http://dx.doi.org/10.5665/sleep.1620

An Adaptive-Duration Version of the PVT Accurately Tracks Changes in Psychomotor Vigilance Induced by Sleep Restriction

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

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

Study Objectives: The Psychomotor Vigilance Test (PVT) is a widely used assay of behavioral alertness sensitive to the effects of sleep loss and circadian misalignment. The standard 10-minute duration of the PVT is often considered impractical for operational or clinical environments. Therefore, we developed and validated an adaptive-duration version of the PVT (PVT-A) that stops sampling once it has gathered enough information to correctly classify PVT performance.

Design: Repeated-measures experiments involving 10-minute PVT assessments every 2 hours across both acute total sleep deprivation (TSD) and 5 days of chronic partial sleep deprivation (PSD).

Setting: Controlled laboratory environment.

Participants: Seventy-four healthy subjects (34 women), aged 22 to 45 years.

Interventions: A TSD experiment involving 33 hours awake (n = 31 subjects), and a PSD experiment involving 5 nights of 4 hours time in bed (n = 43 subjects).

Measurements and Results: The PVT-A algorithm was trained with 527 TSD test bouts and validated with 880 PSD test bouts. Based on our primary outcome measure "number of lapses (response times ≥ 500 ms) plus false starts (premature responses or response times < 100 ms)," 10-minute PVT performance was classified into high (≤ 5 lapses and false starts), medium (> 5 and ≤ 16 lapses and false starts), or low (> 16 lapses and false starts). The decision threshold for PVT-A termination was set so that at least 95% of training data-set tests were classified correctly and no test was classified incorrectly across 2 performance categories (i.e., high as low or low as high), resulting in an average test duration of 6.0 minutes (SD 2.4 min). In the validation data set, 95.7% of test bouts were correctly classified, and there were no incorrect classifications across 2 categories. Agreement corrected for chance was excellent (κ = 0.92). Across the 3 performance categories, sensitivity averaged 93.7% (range 87.2%-100%), and specificity averaged 96.8% (range 91.6%-99.9%). Test duration averaged 6.4 minutes (SD 1.7 min), with a minimum of 27 seconds.

Conclusions: We developed and validated a highly accurate, sensitive, and specific adaptive-duration version of the 10-minute PVT. Test duration of the adaptive PVT averaged less than 6.5 minutes, with 60 tests (4.3%) terminating after less than 2 minutes, increasing the practicability of the test in operational and clinical settings. The adaptive-duration strategy may be superior to a simple reduction of PVT duration in which the fixed test duration may be too short to identify subjects with moderate impairment (showing deficits only later during the test) but unnecessarily long for those who are either fully alert or severely impaired.

Keywords: PVT, psychomotor vigilance, sleep deprivation, alertness, attention, lapse, response speed, response time, sensitivity

Citation: Basner M; Dinges DF. An adaptive-duration version of the PVT accurately tracks changes in psychomotor vigilance induced by sleep restriction. *SLEEP* 2012;35(2):193-202.

INTRODUCTION

Undisturbed sleep of sufficient length is a prerequisite for recuperation and optimal performance during the wake period.[1] Acute total sleep loss, chronic partial sleep deprivation (PSD), and their combination impair performance and increase the risk of sleepiness-related errors and accidents.[2-4] Both intrinsic sleep disorders (such as obstructive sleep apnea[5]) and extrinsic sleep disturbances (e.g., induced by noise[6]) are prevalent and underrecognized in the population. At the same time, shift work affects the sleep and alertness of approximately 1 out of 5 working Americans, and is projected to increase further. Shift work includes working evenings, nights, or rotating shifts and is often associated with shorter-than-normal and disrupted sleep periods at an adverse circadian phase.[7] Finally, in our 24/7 society, sleep is perceived by many as a flexible commodity of lesser value that can be exchanged for waking activities considered more essential or of greater value.[8,9]

All of the above demonstrate the high prevalence of chronic PSD and acute total sleep loss in our society that are associated with a high burden of sleepiness-related errors and accidents. At the same time, sleepy subjects have been shown to be unable to reliably assess their degree of impairment, stressing the need for brief, validated, and objective measures of cognitive performance.[10-12]

Among the most reliable effects of sleep deprivation is degradation of attention,[13,14] especially vigilant attention, as measured by the 10-minute Psychomotor Vigilance Test (PVT).[15,16] The standard PVT records response times to visual stimuli that occur at random 2- to 10-second interstimulus intervals (ISI) over a10-minute period.[16-20] 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, usually defined as response times ≥ 500 ms), and a more modest increase in the number of errors of commission (i.e., responses without a stimulus, or false starts).[10,21] These effects are associated with changes in neural activity in distributed brain regions that can include frontal and parietal control regions, visual and insular corti-

Submitted for publication May, 2011
Submitted in final revised form September, 2011
Accepted for publication September, 2011

Address correspondence to: Mathias Basner, MD, PhD, MSc, Division of Sleep and Chronobiology, Department of Psychiatry, University of Pennsylvania Perelman 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

ces, cingulate gyrus, and the thalamus, and they can increase as time on task increases.[22-26]

PVT performance also has ecologic validity in that it can reflect real-world risks because deficits in sustained attention and timely reactions adversely affect many applied tasks (e.g., all transportation modes, many security-related tasks, and a wide range of industrial tasks).[4,27-29] Therefore, the PVT 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 and its psychometric advantages over other cognitive tests (i.e., negligible aptitude and learning effects).[15,16,20]

However, the 10-minute standard duration of the PVT is regarded by many as too long for applied, operational, or clinical settings. For this reason, shorter PVT versions with durations of 3 and 5 minutes have been developed that usually run on handheld devices.[30-35] However, although some of the shorter-duration PVT versions may track 10-minute PVT performance closely, they usually lose some sensitivity relative to the standard 10-minute PVT. The latter can be explained by the fact that performance on the PVT deteriorates with time on task (so-called vigilance decrement), faster in sleep-deprived than in alert subjects.[20,36] Thus, on the one hand, the shorter PVT versions seem to be too short to detect relevant deterioration in vigilant attention in subjects with moderate impairment whose performances deteriorate only later during the test, whereas, on the other hand, the longer versions may be unnecessarily long for other subjects who are apparently fully alert or severely impaired.

This prompted us to develop an adaptive-duration version of the 10-minute PVT (PVT-A). After each response and depending on the nature of the response, the PVT-A algorithm reevaluates the probability of a subject being a high, medium, or low performer on the full 10-minute PVT. If a predefined decision threshold is exceeded, the PVT-A stops sampling data, as it has determined that it has gathered enough information to make a correct decision.

We argued earlier[37] that 1 threshold dividing outcomes in high and low performance may be insufficient in fit-for-duty paradigms, as it is questionable whether subjects performing just above or below the single decision threshold are really fit or unfit to perform the task. Therefore, we chose instead to divide the dataset into 3 performance categories (high, medium, and low). The medium performance category separates the high-performance category (subjects are fit for the task) from the low-performance category (subjects are unfit for the task and must not perform it). The consequences for subjects falling in the medium-performance category may vary depending on the relevance of the task. If subjects are allowed to perform the task, informing them about their decreased level of alertness may improve their effort and inspire them to apply countermeasures aiming at short-term (e.g., break, caffeine) or long-term (e.g., increasing individual sleep times) improvements of alertness. The latter has been shown in a study of truck drivers.[38] Similar arguments could be brought forward for diagnostic paradigms, in which 2 groups indicating “not impaired” and “impaired” may not be sufficient.

We show here that average duration of the PVT-A decreases substantially relative to the standard 10-minute PVT and that PVT-A predictions are highly accurate with excellent chance-corrected agreement relative to the standard 10-minute PVT.

METHODS

Subjects and Protocol

The following descriptions of the total sleep deprivation (TSD) and partial sleep deprivation (PSD) protocols are in part reproduced from Basner and Dinges.[20]

Acute TSD protocol

TSD data were gathered from 36 subjects in a study on the effects of night work and sleep loss on threat-detection performance on a simulated luggage-screening task (a detailed description of the study is published elsewhere[33]). Study participants stayed in the research lab for 5 consecutive days, which included a 33-hour period of TSD. Data from 4 subjects were excluded from the analysis due to noncompliance and/or excessive fatigue during the first 16 hours of wakefulness. Another subject withdrew after 26 hours awake. Therefore, data from a subset of 31 subjects (mean age ± SD = 31.1 ± 7.3 y, 18 women) contributed to the analyses presented here. The study started at 08:00 on day 1 and ended at 08:00 am on day 5. A 33-hour period of TSD 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-hour sleep opportunities between 00:00 and 08:00. The first sleep period was monitored polysomnographically to exclude possible sleep disorders.

Chronic PSD protocol

The PSD data were obtained from 47 healthy adults in a laboratory protocol involving 5 consecutive nights of sleep restricted to 4 hours per night (04:00 to 08:00 period) following 2 baseline nights with 10 hours time in bed on each night. Data from 3 subjects were excluded from the analysis due to noncompliance and/or excessive fatigue during baseline data collection. One additional subject had no valid baseline data. Therefore, data from 43 subjects (16 women; mean age, 30.5 ± 7.3 y) contributed to the analyses presented here. A detailed description of the experimental procedures is published elsewhere.[39]

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 examinations, 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 and 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°C and 24°C.

In both TSD and PSD experiments, subjects completed 30-minute bouts of a neurobehavioral test battery that included a 10-minute PVT every 2 hours during scheduled wakefulness.

In the TSD experiment, each subject performed 17 PVTs in total (starting at 09:00 after a sleep opportunity from 00:00 to 08:00 with bout #1 and ending at 17:00 on the next day after a night without sleep with bout #17). The data of the TSD protocol were complete, and, thus, 527 test bouts contributed to the analysis. Consistent with previous publications,[20,35] we only used the test bouts administered at 12:00, 16:00, and 20:00 on baseline days 1 and 2 and on days after restriction nights 1 to 5 in the PSD experiment. Of the 903 scheduled test bouts, 23 (2.5%) were missing, and, thus, 880 test bouts contributed to the analysis. Between neurobehavioral test bouts, subjects were permitted to read, watch movies and television, play card or board games, and interact with laboratory staff to help them stay awake, but no naps or 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 Institutional Review Board approval were obtained prior to the start of the study. Subjects 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 studies.

PVT

We utilized a precise computer-based version of the 10-minute PVT, which was performed and analyzed according to the standards set forward in Basner and Dinges.[20] Subjects were instructed to monitor a red rectangular box on the computer screen and press a response button as soon as a yellow stimulus counter appeared on the CRT screen, which stopped the counter and displayed the response time in milliseconds for a 1-second period. The ISI, defined as the period between the last response and the appearance of the next stimulus, varied randomly from 2 to 10 seconds. The subject was instructed to press the button as soon as each stimulus appeared, in order to keep the response time as low as possible but to not press the button too soon (which yielded a false-start warning on the display).

PVT-A Algorithm

Outcome metric and performance group classification

We decided to use the sum of the number of lapses (i.e., errors of omission, defined as a response time ≥ 500 ms) and the number of false starts (i.e., errors of commission, defined as responses without stimulus or responses with response times < 100 ms) as the primary outcome metric. In a systematic comparison of PVT outcome metrics (using the same data set that was used for this analysis), Basner and Dinges[20] found that the number of lapses and false starts scored a high effect size in PSD (0.90) and the highest effect size in TSD (1.94) relative to the other 9 investigated outcome metrics. Also, taking false starts into account may help to identify noncompliant subjects or those who try to prevent lapses by biasing toward false starts, which may be especially important in fit-for-duty contexts.

The TSD study was used to find number of lapses and false-start thresholds that divided 10-minute PVT test bouts into high-, medium-, and low-performance bouts (for the rationale of choosing 3 performance groups: see Introduction). First, 5 or fewer lapses and false starts was identified as the threshold that optimally differentiated test bouts performed until 21:00 (up to 13 hours awake) from test bouts performed at or after 23:00 (15 to 33 h awake). Choosing a cutoff between 21:00 and 23:00 was based on visual inspection of the data and on reports that PVT performance decreases after 16 hours of wakefulness.[10] A second threshold of 16 lapses and false starts was identified by performing a median split on all test bouts with more than 5 lapses and false starts. Therefore, the 3 groups were defined as follows: high (≤ 5 lapses and false starts, n = 193 bouts), medium (> 5 and ≤ 16 lapses and false starts, n = 167 bouts), and low (> 16 lapses and false starts, n = 167 bouts).

PVT-A algorithm description

Similar to an earlier published approach,[40] each response time on the PVT can be thought of as the result of a diagnostic test that will change our confidence in the test bout being a high-, medium-, or low-performance test bout. For example, in case of a lapse or a false start, the probability of being a high-performance test bout decreases, whereas the probability of being a low-performance test bout increases at the same time. Although we may assign equal probabilities to the 3 performance groups ($P_{HIGH} = P_{MEDIUM} = P_{LOW}$) before the subject's first response (termed *prior probability* in Bayesian language), these probabilities change based on the response-time outcome of the first stimulus (i.e., the prior probability is updated to the *posterior probability*). The posterior probability then serves as the new prior probability for the next stimulus, and the process is repeated until 1 of the 3 probabilities (P_{HIGH}, P_{MEDIUM}, or P_{LOW}) exceeds a predefined decision threshold (see below), which is when the test is stopped.

Formally, the posterior probability is calculated by transforming the prior probability (PrP) into the prior odds (PrO) according to equation (1):

$$\mathrm{PrO} = \mathrm{PrP} / (1 - \mathrm{PrP}) \qquad (1).$$

The prior odds is then multiplied with a likelihood ratio (LR) to receive the posterior odds (PstO), which is again transformed into the posterior probability (PstP) according to equation (2):

$$\mathrm{PstP} = \mathrm{PstO} / (1 + \mathrm{PstO}) \qquad (2).$$

The likelihood ratio depends on the response-time outcome of the stimulus. Relative to the prior probability, the posterior probability will increase for likelihood ratios greater than 1, decrease for likelihood ratios less than 1, and remain unchanged for likelihood ratios equal to 1. We only calculated likelihood ratios and posterior probabilities for P_{HIGH} and P_{LOW}. P_{MEDIUM} was then calculated according to equation (3):

$$P_{MEDIUM} = 1 - P_{HIGH} - P_{LOW} \qquad (3).$$

The 4 likelihood ratios LR(High | no lapse or false start), LR(High | lapse or false start), LR(Low | no lapse or false start), and LR(Low | lapse or false start) were calculated based on the TSD data (for a detailed description of likelihood ratio calculations see Hunink et al.).[41] To acknowledge time-on-task effects, we divided the 10-minute task duration into twenty

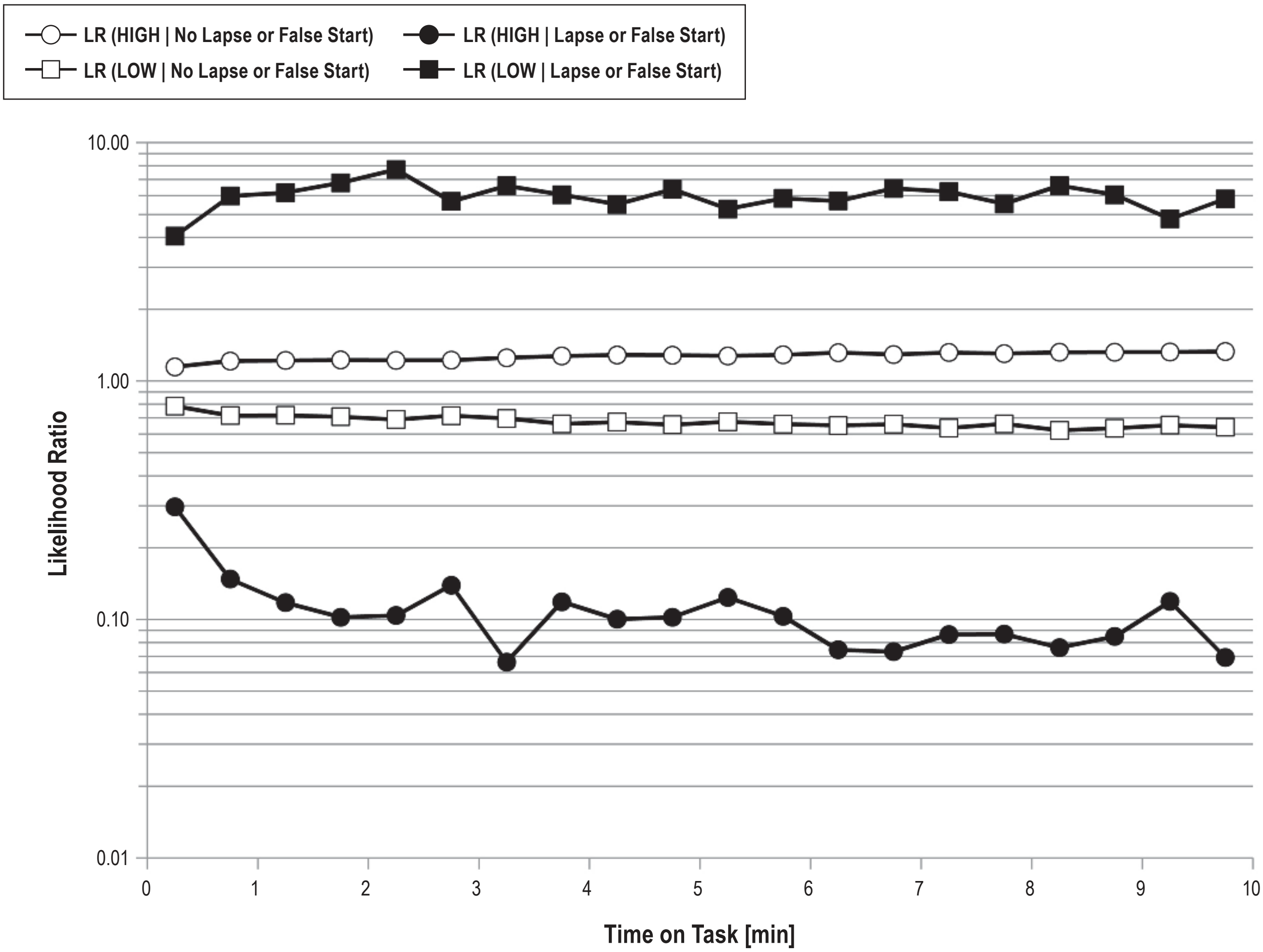

Figure 1—Likelihood ratios (LR) for belonging to the high psychomotor vigilance task (PVT) performance group (HIGH) or to the low PVT performance group (LOW) are given conditional on whether or not a lapse or a false start occurred and depending on time on task. Likelihood ratios are used to update the probability for belonging to a specific PVT performance group from before stimulus presentation (*prior probability*) to the probability after the response to the stimulus (*posterior probability*). The probability for belonging to a specific PVT performance group will increase for likelihood ratios > 1, decrease for likelihood ratios < 1, and remain unchanged for likelihood ratios = 1. The figure shows that in case of a lapse or a false start it will be less likely to belong to the HIGH performance group (black circles, LRs < 1) and more likely to belong to the LOW performance group (black squares, LRs > 1). Accordingly, in case of no lapse or a false start it will be more likely to belong to the HIGH performance group (open circles, LRs > 1) and less likely to belong to the LOW performance group (open squares, LRs < 1). Because lapses and false starts are rare events, they carry more information than do stimuli without lapses and false starts (i.e., likelihood ratios associated with lapses and false starts are more extreme and therefore lead to a greater change in posterior probability than do likelihood ratios with no lapses or false starts). Likelihood ratios were relatively stable across the 10-minute test period.

30-second intervals and calculated likelihood ratios for each interval (see Figure 1).

We assigned equal prior probabilities of 1/3 to P_{HIGH}, P_{MEDIUM}, and P_{LOW}, although the prevalence of high-performance bouts was slightly higher compared with medium- and low-performance bouts in the TSD study. We believe this better reflected the uncertainty of each individual test outcome. In addition to using the basic algorithm described above, after each stimulus, we checked for the following conditions: If the number of lapses and false starts exceeded 5, P_{HIGH} was set to 0, and the probabilities for belonging to the medium- and low-performance group were adjusted accordingly. If the number of lapses and false starts exceeded 16, the algorithm was stopped and the test result was classified as LOW. If the decision threshold (see below) was not exceeded before the full 10 minutes of the PVT, the test was classified according to the actual number of lapses and false starts (i.e., always correct), and the full 10 minutes were recorded for PVT-A duration.

The decision threshold was set to 99.9826% (i.e., once P_{HIGH}, P_{MEDIUM}, or P_{LOW} exceeded this probability, the test was classified accordingly and the algorithm stopped). This choice was made so that, in the training data set (TSD), more than 95% of the 527 decisions were correct, and there were no misclassifications across 2 categories (i.e., HIGH classified as LOW or LOW classified as HIGH). For each test, we recorded: (1) the true test classification based on the result of the full 10-minute PVT, (2) the classification of the PVT-A algorithm, and (3) the time the PVT-A algorithm needed to reach that decision.

The PVT-A algorithm was then validated with the 880 test bouts of the PSD data set using the same procedure and the decision threshold found in the training data set. We want to emphasize that, although each subject contributed several tests to the analysis, the performance classification was always based on a single test, and not on multiple tests, of the same subject.

Data Analysis

For both data sets and for each of the performance classifications HIGH, MEDIUM, and LOW, we calculated accuracy (i.e., percentage of correct decisions), sensitivity, specificity, and positive and negative predictive values (always relative to the remaining 2 categories). We also calculated several statistics for PVT-A test duration for each category. Finally, we calculated κ (a chance-corrected measurement of agreement) across the 3 performance categories.[42] With a nonlinear mixed-effects model controlling for experimental condition (Proc NLMIXED, SAS, Version 9.2, SAS, Inc., Cary, NC), we investigated whether HIGH, MEDIUM, and LOW classifications differed significantly between the full 10-minute PVT and PVT-A across 33 hours of TSD and across the 7 nights of the PSD protocol.

Table 1—Test characteristics and duration of the adaptive-duration PVT (PVT-A) relative to the full 10-minute PVT for the training data set and the validation data set

	Training Data Set (TSD)			Validation Data Set (PSD)		
Performance Category[a]	**HIGH**	**MEDIUM**	**LOW**	**HIGH**	**MEDIUM**	**LOW**
Test characteristics, %						
Accuracy	96.0	95.1	99.1	96.6	95.7	99.1
Sensitivity	100.0	86.8	97.6	100.0	87.2	94.0
Specificity	93.7	98.9	99.7	91.6	98.9	99.9
Positive predictive value	90.2	97.3	99.4	94.6	96.8	99.1
Negative predictive value	100.0	94.2	98.9	100.0	95.3	99.1
Test duration, min						
Mean	6.22	7.78	4.20	5.99	7.63	6.08
SD	1.41	1.26	2.84	1.35	1.23	2.80
Minimum	4.33	5.65	0.61	4.27	5.44	0.45
1st Quartile	5.02	6.59	1.84	4.90	6.62	3.78
Median	5.88	7.93	3.45	5.60	7.42	7.00
3rd Quartile	7.09	8.79	7.19	6.71	8.59	8.67
Maximum	9.85	10.13	9.74	10.08	10.11	9.85
Tests with duration, %						
< 1 min	0.0	0.0	6.7	0.0	0.0	1.8
1 < 2 min	0.0	0.0	22.6	0.0	0.0	9.1
2 < 3 min	0.0	0.0	16.5	0.0	0.0	9.1
3 < 4 min	0.0	0.0	12.2	0.0	0.0	9.1
4 < 5 min	24.3	0.0	9.1	32.5	0.0	9.1
5 < 6 min	29.4	6.0	3.7	29.0	6.4	6.4
6 < 7 min	20.1	30.2	3.7	18.3	28.8	5.5
7 < 8 min	9.8	16.8	7.9	8.7	32.9	15.5
8 < 9 min	12.1	26.8	9.8	7.6	14.2	14.5
≥ 9 min	4.2	20.1	7.9	3.8	17.8	20.0

PVT, psychomotor vigilance task; TSD, total sleep deprivation; PSD, partial sleep deprivation, SD, standard deviation. [a]Performance category is classified based on the number of lapses plus false starts on the full 10-minute PVT.

RESULTS

The likelihood ratios for HIGH and LOW performance groups conditional on the number of lapses and false starts status are shown in Figure 1. Because lapses and false starts are rare events, they carry more information than do stimuli without a lapse or a false start (i.e., likelihood ratios associated with a lapse or a false start are more extreme and therefore lead to a greater change in posterior probability relative to prior probability than do likelihood ratios with no lapse or false start). Figure 1 also illustrates that lapses and false starts during the first 30 seconds of the task were less informative, as they also seem to be more prevalent in the HIGH-performance group during this period relative to the rest of the test. Otherwise, likelihood ratios were relatively stable across the 10 minutes of the test.

Performance of the PVT-A relative to the 10-minute PVT is shown for both the training and the validation data set in Table 1. Overall, there were no relevant differences between training and validation data sets. PVT-A performance was high, with (depending on performance category) 95.1% to 99.1% accuracy, 86.8% to 100% sensitivity, 91.6% to 99.9% specificity, and positive and negative predictive values ranging from 90.2% to 99.4% and 94.2% to 100%, respectively. According to Landis and Koch,[43] chance-corrected agreement was excellent for both the training data set ($\kappa = 0.93$) and the validation data set ($\kappa = 0.92$). Test duration averaged 6.0 minutes (SD 2.4 min, minimum 37 s) for the training data set and 6.4 minutes (SD 1.7 min, minimum 27 s) for the validation data set. Average test duration was longest for MEDIUM-performance bouts relative to HIGH- and LOW-performance bouts (7.63-7.78 min vs 4.20-6.22 min).

Figure 2 illustrates, for each test bout and for both the training and the validation data set, the number of lapses and false starts on the full 10-minute PVT (abscissa), the classification of the test bout according to PVT-A (represented by different symbols), and the duration of PVT-A (ordinate). PVT-A duration was highest for test bouts with the number of lapses and false starts on the 10-minute PVT near the category boundaries of 5 lapses and false starts and 16 lapses and false starts. It decreased with increasing distance from these 2 boundaries. Even for test bouts with no lapse or false start on the 10-minute PVT, PVT-A duration was still longer than 4 minutes, whereas PVT-A duration decreased continuously to values of less than 1 minute with an increasing number of lapses and false starts on the 10-minute PVT. Misclassifications tended to be close to the category boundaries (± 5 lapses and false starts, with the exception of 1 low-performance bout with 25 lapses and false starts that was classified as MEDIUM in PSD).

Figure 3 compares the percentage of test bouts classified as HIGH, MEDIUM, and LOW between the full 10-minute PVT and PVT-A across 33 hours of TSD and across PSD. In general, agreement between the 2 versions of the test was high.

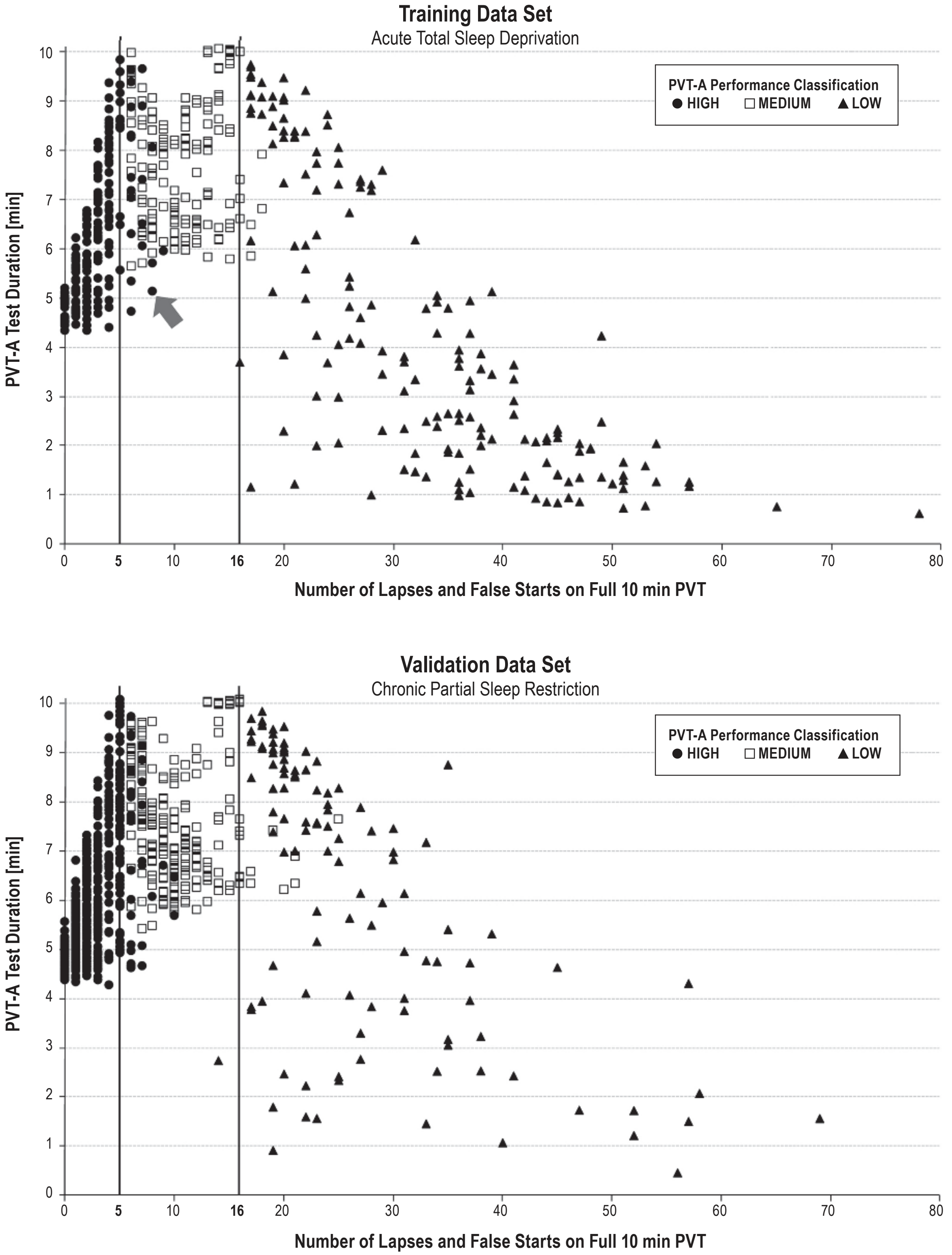

Figure 2—For each test bout and for both the training data set (above, n = 527 test bouts) and the validation data set (below, n = 880 test bouts), the number of lapses and false starts on the full 10-minute psychomotor vigilance task (PVT) (abscissa), the classification of the test bout according to the adaptive-duration PVT (PVT-A, represented by different symbols), and the duration of PVT-A (ordinate) are plotted. The vertical lines represent the category boundaries separating HIGH from MEDIUM (≤ 5 lapses and false starts) and MEDIUM from LOW (≤ 16 lapses and false starts) performance groups based on the full 10-minute PVT. As an example, the arrow points to a test bout with 8 lapses and false starts on the 10-minute PVT that was wrongly classified by PVT-A as a HIGH-performance bout (PVT-A stopped after 5.1 min to reach this decision; 7 of the 8 lapses or false starts occurred after an elapsed time of 5.1 min). PVT-A duration was highest for test bouts with the number of lapses and false starts on the 10-minute PVT near the category boundaries of 5 and 16. Even for test bouts with no lapses or false starts on the 10-minute PVT, PVT-A duration was still longer than 4 minutes, whereas PVT-A duration decreased continuously to values shorter than 1 minute with an increasing number of lapses and false starts on the 10-minute PVT. Misclassifications tended to be close to the category boundaries.

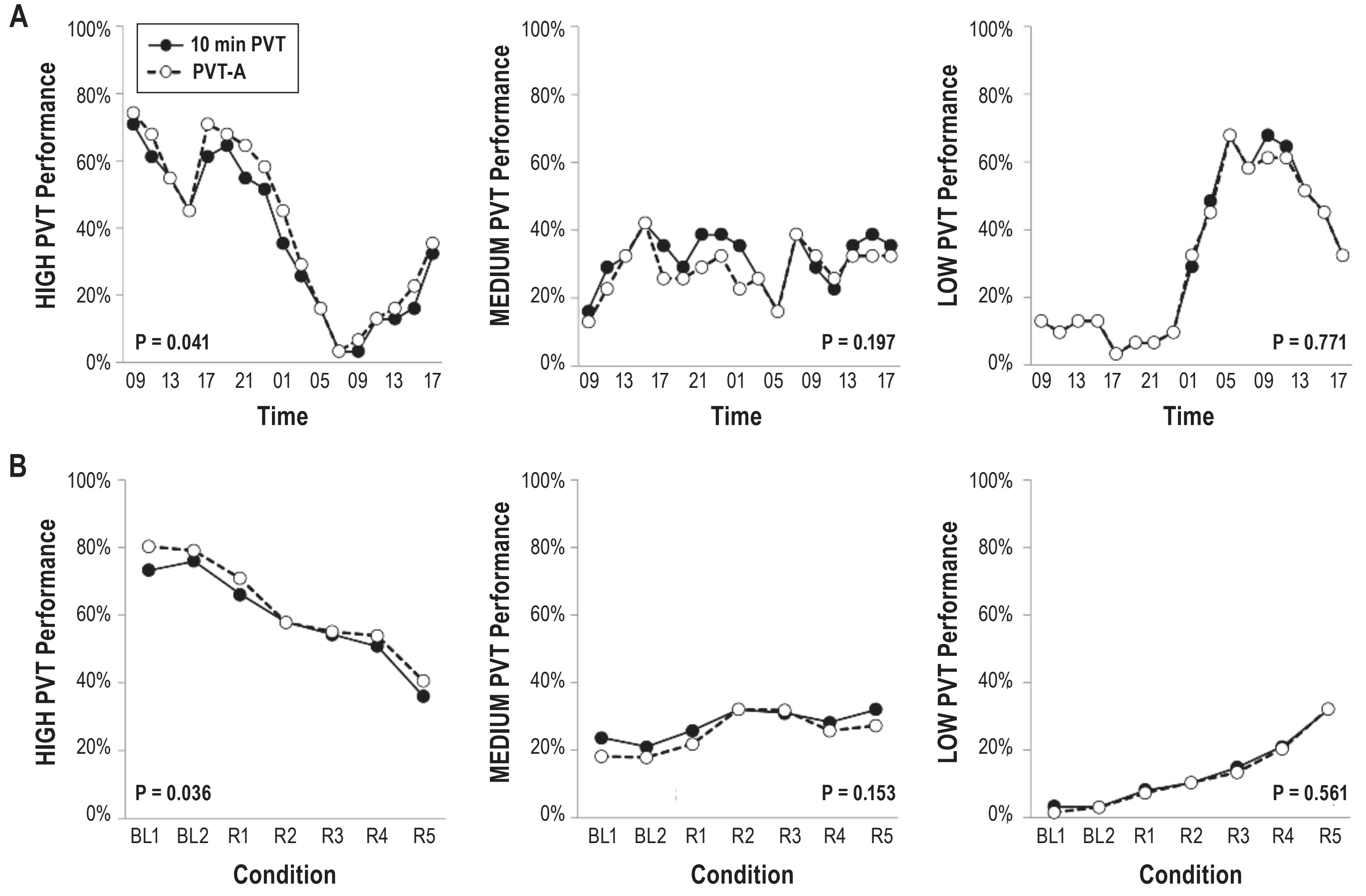

Figure 3—The proportion of performance bouts classified as HIGH, MEDIUM, or LOW is shown for the training data set (**A**, across 33 h of acute total sleep deprivation) and for the validation data set (**B**, chronic partial sleep deprivation with 2 baseline nights BL1 and BL2 and 5 nights of sleep restricted to 4 h time in bed—R1 to R5). Classifications based on the full 10-minute psychomotor vigilance task (PVT) are shown as black circles, whereas classifications based on the adaptive-duration version of the PVT (PVT-A) are shown as open circles. In general, agreement between the 2 versions of the test was high. PVT-A significantly overestimated the percentage of HIGH-performance bouts in both the training and the validation data set (mean difference 3.8%, range 0%-9.7%). The percentage of performance bouts classified as MEDIUM or LOW did not differ significantly between PVT-A and the full 10-minute PVT. P values are based on nonlinear mixed-effects models testing for differences between the 10-minute PVT and PVT-A controlling for experimental condition.

PVT-A significantly overestimated the percentage of HIGH-performance bouts in both the training and the validation data set (mean difference 3.8%, range 0%-9.7%). The percentage of performance bouts classified as MEDIUM or LOW did not differ significantly between PVT-A and the full 10-minute PVT. Furthermore, the percentage of subjects classified as high performers and the percentage of subjects classified as low performers reflected both homeostatic and circadian influences on psychomotor vigilance performance during TSD and continuously increasing impairment during PSD.

DISCUSSION

Using a Bayesian approach and comparable to sequentially applied diagnostic tests, we developed an adaptive-duration version of the PVT. The algorithm was trained with 527 test bouts from 31 subjects undergoing 33 hours of acute TSD and validated with 880 test bouts of 43 subjects being restricted to 5 nights of 4 hours of time in bed. Compared with the full 10-minute PVT, the average duration of the adaptive-duration PVT was markedly reduced to less than 6.5 minutes, with minimum test durations of less than 0.5 minutes. At the same time, PVT-A was highly accurate, sensitive, and specific both in the training and in the validation data set with excellent chance-corrected agreement relative to the full 10-minute PVT.

PVT-A duration was highest near the category boundaries of 5 and 16 lapses and false starts because misclassifications are more likely near the category boundaries and PVT-A has to sample more information to correctly classify the test. For the same reason, misclassifications were usually close to the category boundaries (± 5 lapses and false starts, with 1 exception). Even in test bouts with no lapse or false start on the full 10-minute PVT, PVT-A needed more than 4 minutes to classify the test bout as a high-performance bout. This can be explained by likelihood ratios close to 1 for both LOW and HIGH performance categories for the no lapse or false start condition (see Figure 1). Therefore, many stimuli with no lapse or false start are needed to push P_{HIGH} above the decision threshold. For the same reason, only a few stimuli with lapse or false start are needed to push P_{LOW} above the decision threshold, which is why some of the tests were classified as low-performance bouts in less than 1 minute.

This difference between minimum test durations for HIGH and LOW test bouts makes sense in light of the time-on-task effect. Although it is possible that a subject without a lapse or a false start during the first minutes of the task deteriorates later during the task, a subject with many lapses during the first minutes of the task very likely only deteriorates further with time on task.[20] Subjects deteriorating only late during the task may also explain the comparatively low positive predictive values for the HIGH-performance category (i.e., the observed tendency of PVT-A to misclassify medium-performance bouts as high-performance bouts; ca. 80% of all misclassifications), whereas only 2 medium-performance test bouts were wrongly classified as a low-performance test bout.

Obviously, the choice of the decision threshold (i.e., the posterior probability at which PVT-A stops sampling data) affects both test performance and duration. Based on the data of the training data set, we chose a decision threshold of 99.9826% that led to a correct classification in 95.1% of test bouts and to an average test duration of 6.0 minutes. Another sensible choice for the decision threshold would have been one that led to no misclassifications across 2 categories (i.e., HIGH classified as LOW or LOW classified as HIGH). In the training data set, this decision threshold (99.359%) was associated with an average test duration of 4.3 minutes and 86.1% correct decisions.

For the PVT-A algorithm, we assumed that consecutive tests (each based on the response to 1 stimulus) were conditionally independent given the performance status. We checked this assumption by calculating likelihood ratios conditional on the outcome of 2 consecutive responses and comparing them to the product of 2 single-response likelihood ratios. In the case of conditional independence, these likelihood ratios should be equal. We found that the product of 2 single response likelihood ratios only slightly overestimated likelihood ratios greater than 1 and only slightly underestimated likelihood ratios less than 1 (i.e., were more extreme) compared with likelihood ratios based on the outcome of 2 consecutive responses, confirming that conditional independence was a reasonable assumption. Because 2 consecutive stimuli with a lapse or a false start were extremely rare events in the high-performance group, it was impossible to calculate likelihood ratios for 15 out of the 20 time-on-task periods for the high-performance group, which also prevented us from using an algorithm based on the outcome of 2 (or more) consecutive stimuli.

The fact that 9.3% of test bouts were classified as low-performance bouts by both PVT and PVT-A during the first 16 hours of wakefulness in the TSD protocol warrants further discussion. Noncompliance or prior sleep debt are possible explanations. We asked subjects to adhere to a regular sleep schedule in the week prior to the start of the study, but actigraphy and time-stamped telephone logs showed that adherence to these instructions was not always perfect. In the TSD protocol, 1 night (n = 22 subjects) or 2 nights (n = 9 subjects) with an 8-hour sleep opportunity may simply not have been enough to counter any preexisting sleep debt. Also, compliance in the TSD protocol seems to have been somewhat lower than in the PSD protocol. In the TSD protocol, data from 4 out of 36 subjects were excluded from the analysis due to noncompliance, excessive fatigue, or both during the first 16 hours of wakefulness. In the PSD protocol, only 3 out of 47 subjects were excluded for these reasons. Overall, only 3.1% (PVT) and 2.3% (PVT-A) of the test bouts were classified as low-performance bouts during baseline days 1 and 2 of the PSD protocol, after 1 or 2 nights with 10-hour sleep opportunities, respectively.

Limitations

Several limitations have to be taken into account when interpreting the results of these analyses. First, categorizing the continuous outcome metric number of lapses and false starts on the 10-minute PVT into 3 discrete outcomes (groups with high, medium, and low performance) for the PVT-A constitutes a data reduction and, therefore, a loss of information. However, as explained above, we believe that it will be sufficient for many applied contexts to know whether a test bout qualifies as a high-, medium-, or low-performance test bout, and that the absolute number of lapses would not add relevant information. Also, it would be easy to report the number of lapses and false starts and other common PVT outcome metrics at PVT-A termination and the projected number of lapses and false starts for the full 10-minute PVT in addition to the performance category, if this was desired by the end user. On the same note, it would be possible to introduce more than 3 performance categories.

Second, the 10-minute PVT is a work-paced task (i.e., the behavior of the tested individual does not influence task duration). In contrast, PVT-A termination will depend on the response behavior of the tested subject. Test duration will be short in test bouts with either a very low or a very high number of lapses and false starts. Although the first may not be problematic because it is impossible to fake high performance, the latter is a more severe threat to the validity of the test because noncompliant or poorly motivated subjects may choose to lapse frequently or bias toward false starts to stop the test early. This is unlikely to happen in fit-for-duty contexts (as the subjects are usually highly motivated to achieve high-performance levels). However, in cases of repeated low-performance levels, the response data should be checked for plausibility.

Third, in our analyses, the PVT-A algorithm was applied posthoc to data collected with the standard 10-minute PVT. However, we do not see major obstacles in implementing the PVT-A algorithm into our current PVT to achieve online, real-time analysis.

Fourth, the investigated subjects were healthy, had a restricted age range, and were investigated in a controlled laboratory environment. The results may therefore not generalize to non-healthy, older or younger groups of subjects and to operational environments.

Finally, it has to be stressed that performance impairment on the PVT-A indicates reduced vigilant attention due to fatigue or other reasons. Because vigilant attention is instrumental for many cognitive and more complex tasks, it is likely that these will also be affected if vigilant attention is low (as shown for a 3-minute PVT for a simulated luggage-screening task[37]). However, it is unknown how PVT-A specifically relates to many other tasks and how well it predicts performance on these tasks.

CONCLUSIONS

We developed and validated a highly accurate, sensitive, and specific adaptive-duration version of the 10-minute PVT. Test duration of the adaptive PVT averaged less than 6.5 minutes, with some tests terminating after less than a minute, increasing the practicability of the test in operational and clinical settings. The adaptive-duration strategy may be superior to a simple reduction of PVT duration in which the fixed test duration may be too short to identify subjects with moderate impairment (showing deficits only later during the test) but unnecessarily long for those who are either fully alert or severely impaired.

ACKNOWLEDGMENTS

This investigation was sponsored by the Human Factors Program of the Transportation Security Laboratory, Science and Technology Directorate, US 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. The authors 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 the journal *SLEEP*. Dr. Basner has received compensation for consulting from 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 Associated Professional Sleep Societies, LLC, 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. Basner M, Rubinstein J, Fomberstein KM, et al. Effects of night work, sleep loss and time on task on simulated threat detection performance. Sleep 2008;31:1251-9.
3. Barger LK, Cade BE, Ayas NT, et al. Extended work shifts and the risk of motor vehicle crashes among interns. N Engl J Med 2005;352:125-34.
4. Dinges DF. An overview of sleepiness and accidents. J Sleep Res 1995;4:4-14.
5. Finkel KJ, Searleman AC, Tymkew H, et al. Prevalence of undiagnosed obstructive sleep apnea among adult surgical patients in an academic medical center. Sleep Med 2009;10:753-8.
6. Basner M, Müller U, Elmenhorst E-M. Single and combined effects of air, road, and rail traffic noise on sleep and recuperation. Sleep 2011;34:11-23.
7. Driscoll TR, Grunstein RR, Rogers NL. A systematic review of the neurobehavioural and physiological effects of shiftwork systems. Sleep Med Rev 2007;11:179-94.
8. Basner M, Dinges DF. Dubious bargain: trading sleep for Leno and Letterman. Sleep 2009;32:747-52.
9. Basner M, Fomberstein KM, Razavi FM, et al. American time use survey: sleep time and its relationship to waking activities. Sleep 2007;30:1085-95.
10. 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.
11. Van Dongen HP, Belenky G, Vila BJ. The efficacy of a restart break for recycling with optimal performance depends critically on circadian timing. Sleep 2011;34:917-29.
12. Zhou X, Ferguson SA, Matthews RW, et al. Mismatch between subjective alertness and objective performance under sleep restriction is greatest during the biological night. J Sleep Res 2011 May 13. doi: 10.1111/j.1365-2869.2011.00924.x.
13. Goel N, Rao H, Durmer JS, Dinges DF. Neurocognitive consequences of sleep deprivation. Semin Neurol 2009;29:320-39.
14. Lim J, Dinges DF. A meta-analysis of the impact of short-term sleep deprivation on cognitive variables. Psychol Bull 2010;136:375-89.
15. Lim J, Dinges DF. Sleep deprivation and vigilant attention. Molecular and Biophysical Mechanisms of Arousal, Alertness, and Attention. Annals of the New York Academy of Sciences. Oxford: Blackwell Publishing; 2008:305-22.
16. 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.
17. 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.
18. Dinges DF, Kribbs NB. Performing while sleepy: Effects of experimentally-induced sleepiness. In: Monk TH, ed. Sleep, Sleepiness and Performance. Chichester, UK: John Wiley and Sons, Ltd; 1991:97-128.
19. Warm JS, Parasuraman R, Matthews G. Vigilance requires hard mental work and is stressful. Hum Factors 2008;50:433-41.
20. Basner M, Dinges DF. Maximizing sensitivity of the Psychomotor Vigilance Test (PVT) to sleep loss. Sleep 2011;34:581-91.
21. Doran, S.M., Van Dongen, H.P., Dinges, D.F.: Sustained attention performance during sleep deprivation: evidence of state instability. Archives of Italian Biology: Neuroscience 139 (3):253-267, 2001.
22. 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, DC: American Psychological Association; 2010:83-101.
23. 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.
24. 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.
25. Tomasi D, Wang RL, Telang F, et al. Impairment of attentional networks after 1 night of sleep deprivation. Cereb Cortex 2009;19:233-40.
26. 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.
27. Philip P, Akerstedt T. Transport and industrial safety, how are they affected by sleepiness and sleep restriction? Sleep Med Rev 2006;10:347-56.
28. Van Dongen HP, Dinges DF. Sleep, circadian rhythms, and psychomotor vigilance. Clin Sports Med 2005;24:237-49, vii-viii.
29. 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. Austin, TX: 30th Annual Meeting of the Cognitive Science Society, 2008:2017-22.
30. 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.
31. 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.
32. 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.
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;40:347-52.
34. 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.
35. Basner M, Mollicone DJ, Dinges DF. Validity and sensitivity of a brief Psychomotor Vigilance Test (PVT-B) to total and partial sleep deprivation. Acta Astronaut 2011; 69: 949-59.

36. 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.
37. Basner M, Rubinstein J. Fitness for duty: a 3 minute version of the Psychomotor Vigilance Test predicts fatigue related declines in luggage screening performance. J Occup Environ Med 2011;53:1146-54.
38. Dinges DF, Maislin G, Brewster RM, Krueger GP, Carroll RJ. Pilot test of fatigue management technologies. Transportation Research Record: J Transport Res Board 2005;1922:175-82.
39. 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.
40. Basner M, Griefahn B, Müller U, Plath G, Samel A. An ECG-based algorithm for the automatic identification of autonomic activations associated with cortical arousal. Sleep 2007;30:1349-61.
41. Hunink M, Glasziou P, Siegel J, Elstein A, Weinstein M. Decision Making in Health and Medicine: Integrating Evidence and Values. Cambridge, UK: University Press; 2001.
42. Fleiss J, Levin B, Paik MC. Statistical Methods for Rates and Proportions. Hoboken, NJ: John Wiley & Sons, Inc; 2003.
43. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-74.