired t-tests based on outcome variables centered around BL2 PVT performance (six graphs on the right of Figure 4) found significant differences between PVT-B and PVT performance during partial sleep restriction for the fastest 10% RT and the number of lapses based on the 500 ms definition for PVT-B. Additionally, the performance score differed significantly after restriction night 2 (R2) between PVT-B and PVT. The main effect of time was found to be highly significant (all $p<0.0001$) for all outcome variables during both total and partial sleep deprivation.

4. DISCUSSION

In an earlier analysis, we systematically compared performance on the 10-min PVT to the first 1 to 9 min of the same test, and found that the highest effect sizes were often found for shorter than 10 min test durations, especially for outcome variables that did not involve lapses. [12] This underlined the feasibility of shorter than 10-min PVTs and motivated us to develop a modified 3-min version of the test. A 5-min version had already been shown to reach similar degrees of sensitivity to sleep deprivation as the standard 10-min PVT, albeit only in TSD paradigms [33, 34, 35]. Roach et al. [34] concluded that a 90 s version of the PVT may not provide a reasonable substitute for the 10-min PVT.

This is the first study to systematically compare a modified brief 3-min version of the PVT with the standard 10-min version during both TSD and PSD. For this purpose, 74 subjects contributed 1,656 pairs of both versions of the test that were performed in close temporal proximity. However, we did not simply shorten test duration. We also decreased ISIs from the standard 2 to 10 s to 1 to 4 s for the following reasons: First, we wanted to get more precise estimates of our outcome variables by lowering ISIs and therefore by sampling more behavior. Thus, although the duration of the test decreased by 70%, the sampling rate only

decreased by 32.4%. Second, we observed in our PVT databases (and were also able to show in this analysis) that short ISIs were associated with longer RTs, which we hypothesized was due to a central nervous system refractory period following a response and in preparation for the next stimulus. By capitalizing on this effect, we intended to increase sensitivity of the PVT-B. Third, we hypothesized that the higher cognitive workload associated with an increased stimulus density would increase the time on task effect, and therefore more quickly unmask sleepiness compared to a 3-min version with standard ISIs.

A comparison of RT distributions of the PVT-B and the PVT revealed that, both in alert and sleep deprived subjects, RTs were shorter, false start rate was higher, and lapse frequency was lower on the PVT-B compared to the PVT. This could be explained by differences in hardware (personal computer versus PVT-192 device), knowledge of test duration, increased stimulus density in the PVT-B, and by the fact that in > 90% of the trials the PVT-B was performed after the PVT. Even after controlling for differences in ISI, time on task, and test order, RTs on the PVT-B were still significantly faster compared to the PVT. Therefore, faster RTs on the PVT-B were likely due to hardware differences (stimulus presentation, response buttons, hardware response latencies) or knowledge of test duration. Systematic comparisons on the same hardware platform are needed to investigate the magnitude of the effect of knowledge of test duration.

We operationalized effect size as a measure of the PVT's ability or power to differentiate alert from sleep deprived subjects [12]. Effect size addresses more than just the sensitivity of the PVT, which was used as the validation criterion by other authors. [32, 33, 36] The PVT has to be sensitive (indicate high levels of sleepiness in sleep deprived subjects), specific (indicate low levels of sleepiness in alert subjects) and do this consistently in order to achieve high effect sizes. Our analyses showed that effect sizes of the PVT-B were consistently lower compared to the PVT. This is only partially in line with our previous work where we compared the 10-min PVT to the first 3 min of the same 10-min PVT test bout for 10 different PVT outcome metrics. [12] That analysis found lower effect sizes for the first 3 min of the PVT only in 70% of the outcome metrics in both TSD and PSD. The fact that in this study subjects performed two distinct tests on different hardware platforms with altered ISIs in the 3-min version of the test and with the knowledge of different test durations may have contributed to this discrepancy. More studies using the same hardware for both tests and counter-balancing the order of test administration in a cross-over fashion are needed to elucidate the differences in effect sizes found between the two versions of the test.

Despite the above factors, effect sizes for the PVT-B were still substantial, and compared to the 70% decrease in test duration, the 22.7% average decrease in effect size was acceptable. According to Cohen's criteria [41], all outcome metrics scored large effect sizes (>0.8) on the PVT-B in TSD. In the PSD study, all outcome metrics scored large effect sizes on the PVT. On the PVT-B, only mean 1/RT and slowest 10% 1/RT still scored large effect sizes. The effect sizes of lapses (both 500 ms and 355 ms definitions) and the performance score dropped to medium (>0.5 and <0.8), while the effect size of fastest 10% RT dropped to low (>0.2 and <0.5). Thus, it was shown that the utility of the PVT-B depends on the outcome metric.

Comparable to our analysis on optimal outcome metrics and task durations of the PVT [12], the highest effect sizes were observed for the reciprocal measures 1/RT and slowest 10% 1/RT for both the PVT-B and the PVT, and during both TSD and PSD (with the exception that the performance score's effect size was higher than that of the slowest 10% 1/RT on PVT-B in TSD). This highlights the favorable properties of the reciprocal outcomes, which reflect response slowing in the pre-lapse domain (i.e. RTs < 500 ms) and effectively remove the

influence of outlying long RTs. The reciprocal outcomes also showed very good coherence between PVT-B and the PVT with high p-values for the interaction between test version and test time, and they were the only two variables scoring large effect sizes on the PVT-B in PSD.

One advantage of the newly developed performance score is its easy interpretability. Although in terms of effect size it ranked only in 5th (TSD) and 6th (PSD) position on the PVT, the differences in effect size between the PVT and the PVT-B were lowest for this outcome measure, which is why it ranked in 2nd (TSD) and 3rd (PSD) position on the PVT-B. This is probably due to the fact that it takes both errors of omission (lapses) and errors of commission (false starts) into account, and therefore penalizes the bias towards faster RTs observed in PVT-B performance. Both the easy interpretability and these favorable statistical properties make the performance score a potential candidate for a primary outcome measure of the PVT-B. It is currently used to give astronauts feedback on their Reaction Self Test performance, a Microsoft Windows based version of the PVT-B, on board the International Space Station.

The PVT-B tracked the PVT closely over time in both TSD and PSD, especially for the reciprocal outcome measures (i.e., response speed). The increase in the frequency of 500 ms lapses during SD was less pronounced for the PVT-B compared to the PVT, as indicated by a significant interaction between test version and test time for this outcome metric. This is most likely a side effect of the overall decrease in PVT-B response times, as lowering the lapse threshold for the PVT-B diminished the difference in the number of lapses between tests to nonsignificant levels, even though the number of stimuli was lower for the PVT-B. Also, the increase in the fastest 10% RT was less pronounced for the PVT-B during both TSD and PSD compared to the PVT, with the highest observed drop in effect size of 67.8% for this measure during PSD. This could be explained by a general response bias towards faster RTs in the PVT-B, which would even be enhanced by increased compensatory effort during SD. The latter may be sufficient to keep the fastest 10% RTs low during a 3-min, but not during a 10-min version of the test.

4.1 Limitations

Several limitations have to be considered when interpreting the findings from this analysis. First, test duration, hardware, and ISIs were changed simultaneously for PVT-B relative to the PVT. Although we were able to shed some light onto the contributions of these factors to the observed differences in response times between both test versions, it would still be very valuable to perform a counterbalanced cross-over study comparing both versions of the PVT using the same hardware. Second, the 355 ms lapse threshold for the PVT-B was found post-hoc in the TSD experiment, and may depend on PVT hardware. Although its utility was confirmed in the PSD experiment, further studies are needed to confirm the adequacy of the 355 ms lapse threshold for the PVT-B. Third, the PVT was performed once every 2 hours while the PVT-B was performed only once every 4 hours in the PSD protocol, which probably affected the comparison of both tests. However, we believe that our results are conservative as the difference in test frequency most likely decreased rather than increased the agreement between both versions of the PVT. Finally, our subject sample consisted of healthy, young to middle-aged subjects. Our findings may therefore not generalize to other populations.

4.2 Conclusions

This is the first time a modified 3-min version of the PVT was validated against the standard 10-min PVT during both TSD and PSD. Our analyses show that the PVT-B differentiated alert from sleep deprived subjects somewhat less than the standard 10-min PVT for all

investigated outcome variables and during both TSD and PSD. However, effect sizes of the PVT-B were still large for all outcome metrics in TSD and (with the exception of fastest 10% RT) medium to large in PSD. Relative to the 70% decrease in test duration the 22.7% average decline in effect sizes of the PVT-B was deemed an acceptable trade-off between duration and sensitivity. The reciprocal outcome metrics mean 1/RT and slowest 10% 1/RT and the performance score were identified as candidates for primary outcome metrics for the PVT-B as they scored the largest effects sizes and/or the decrease in effect size compared to the PVT was relatively minor. Also, with the exception of fastest 10% RT in PSD and after lowering the lapse threshold for PVT-B from 500 ms to 355 ms, no statistical differences were found between both tests for all outcome variables and during both TSD and PSD. Therefore, we were able to show that the 3-min PVT-B remains a sensitive and specific assay for detecting wake-state instability induced by both total and partial sleep deprivation [14]. It may be a useful tool in applied settings where use of the standard 10-min PVT is not feasible or undesirable. The validity of the PVT-B still needs to be established in such settings.

Research Highlights

- The Psychomotor Vigilance Test (PVT) measures behavioral alertness.
- A brief 3 min version of the PVT remains sensitive to the effects of sleep loss.
- Its utility is currently evaluated on astronauts on board the International Space Station ISS.
- The brief PVT may be practical for many operational and clinical environments.

Acknowledgments

This investigation was sponsored by the National Space Biomedical Research Institute through NASA NCC 9–58, and in part by NASA grant NNX08AY09G, NIH grants M01-RR00040, NR04281 and CTRC UL1 RR0241340, and by the Department of Homeland Security's Transportation Security Laboratory Human Factors Program (FAA #04-G-010).

LITERATURE

1. Committee on Sleep Medicine Research. Sleep disorders and sleep deprivation: an unmet public health problem. Washington, DC: National Academic Press; 2006.
2. Banks S, Dinges DF. Behavioral and physiological consequences of sleep restriction. J.Clin.Sleep Med. 2007; 3:519–528. [PubMed: 17803017]
3. Basner M, Dinges DF. Dubious bargain: trading sleep for Leno and Letterman. Sleep. 2009; 32:747–752. [PubMed: 19544750]
4. Basner M, Fomberstein KM, Razavi FM, Banks S, William JH, Rosa RR, Dinges DF. American time use survey: sleep time and its relationship to waking activities. Sleep. 2007; 30:1085–1095. [PubMed: 17910380]
5. Smith L, Macdonald I, Folkard S, Tucker P. Industrial shift systems. Appl Ergon. 1998; 29:273–280. [PubMed: 9701542]
6. Philip P, Akerstedt T. Transport and industrial safety, how are they affected by sleepiness and sleep restriction? Sleep Medicine Reviews. 2006; 10:347–356. [PubMed: 16920370]
7. Dinges DF. An overview of sleepiness and accidents. J.Sleep Res. 1995; 4:4–14. [PubMed: 10607205]
8. Frey DJ, Badia P, Wright KP Jr. Inter- and intra-individual variability in performance near the circadian nadir during sleep deprivation. J.Sleep Res. 2004; 13:305–315. [PubMed: 15560765]

9. 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; 27:423–433. [PubMed: 15164894]
10. Dinges, DF.; Mallis, M. Managing fatigue by drowsiness detection: can technological promises be realized?. Hartley, L., editor. Pergamon; 1998. p. 209-229.
11. Gilliland, K.; Schlegel, RE. Readiness to perform: A critical analysis of the concept and current practices. Office of Aviation Medicine, Federal Aviation Administration; 1993.
12. Basner M, Dinges DF. Maximizing sensitivity of the Psychomotor Vigilance Test (PVT) to sleep loss. Sleep. 2011; 34:581–591. [PubMed: 21532951]
13. Dorrian, J.; Rogers, NL.; Dinges, DF.; Kushida, CA. Sleep Deprivation: Clinical Issues, Pharmacology and Sleep Loss Effects. New York, NY: Marcel Dekker, Inc.; 2005. Psychomotor vigilance performance: Neurocognitive assay sensitive to sleep loss; p. 39-70.
14. Lim, J.; Dinges, DF. Molecular and Biophysical Mechanisms of Arousal, Alertness, and Attention. Oxford: Annals of the New York Academy of Sciences, Blackwell Publishing; 2008. Sleep deprivation and vigilant attention; p. 305-322.
15. Patrick GTW, Gilbert JA. On the effects of sleep loss. Psychol.Rev. 1896; 3:469–483.
16. Dinges DF, Powell JW, Microcomputer analysis of performance on a portable. simple visual RT task during sustained operations. Behav.Res.Methods Instrum.Comput. 1985; 6:652–655.
17. 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.
18. 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; 49:3426–3435. [PubMed: 19925871]
19. Dinges DF, Pack F, Williams K, Gillen KA, Powell JW, Ott GE, Aptowicz C, Pack AI. Cumulative sleepiness, mood disturbance, and psychomotor vigilance performance decrements during a week of sleep restricted to 4–5 hours per night. Sleep. 1997; 20:267–277. [PubMed: 9231952]
20. Jewett ME, Dijk DJ, Kronauer RE, Dinges DF. Dose-response relationship between sleep duration and human psychomotor vigilance and subjective alertness. Sleep. 1999; 22:171–179. [PubMed: 10201061]
21. 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; 26:117–126. [PubMed: 12683469]
22. Belenky G, Wesensten NJ, Thorne DR, Thomas ML, Sing HC, Redmond DP, Russo MB, Balkin TJ. Patterns of performance degradation and restoration during sleep restriction and subsequent recovery: a sleep dose-response study. J.Sleep Res. 2003; 12:1–12. [PubMed: 12603781]
23. 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–R1163. [PubMed: 10516257]
24. 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. [PubMed: 14984804]
25. Neri DF, Oyung RL, Colletti LM, Mallis MM, Tam PY, Dinges DF. Controlled Breaks as a Fatigue Countermeasure on the Flight Deck. Aviation Space and Environmental Medicine. 2002; 73:654–664.
26. Kribbs NB, Pack AI, Kline LR, Getsy JE, Schuett JS, Henry JN, Maislin G, Dinges DF. Effects of one night without nasal CPAP treatment on sleep and sleepiness in p patients with obstructive sleep apnea. Am Rev Respir Dis. 1993; 147:1162–1168. [PubMed: 8484626]
27. Wesensten NJ, Belenky G, Thorne DR, Kautz MA, Balkin TJ. Modafinil vs. caffeine: effects on fatigue during sleep deprivation. Aviat Space Environ Med. 2004; 75:520–525. [PubMed: 15198278]

28. Wyatt JK, Cajochen C, Ritz-De Cecco A, Czeisler CA, Dijk DJ. Low-dose repeated caffeine administration for circadian-phase-dependent performance degradation during extended wakefulness. Sleep. 2004; 27:374–381. [PubMed: 15164887]
29. Signal TL, Gander PH, Anderson H, Brash S. Scheduled napping as a countermeasure to sleepiness in air traffic controllers. J.Sleep Res. 2009; 18:11–19. [PubMed: 19250171]
30. Balkin TJ, Bliese PD, Belenky G, Sing H, Thorne DR, Thomas M, Redmond DP, Russo M, Wesensten NJ. Comparative utility of instruments for monitoring sleepiness-related performance decrements in the operational environment. J Sleep Res. 2004; 13:219–227. [PubMed: 15339257]
31. Anderson C, Wales AW, Horne JA. PVT lapses differ according to eyes open, closed, or looking away. Sleep. 2010; 33:197–204. [PubMed: 20175403]
32. 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; 36:339–346. [PubMed: 15354700]
33. 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; 40:347–352. [PubMed: 18411559]
34. 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; 23:1379–1387. [PubMed: 17190720]
35. Lamond N, Dawson D, Roach GD. Fatigue assessment in the field: validation of a handheld electronic psychomotor vigilance task. Aviat.Space Environ.Med. 2005; 76:486–489. [PubMed: 15892548]
36. Thorne DR, Johnson DE, Redmond DP, Sing HC, Belenky G, Shapiro JM. The Walter Reed palm-held psychomotor vigilance test. Behavior Research Methods. 2005; 37:111–118. [PubMed: 16097350]
37. Basner M, Rubinstein J, Fomberstein KM, Coble M, Ecker AJ, Avinash D, Dinges DF, Effects of night work. sleep loss and time on task on simulated threat detection performance. Sleep. 2008; 31:1251–1259. [PubMed: 18788650]
38. 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; 33:1013–1026. [PubMed: 20815182]
39. Efron, B.; Tibshirani, RJ. An introduction to the bootstrap. first ed. New York, NY: Chapman & Hall; 1993.
40. Curran-Everett D. Multiple comparisons: philosophies and illustrations. Am.J Physiol Regul.Integr.Comp Physiol. 2000; 279:R1–R8. [PubMed: 10896857]
41. Cohen, J. Statistical power analysis for the behavioral sciences. 2nd edition ed. Hillsdale, NJ: Lawrence Erlbaum; 1988.

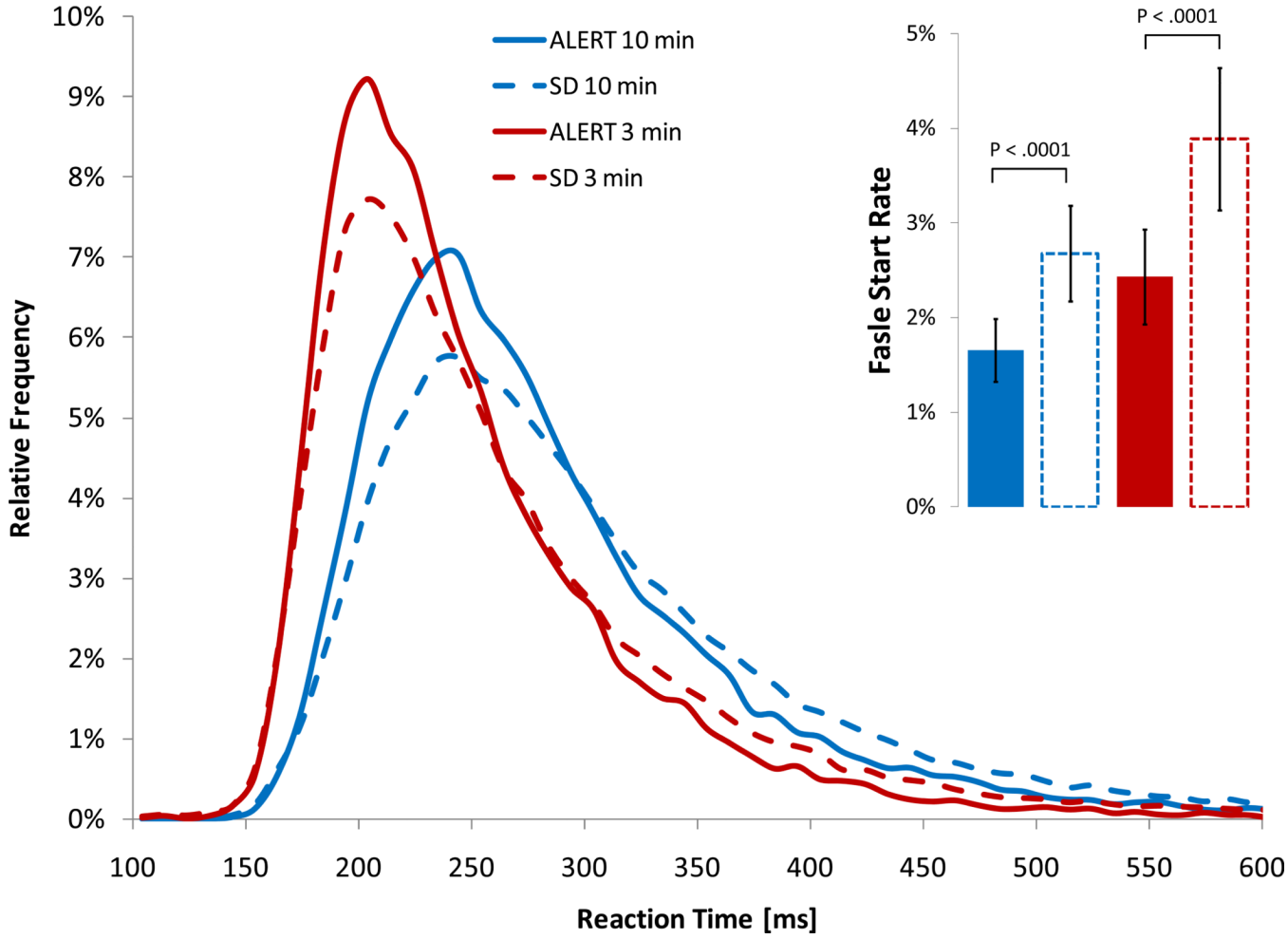

Figure 1.

Relative frequency distributions of PVT response time are shown for alert (bouts 1 to 7 during total sleep deprivation and baseline 2 during partial sleep deprivation) and sleep deprived states (bouts 8 to 17 during total sleep deprivation and restriction nights 1 to 5 during partial sleep deprivation) for both the modified 3-min (PVT-B) and the 10-min version of the PVT. The insert shows the frequency of false starts (errors of commission) including 95% confidence intervals.

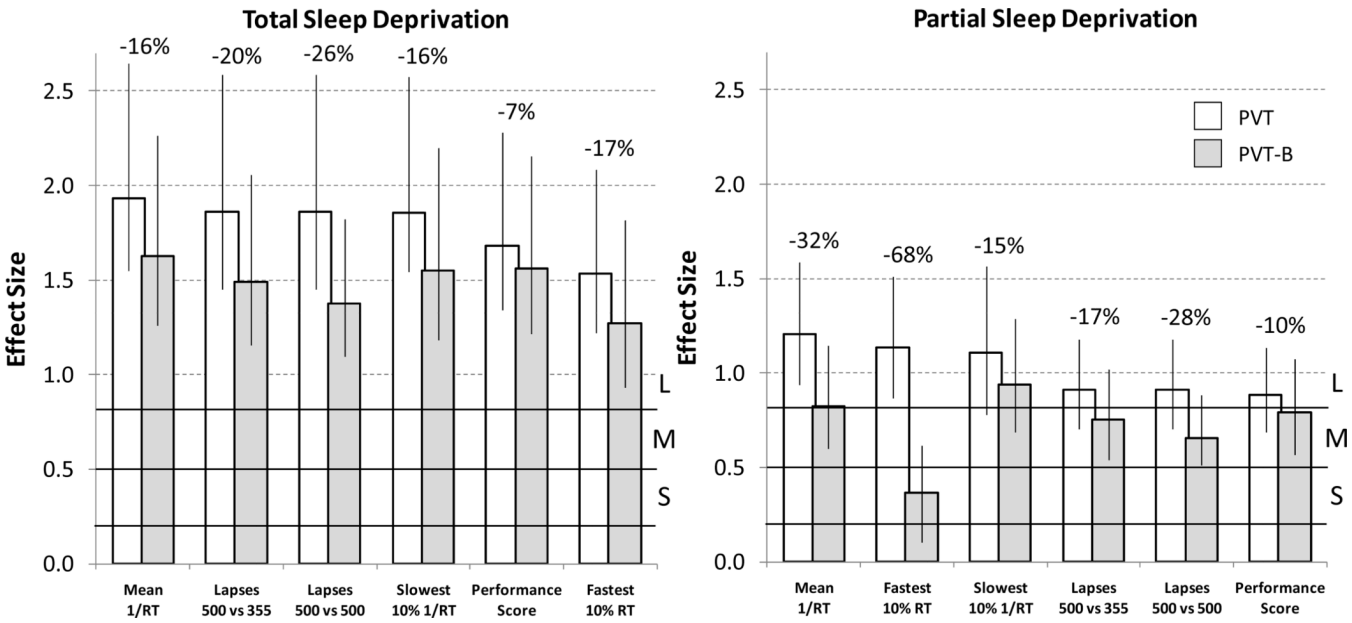

Figure 2.

Effect sizes ± 95% bootstrap confidence intervals are compared between the standard 10-min PVT (PVT) and the modified 3-min version of the PVT (PVT-B) for 5 outcome metrics and for both total (left) and partial (right) sleep deprivation. For PVT-B, both standard 500 ms and modified 355 ms lapse thresholds were applied. Ranges representing small (S, >0.2 and <0.5), medium (M, >0.5 and <0.8), and large (L, >0.8) effect sizes according to Cohen [41] are indicated by black horizontal lines. The relative decrease in effect size from PVT to PVT-B is indicated as percentages above each outcome metric.

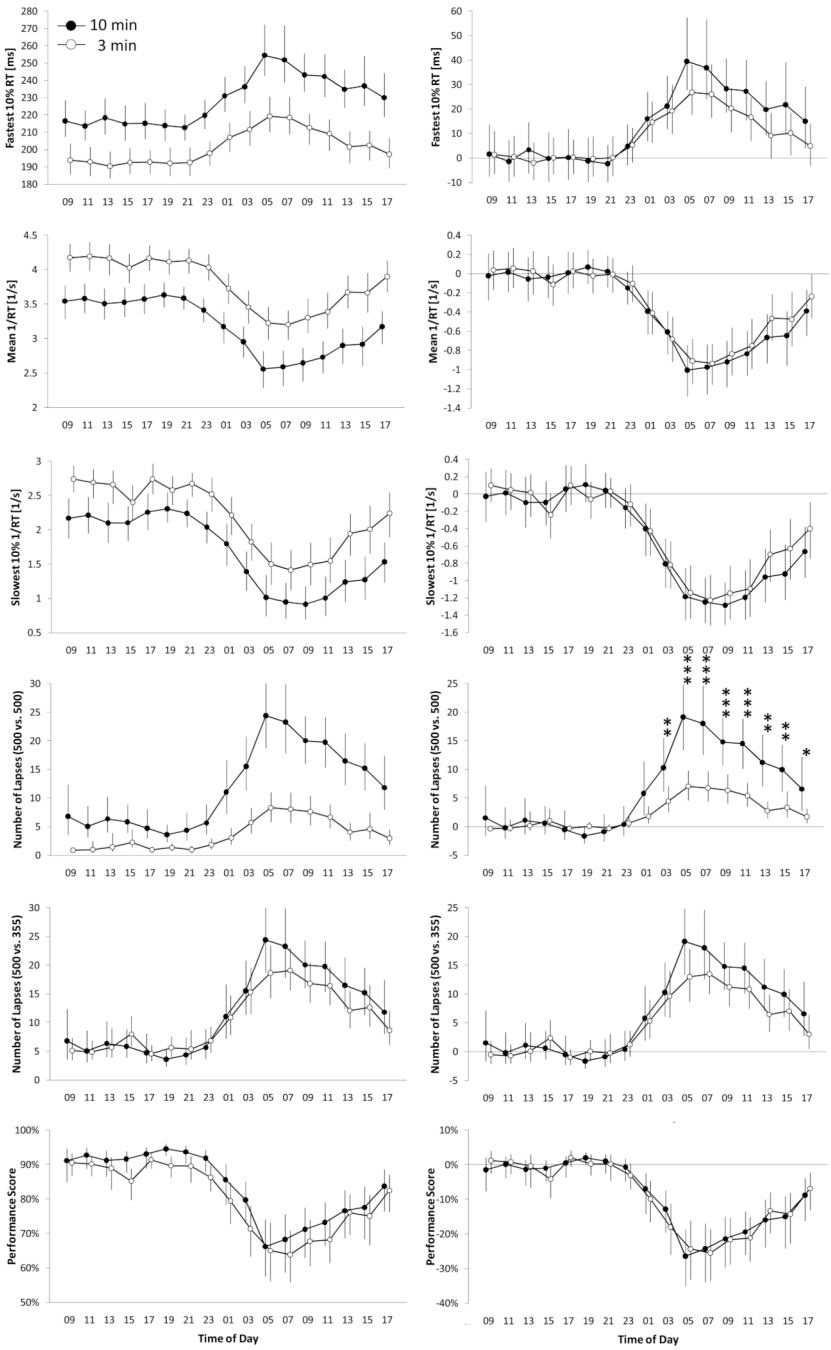

Figure 3.

For each of the 6 outcome variables, between-subject averages (N=31 subjects) are shown for each of the 17 tests performed during a 33 h period of total sleep deprivation for both the 10-min (black circles) and the 3-min (open circles) PVT. Error bars represent 95% BCa confidence intervals based on a bootstrap sample with 1,000,000 replications. In the right column of the figure, the 5 outcome variables of the 3-min and the 10-min PVT were centered around alert performance (average of test bouts 1 to 7). Paired t-tests were performed on each of test bouts 8 to 17 during sleep deprivation to test whether the modified 3-min (PVT-B) and the 10-min PVT differed statistically significantly. * $p<0.05$, ** $p<0.01$, *** $p<0.001$ (adjusted for multiple testing)

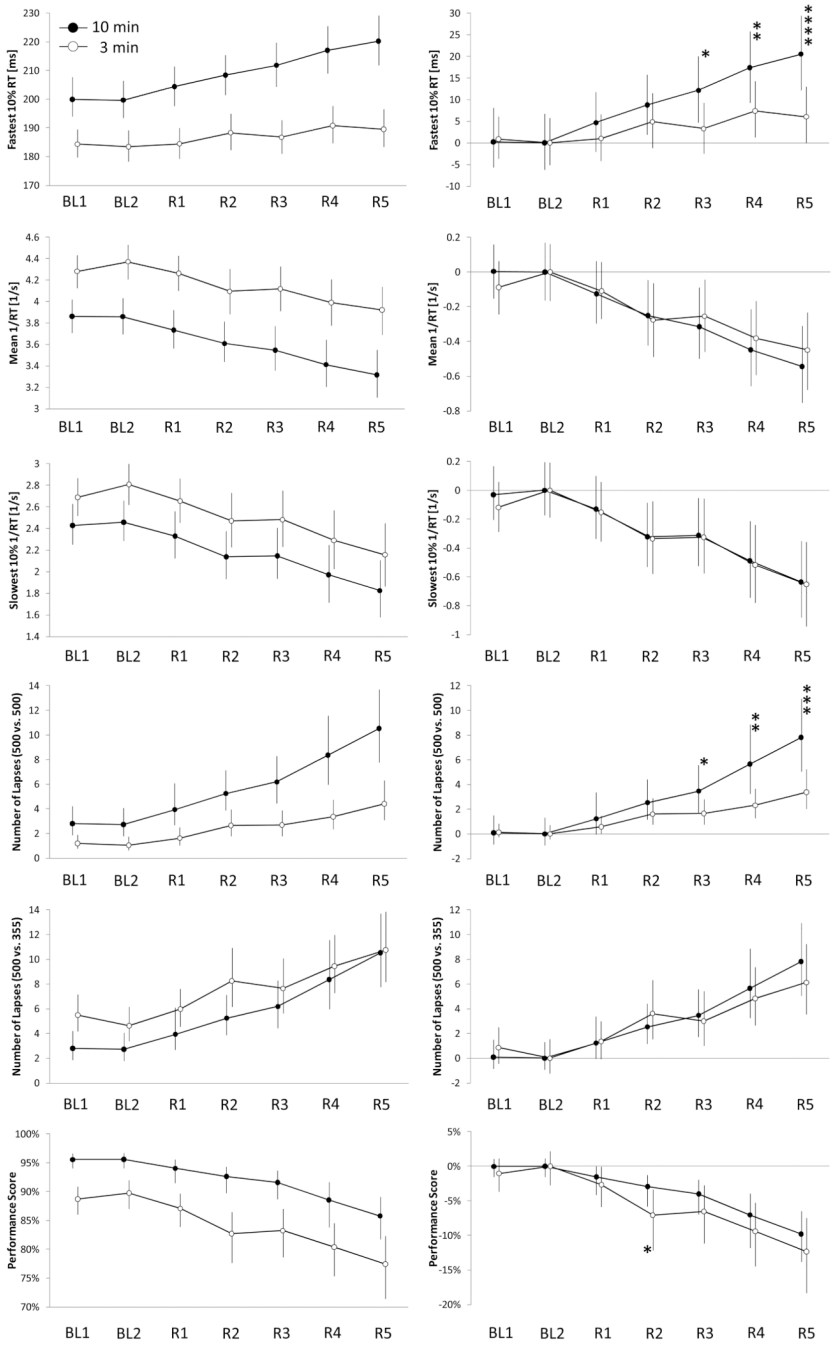

Figure 4.

For each of the 6 outcome variables, between-subject averages (N=43 subjects) are shown for each of the 7 conditions of the partial sleep deprivation protocol (BL = baseline, R = restriction) for both the modified 3-min PVT-B (open circles) and the 10-min PVT (black circles). Error bars represent 95% BCa confidence intervals based on a bootstrap sample with 1,000,000 replications. In the right column of the figure, the 5 outcome variables of the 3-min PVT-B and the 10-min PVT were centered around alert performance (BL2). Paired t-tests were performed on each of the five sleep restriction conditions (R1 – R5) to test whether the 3-min and the 10-min PVT differed statistically significantly. * $p<0.05$, ** $p<0.01$, *** $p<0.001$ (adjusted for multiple testing)

Table 1

Results of a mixed model ANOVA with random subject effect investigating the influence of several variables on log-transformed response time in the total sleep deprivation protocol.

Variable	Degrees of Freedom	F-Value	P-Value
Hours awake	16; 80372	646.5	<.0001
Inter stimulus interval	9; 80372	392.8	<.0001
Time on task	9; 80372	127.9	<.0001
Sleep deprivation start day	1; 28	0.6	0.4286
Test order	1; 28	3.7	0.0661
Test version	1; 80372	3450.2	<.0001

Degrees of Freedom (DF) shown as numerator DF; denominator DF. Denominator DF were calculated with Satterthwait's method (unadjusted denominator DF are reported). P-values of type 3 effects are shown.

Table 2

Results of a mixed model ANOVA with random subject effect investigating differences between the modified brief and the full 10-min PVT for 5 outcome metrics.

Outcome Metric	Total Sleep Deprivation			Partial Sleep Deprivation		
	Version (DF 1; 990)	Time (DF 16; 990)	Version*Time (DF 16; 990)	Version (DF 1; 546)	Time (DF 6; 546)	Version*Time (DF 6; 546)
Fastest 10% RT	<.0001	<.0001	0.2452	<.0001	<.0001	**<.0001**
Mean RRT	<.0001	<.0001	0.9217	<.0001	<.0001	0.3900
Slowest 10% RRT	<.0001	<.0001	0.7735	<.0001	<.0001	0.9884
Number of Lapses (500 ms vs. 500 ms)	<.0001	<.0001	**<.0001**	<.0001	<.0001	**<.0001**
Number of Lapses (500 ms vs. 355 ms)	0.0056	<.0001	0.1554	<.0001	<.0001	0.3531
Performance Score (500 ms vs. 355 ms)	<.0001	<.0001	0.9462	<.0001	<.0001	0.6084

Degrees of Freedom (DF) shown as numerator DF; denominator DF. Denominator DF were calculated with Satterthwait's method (unadjusted denominator DF are reported). P-values of type 3 effects are shown.