Sleep Advances, 2022, **3**, 1–10
https://doi.org/10.1093/sleepadvances/zpac038
Advance access publication 20 October 2022
Original Article

Original Article

Ultra-short objective alertness assessment: an adaptive duration version of the 3 minute PVT (PVT-BA) accurately tracks changes in psychomotor vigilance induced by sleep restriction

Mathias Basner*

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

*Corresponding author. Mathias Basner, Unit for Experimental Psychiatry, Division of Sleep and Chronobiology, Department of Psychiatry, University of Pennsylvania Perelman School of Medicine, 1019 Blockley Hall, 423 Guardian Drive, Philadelphia, PA 19104-6021, USA. Email: basner@pennmedicine.upenn.edu.

Abstract

Study Objectives: The Psychomotor Vigilance Test (PVT) is a widely used and sensitive assay of the cognitive deficits associated with sleep loss and circadian misalignment. As even shorter versions of the PVT are often considered too long, I developed and validated an adaptive duration version of the 3 min PVT (PVT-BA).

Methods: The PVT-BA algorithm was trained on data from 31 subjects participating in a total sleep deprivation protocol and validated in 43 subjects undergoing 5 days of partial sleep restriction under controlled laboratory conditions. With each subject response, the algorithm updated the odds of the test being high, medium or low performance based on lapses plus false starts on the full 3 min PVT-B.

Results: With a decision threshold of 99.619%, PVT-BA classified 95.1% of training data tests correctly without incorrect classifications across two performance categories (i.e. high as low or low as high) and resulted in an average test duration of 1 min 43 s with a minimum duration of 16.4 s. Agreement corrected for chance between PVT-B and PVT-BA was "almost perfect" for both the training (kappa = 0.92) and validation data (kappa = 0.85). Across the three performance categories and data sets, sensitivity averaged 92.2% (range 74.9–100%) and specificity averaged 96.0% (range 88.3–99.2%).

Conclusions: PVT-BA is an accurate adaptive version of PVT-B and, to my knowledge, the shortest version to date that maintains key properties of the standard 10 min duration PVT. PVT-BA will facilitate the use of the PVT in settings in which it was previously considered impractical.

Key words: circadian rhythms; cognitive function; mathematical modeling; neurobehavioral performance; sleep/wake cognition; sleep deprivation; sleepiness

Statement of Significance

The Psychomotor Vigilance Test (PVT) is the de-facto gold standard for assessing the adverse cognitive effects of sleep loss and circadian misalignment. However, it is considered too long for many applied, operational or clinical settings. Here I developed an adaptive duration version of the 3 min PVT (PVT-BA) and show that it maintains sensitivity and specificity relative to both the 3 min PVT (PVT-B) and the standard 10 min PVT. Average test duration decreased from 3 min to 1 min 43 s, some tests being shorter than 20 s. The short duration of PVT-BA paired with its high accuracy will facilitate the use of the PVT in settings in which it was previously considered too long and impractical.

Introduction

The Psychomotor Vigilance Test (PVT), invented in 1985 by David F. Dinges and John W. Powell [1], is arguably the most widely used assay in research on the cognitive effects of sleep loss and circadian misalignment [2]. The standard PVT records response times to visual stimuli that occur at random 2–10 s inter-stimulus intervals (ISI) over a 10 min period [1–5]. Sleep deprivation and circadian misalignment induce reliable changes in PVT

Submitted for publication: June 2, 2022; **Revised:** October 11, 2022

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 errors of commission (i.e. responses without a stimulus, or false starts) [6, 7]. These effects are associated with changes in neural activity in distributed brain regions [8–12].

The PVT has several attributes that explain its popularity in sleep and circadian research. In contrast to more complex cognitive tests, the PVT has little to no aptitude and practice effects that may otherwise mask sleep deprivation effects and require careful study designs with adequate control groups [13, 14]. Degradation of attention is among the most reliable effects of sleep deprivation, which explains the high sensitivity of the PVT [15–17] and that it is often used as a de-facto gold standard for assessing the cognitive effects of sleep deprivation [18]. PVT performance also has ecological validity relative to real-world risks, as deficits in sustained attention and timely reactions adversely affect many applied tasks (e.g. industrial tasks, transportation, security-related tasks) [19–22]. Finally, sleep deprived subjects are unable to reliably assess their degree of impairment—especially in chronic exposure situations and at the circadian low—stressing the need for objective measures of cognitive performance deficits induced by sleep loss [7, 23, 24].

While reliable and portable assessments of alcohol intoxication have been available to law enforcement for years, a similarly sensitive and specific assay of sleep loss is still missing and, given the complexity of the biological processes involved, may not be available in the near future. In the meantime, objective behavioral assays could be used to assess the degree of sleep loss-induced impairment, and—due to its favorable properties outlined above—the PVT is currently the best candidate for such an assay [25]. The PVT could also be a valuable fit-for-duty or readiness-to-perform tool in workplace settings [26].

However, the 10 min standard duration of the PVT is regarded by many as too long for applied, operational, or clinical settings. For this reason, shorter versions with durations of 3 min and 5 min have been developed [27–32]. These track 10 min PVT performance closely [33], but typically lose some sensitivity relative to the standard 10 min PVT, likely due to the fact that performance on the PVT deteriorates with time-on-task (so-called vigilance decrement), faster in sleep deprived than in alert subjects [2, 34]. 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 who deteriorate only later during the test, while, on the other hand, they may be unnecessarily long for other subjects who are apparently fully alert or severely impaired.

This prompted me to develop an adaptive duration version of the 10-min PVT (PVT-A) [35]. After each response and depending on the nature of the response, the PVT-A algorithm re-evaluates the probability of a subject being a HIGH, MEDIUM, or LOW performer on the full 10-min PVT. If a pre-defined decision threshold is exceeded, PVT-A stops sampling data, as it has determined to have gathered enough information to make a correct performance categorization.

However, despite the fact that individual tests can last less than 30 s, administration time on PVT-A still averaged six minutes, which is twice as long as a 3 min brief non-adaptive version of the PVT (PVT-B) that we previously developed and validated [35]. Furthermore, based on feedback from colleagues who considered using PVT-B, many even considered 3 min too long for applied, operational, or clinical settings. This prompted me to develop an adaptive version of PVT-B, abbreviated PVT-BA, with the goal to provide an ultra-short assay of vigilant attention that nevertheless retains sensitivity to acute total sleep deprivation and chronic partial sleep restriction and closely tracks both the full 3 min PVT-B and the standard 10 min 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 [2, 35].

Acute TSD protocol.

TSD data were gathered on $N = 36$ 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 five consecutive days, which included a 33 h period of TSD. Four subjects were excluded from the analysis due to non-compliance and/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 ± *SD* 31.1 ± 7.3 years; 18 female; 51.6% black, 32.3% white, 16.1% other race) contributed to the analyses presented here. The study started at 8 am on day 1 and ended at 8 am on day 5. A 33 h period of total sleep deprivation started either on day 2 ($N = 22$) or on day 3 ($N = 9$) of the study. Except for the sleep deprivation period, subjects had 8 h sleep opportunities between 12 pm and 8 am. The first sleep period was monitored polysomnographically to exclude possible sleep disorders.

Chronic partial sleep deprivation (PSD) protocol.

PSD data were obtained from $N = 47$ healthy adults in a laboratory protocol involving 5 consecutive nights of sleep restricted to 4 h per night (4 am to 8 am period) following two baseline nights with 10 h time in bed each. Three subjects were excluded from the analysis due to non-compliance and/or excessive fatigue during baseline data collection. One additional subject had no valid baseline data. Therefore, $N = 43$ subjects (16 females; 69.8% black, 25.6% white, 4.6% other race) who averaged 30.5 ± 7.3 years (mean ± *SD*) contributed to the analyses presented here. A detailed description of the experimental procedures is published elsewhere [36].

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

In both TSD and PSD experiments subjects completed 30 min bouts of a neurobehavioral test battery that included the 10 min

PVT and the 3 min PVT-B every 2 h during scheduled wakefulness. In the TSD experiment, each subject performed 17 PVTs in total (starting at 9 am after a sleep opportunity from midnight to 8 am with bout #1 and ending at 5 pm 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 [2, 32], I only used the test bouts administered at 12:00, 16:00, and 20:00 on baseline days 1 and 2 and days after restriction nights 1–5 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/board games and interact with laboratory staff to help them stay awake, but no naps/sleep or vigorous activities (e.g. exercise) were allowed.

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

PVT

We utilized a precise computer-based version of the 10 min PVT, that was performed and analyzed according to the standards set forward in Basner and Dinges [2]. 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 response time (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 2–10 s (10 min PVT) and 1–4 s (PVT-B, both 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 on the CRT screen (10 min PVT) or LED display (PVT-B), 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 RT. In the TSD protocol, PVT-B was performed on average 22.9 min before (N = 136) or immediately after (N = 391) the 10 min PVT. In the PSD protocol, PVT-B was performed on average 16 min after the 10 min PVT in all subjects. In a previous analysis of the TSD data it was shown that test administration order had no statistically significant effect on PVT RTs [32].

PVT-BA algorithm

Outcome metric and performance group classification.

I decided to use the sum of the number of lapses (i.e. errors of omission, defined as a RT ≥500 ms on the 10 min PVT and ≥355 ms on PVT-B) plus the number of false starts (i.e. errors of commission, defined as responses without stimulus or responses with RTs <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 [2] found that the number of lapses and false starts (LpFS) 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 (the highest effect size in the PSD protocol was achieved by response speed with 1.21). Also, taking false starts into account may help to identify non-compliant subjects or those who try to prevent lapses by biasing toward false starts, which may be especially important in fit-for-duty contexts. Essentially, by accounting for both false starts and lapses, subjects performing the PVT consistently have to deliver RTs within a tight window (above the coincident false start threshold and below the lapse threshold) in order to qualify as high performers. Due to the lower lapse threshold, this is even harder on PVT-B compared to the standard 10 min PVT. The lower lapse threshold was introduced on PVT-B to increase its sensitivity and comparability to the 10 min PVT despite its shorter duration [32].

The TSD study was used to find two LpFS thresholds that divided PVT-B test bouts into HIGH, MEDIUM, and LOW performance bouts. We argued earlier [25] that one 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, I rather chose to divide the dataset into three 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 was shown in a study of truck drivers [37]. Similar arguments could be brought forward for diagnostic paradigms, where two groups indicating "not impaired" and "impaired" may not be sufficient.

First, ≤6 LpFS was identified as the threshold that optimally differentiated test bouts performed until 21:00 (up to 13 h awake) from test bouts performed at or after 23:00 (15–33 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 h of wakefulness [7]. This threshold is only one LpFS higher than the threshold identified for PVT-A [35]. A second threshold of 16 LpFS was identified by performing a median split on all test bouts with more than 6 LpFS. This threshold is identical to the one identified for PVT-A [35]. Therefore, the three performance groups were defined as follows: HIGH (≤6 LpFS, N = 220 bouts), MEDIUM (>6 and ≤16 LpFS, N = 158 bouts), and LOW (>16 LpFS, N = 149 bouts).

PVT-BA algorithm description.

Similar to an earlier published approach [38], each RT 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, while the probability of being a LOW performance test bout increases at the same time. While one may assign equal probabilities to the three 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 RT 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 one of the three probabilities P_{HIGH}, P_{MEDIUM}, or P_{LOW} exceeds a pre-defined 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):

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

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):

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

The LR depends on the RT outcome of the stimulus. Relative to the prior probability, the posterior probability will increase for LRs >1, decrease for LRs <1, and remain unchanged for LRs = 1. I only calculated LRs and posterior probabilities for P_{HIGH} and P_{LOW}. P_{MEDIUM} was then calculated according to equation (3):

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

The 4 likelihood ratios LR(High | no LpFS), LR(High | LpFS), LR(Low | no LpFS), and LR(Low | LpFS) were calculated based on the TSD data (for a detailed description of LR calculations see Hunink et al. [39]) . To acknowledge time on task effects, I divided the 3 min task duration into six 30 s intervals and calculated LRs for each interval (see Figure 1).

I 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 to MEDIUM and LOW performance bouts in the TSD study. I believe this better reflects the uncertainty of each individual test outcome. In addition to the basic algorithm described above, after each stimulus I checked for the following conditions: If LpFS exceeded 6, P_{HIGH} was set to zero, and the probabilities for belonging to the MEDIUM and LOW performance group were adjusted accordingly:

$$P_{MED} = P_{MED}/(P_{MED} + P_{LOW}). \quad (4)$$

$$P_{LOW} = P_{LOW}/(P_{MED} + P_{LOW}). \quad (5)$$

If LpFS 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 3 min of the PVT-B had elapsed, the test was classified according to the actual LpFS (i.e. always correct), and the full 3 min were recorded for PVT-BA duration (4.7% of TSD tests and 3.6% of PSD tests did not exceed the decision threshold before the full 3 min of PVT-B elapsed, respectively).

The decision threshold was set to 99.619% (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) >95% of the 527 decisions were correct, and there were no misclassifications across two categories (i.e. HIGH classified as LOW or LOW classified as HIGH). For each test, I recorded:

(a) The true test classification based on the result of the full 3-min PVT-B;
(b) The classification of the PVT-BA algorithm; and
(c) The time the PVT-BA algorithm needed to reach that decision.

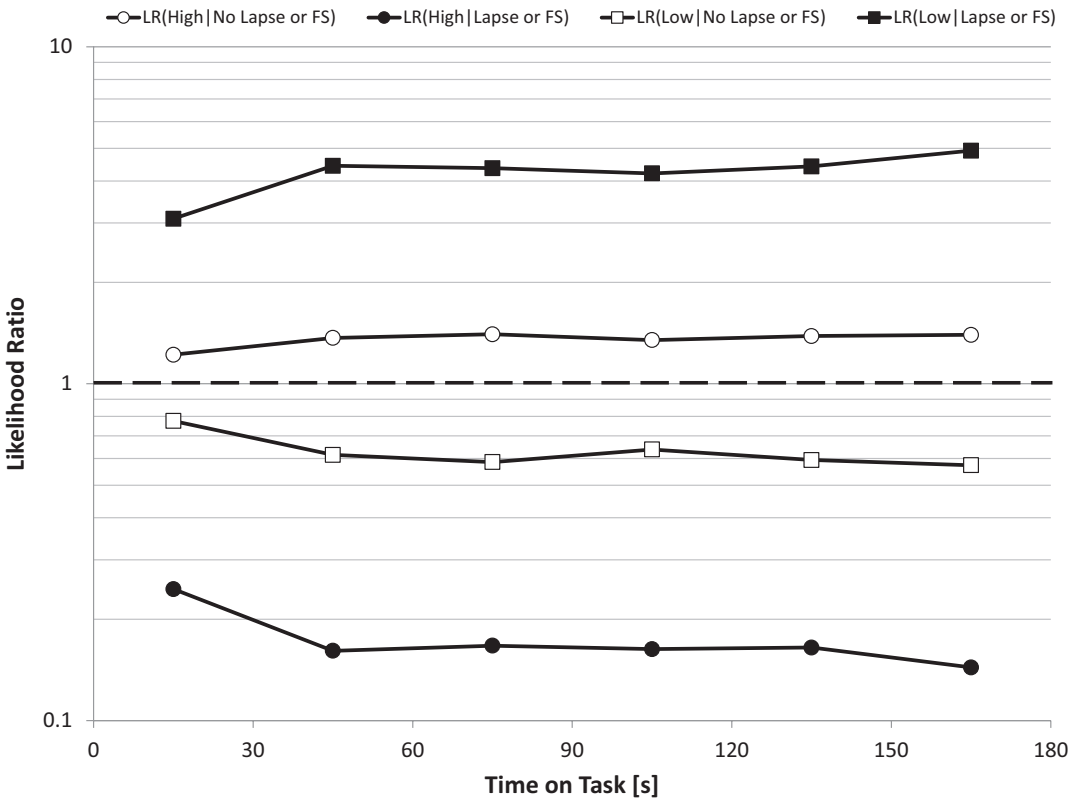

Figure 1. Likelihood ratios (LR) for belonging to the high PVT performance group (HIGH) or to the low PVT performance group (LOW) are given conditional on whether or not a lapse or a false start occurred and depending on time on task. Posterior probability for belonging to a specific PVT performance group will increase for LRs >1, decrease for LRs <1, and remain unchanged for LRs = 1. As lapses and false starts are rare events, they carry more information than stimuli without lapses and false starts (i.e. LRs associated with lapses and false starts are more extreme and therefore lead to a greater change in posterior probability than LRs with no lapses or false starts). LRs were relatively stable across the 3 min of the test, but were closer to 1 during the first 30 s of the test.

Table 1. Test characteristics and duration of the Adaptive Brief Duration PVT (PVT-BA) relative to the full 3 min Brief Duration PVT (PVT-B) for the training data set and the validation data set

	Training Data Set (TSD)			Validation Data Set (PSD)		
Performance category[1]	**HIGH**	**MEDIUM**	**LOW**	**HIGH**	**MEDIUM**	**LOW**
Test characteristics						
Accuracy [%]	97.5	95.1	97.5	93.8	90.5	96.5
Sensitivity [%]	100.0	85.4	98.0	100.0	74.9	94.7
Specificity [%]	95.8	99.2	97.4	88.3	98.6	96.9
Positive predictive value [%]	94.4	97.8	93.6	88.1	96.6	87.9
Negative predictive value [%]	100.0	94.1	99.2	100.0	88.2	98.7
Test duration						
Average [s]	97.8	143.6	75.7	95.2	140.9	77.0
Standard deviation [s]	31.7	23.2	45.2	30.5	24.3	43.7
Minimum [s]	59.4	102.9	16.4	57.0	89.2	16.7
1st Quartile [s]	72.4	125.0	40.9	70.1	121.3	42.3
Median [s]	86.2	137.6	61.0	85.7	136.4	60.9
3rd Quartile [s]	118.8	162.7	99.4	110.1	160.8	109.4
Maximum [s]	181.8	181.9	175.2	180.8	182.7	178.3
<30 s [%]	0.0	0.0	7.1	0.0	0.0	7.1
30 < 60 s [%]	0.4	0.0	42.3	0.4	0.0	42.3
60 < 90 s [%]	57.5	0.0	23.1	55.9	0.4	16.5
90 < 120 s [%]	17.6	13.0	7.1	22.7	22.1	13.7
120 < 150 s [%]	14.2	50.0	7.7	12.5	40.4	11.5
≥ 150 s [%]	10.3	37.0	12.8	8.4	37.0	8.8

TSD, total sleep deprivation; PSD, partial sleep deprivation.
[1]Performance category is classified based on the number of lapses plus false starts on the full 3 min PVT-B.

The PVT-BA 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. 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, I calculated the following test performance metrics always relative to the remaining two categories (TP = true positive; TN = true negative; FP = false positive; FN = false negative):

- Accuracy = (TP + TN)/ (TP + TN + FP + FN).
- Sensitivity = TP/ (TP + FN).
- Specificity = TN/ (TN + FP).
- Positive Predictive Value = TP/ (TP + FP).
- Negative Predictive Value = TN/ (TN + FN).

Furthermore, I calculated accuracy for each individual to investigate individual differences in PVT-BA performance. I also calculated average, standard deviation, minimum, maximum and quartiles for PVT-BA test duration and for each category. Finally, I calculated kappa (a chance-corrected measurement of agreement) across the three performance categories [40]. With a non-linear mixed effects model controlling for experimental condition (time of day for the TSD protocol and study day for the PSD protocol) I investigated whether HIGH, MEDIUM, and LOW classifications differed significantly between the PVT-BA, PVT-B and standard 10 min PVT across 33 h of TSD and across the 7 nights of the PSD protocol, and if so by how much (Proc NLMIXED, SAS, Cary, NC, Version 9.4).

Results

The LRs for HIGH and LOW performance groups conditional on whether a response was classified as LpFS and depending on time-on-task are shown in Figure 1. A lapse or false start increased the likelihood of being a LOW performer 3- to 5-fold, while it lowered the likelihood of being a HIGH performer by 75–85%. In contrast, responses neither classified as a lapse nor as a false start decreased the likelihood of being a LOW performer only by 22–43% and increased the likelihood of being a HIGH performer only by 22–40%. Figure 1 also illustrates that lapses and false starts during the first 30 s of the task are 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, LRs were relatively stable across the 3 min of the test.

Performance of PVT-BA relative to PVT-B is shown for both the training and the validation data set in Table 1. Overall, differences in PVT-BA performance between training and validation data sets were modest. PVT-BA performance was high, with (depending on performance category) 90.5–95.5% accuracy, 74.9–100% sensitivity, 88.3–99.2% specificity, and positive and negative predictive values ranging from 87.9–97.8% to 88.2–100%, respectively. According to Landis and Koch [41], chance-corrected agreement was “almost perfect” for the training data set (kappa = 0.92) and

the validation data set (kappa = 0.85). Test duration averaged 103.2 s (57.2% of full duration; SD 43.0 s; minimum 16.4 s) for the training data set and 103.6 s (57.4% of full duration; SD 39.9 s, minimum 16.7 s) for the validation data set. Average test duration decreased in the order MEDIUM (140.9–143.6 s; most tests 120 < 150 s), HIGH (95.2–97.8 s; most tests 60 < 90 s) and LOW (75.7–77.0 s; most tests 30 < 60 s) performance bouts.

Of PVT-BA accuracy scores calculated individually for the 74 participants across TSD and PSD protocols, 31.1% were 100%; 43.2% were ≥90% and <100%; 18.9% were ≥80% and <90%; and 6.8% were ≥65% and <80%, respectively. Further analyses showed that percent PVT-B classified as MEDIUM performance was highly negatively correlated with PVT-BA accuracy (r = –0.72, p < .001), while percent PVT-B classified as HIGH performance was moderately positively correlated with PVT-BA accuracy (r = 0.43, p < .001) and percent PVT-B classified as LOW performance was not correlated with PVT-BA accuracy (r = 0.08, p = .49). For individuals with PVT-BA accuracy scores of 100%, 13.1% of PVT-B were classified as MEDIUM performance; for those with PVT-BA accuracy scores ≥90% and <100%, 33.3% of PVT-B were classified as MEDIUM performance; for those with PVT-BA accuracy scores ≥80% and <90%, 49.8% of PVT-B were classified as MEDIUM performance; and for those with PVT-BA accuracy scores ≥65% and <80%, 67.3% of PVT-B were classified as MEDIUM performance, respectively.

Category boundaries for PVT-BA (6 and 16 lapses and false starts) were similar to those for PVT-A (5 and 16 lapses and false starts) [35]. Figure 2 illustrates, for each test bout and for both the training and the validation data set, LpFS on the full 3-min PVT-B (abscissa), the classification of the test bout according to PVT-BA (represented by different symbols), and the duration of PVT-BA (ordinate). PVT-BA duration was highest for test bouts with LpFS on the 3-min PVT-B near the category boundaries 6 and 16 LpFS. It decreased with increasing distance from these two boundaries. Even for test bouts with no lapse or false start on PVT-B, PVT-BA duration was still 60 s or longer, whereas PVT-BA duration decreased continuously to values below 30 s with LpFS increasing on the PVT-B. With few exceptions, misclassifications tended to be close to the category boundaries (Figure 3).

Figure 4 compares the percentage of test bouts classified as HIGH, MEDIUM, and LOW between the full 10-min PVT, PVT-B and PVT-BA across 33 h of total sleep deprivation and during partial sleep restriction. All three versions of the test showed the characteristic decline in vigilant attention after 16 h awake and improving performance after one night without sleep during late morning and afternoon hours (i.e. circadian rescue) in the total sleep deprivation protocol and the characteristic continuous decline in performance across days of sleep restriction in the partial sleep restriction protocol.

In general, agreement between the three versions of the test was high. In the training data set, PVT-BA did not differ from either the PVT-B or the full 10-min PVT in any of the performance categories (all p > .09). In the validation data set, PVT-BA significantly overestimated HIGH performance bouts relative to the PVT-B (mean difference +6.2%; p = .004) and significantly underestimated HIGH performance bouts relative to the 10 min PVT (mean difference –6.7%; p < .001). PVT-BA significantly underestimated MEDIUM performance bouts relative to the PVT-B (mean difference –7.7%; p < .001) but did not differ from the 10 min PVT (mean difference –0.8%; p = .366). PVT-BA significantly overestimated LOW performance bouts relative to the 10 min PVT (mean difference +7.1%; p < .001) but did not differ from PVT-B (mean difference +0.8%; p = .391).

DISCUSSION

Using a Bayesian approach and comparable to sequentially applied diagnostic tests, I developed an adaptive duration version of the 3 min PVT-B, similar to an earlier study that created an adaptive version of the standard 10 min PVT [35]. The algorithm was trained with 527 test bouts of 31 subjects undergoing 33 h of acute total sleep deprivation and validated with 880 test bouts of 43 subjects being restricted to 5 nights of 4 h time in bed. Compared to PVT-B, average duration of PVT-BA was markedly reduced to 1 min 43 s (less than 60% of the original duration), with minimum test durations below 20 s. At the same time, PVT-BA was highly accurate, sensitive, and specific with "almost perfect" chance-corrected agreement relative to PVT-B. Test performance was only slightly reduced in the validation data set relative to the training data set.

PVT-BA category boundaries (6 and 16 lapses) were close to PVT-A category boundaries (5 and 16 lapses), which is not surprising as development of PVT-B was based on the same two data sets used here [32]. PVT-BA duration was highest near the category boundaries as misclassifications were more likely near the category boundaries where PVT-BA had to sample more information to correctly classify the test. Even in test bouts with no lapses and false starts on PVT-B, PVT-BA needed 57 s or more to classify the test bout as a HIGH performance bout. This can be explained by LRs close to 1 both for LOW and HIGH performance categories for the no LpFS condition (see Figure 1). Therefore, many stimuli with no lapses and false starts are needed to push P_{HIGH} above the decision threshold. For the same reason, only a few stimuli with lapses and false starts 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 20 s. This difference between minimum test durations for HIGH and LOW test bouts makes sense in light of the time-on-task effect. While it is possible that a subject without LpFS 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 [2].

Astonishingly, not a single of the 628 HIGH performance tests on PVT-B was misclassified by PVT-BA (Figure 3). Misclassifications usually occurred close to the category boundaries with one notable exception: a single test of the validation data set performed at 9:04 pm on the day following the 4th sleep restriction night with 21 LpFS on PVT-B was wrongly classified as HIGH performance on PVT-BA. On closer inspection of the data, PVT-BA made this wrong classification 84.6 s into the test. Within this period, the participant committed a single lapse 6.6 s into the test. The subject subsequently committed 20 additional lapses after PVT-BA had classified the test as HIGH performance. This example illustrates a potential weakness of PVT-BA, as its decisions are necessarily based on responses early during the test and thus may miss performance decrements towards the end of the test unmasked by time-on-task. However, the data still indicate that PVT-BA misclassifications across two categories are very rare events (1 out of 946 HIGH and LOW performance tests across both training and validation data sets). Also, a closer look at the above mentioned only test with a misclassification across two categories showed that 13 out of the 20 lapses (65%) committed after PVT-BA had already made its decision fell within the traditional 500 ms lapse threshold, and that the longest RT was modest with 1433 ms. This likely explains why the participant was able to sustain attention during the first half of PVT-B.

PVT-BA accuracy scores calculated for each individual averaged 92.3% and ranged between 65 and 100%. Correlation

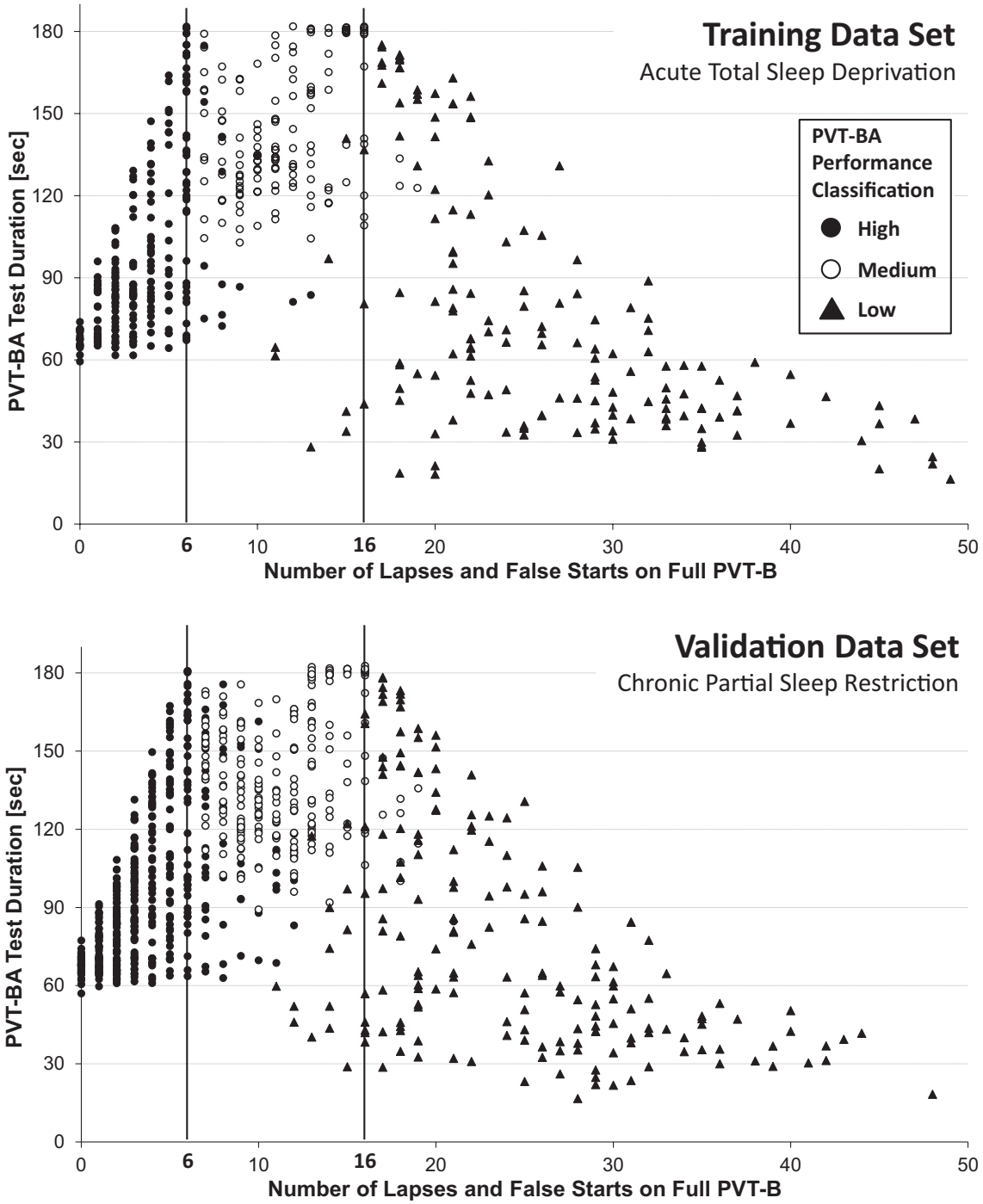

Figure 2. For each test bout and for both the training data set (above, N = 527 test bouts) and the validation data set (below, *N* = 880 test bouts), the number of lapses and false starts on the full 3 min PVT-B (abscissa), the classification of the test bout according to the adaptive duration PVT-BA (represented by different symbols), and the duration of PVT-BA (ordinate) are plotted. The vertical lines represent the category boundaries separating HIGH from MEDIUM (≤6 lapses and false starts) and MEDIUM from LOW (≤16 lapses and false starts) performance groups based on the full PVT-B. PVT-BA duration was highest for test bouts with number of lapses and false starts on the PVT-B near the category boundaries 6 and 16. Even for test bouts with no lapse or false start on the PVT-B, PVT-BA duration was still 57 s or longer, whereas PVT-BA duration decreased continuously to values below 20 s with an increasing number of lapses and false starts on the PVT-B. Misclassifications tended to be close to the category boundaries.

analyses suggest that PVT-B performance classified as MEDIUM performance was a main driver for low PVT-BA accuracy, while PVT-B performance classified as HIGH performance was a driver for high PVT-BA accuracy. This is in line with the finding that misclassifications across all tests were most likely to occur in the MEDIUM performance category. However, as shown in Figure 3, even in subjects with low PVT-BA accuracy and a high percentage of PVT-B tests classified as MEDIUM performance, misclassifications tended to be close to category boundaries: no test with 14–16 LpFS was misclassified as HIGH performance and no test with 7–10 LpFS was misclassified as LOW performance.

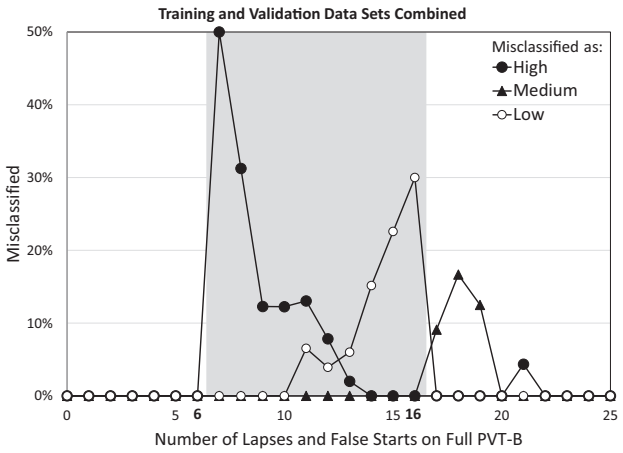

Figure 3. Percent of tests misclassified as HIGH (closed circles), MEDIUM (closed triangles), or LOW (open circles) is shown for the combined training and validation data sets depending on the number of lapses and false starts (LpFS) on the full 3 min PVT-B. The gray area shows the range of number of LpFS that should have been classified as MEDIUM performance. The percentage of tests misclassified was highest in the vicinity of the category boundaries of 6 LpFS and 16 LpFS (bolded). All HIGH performance bouts (≤6 LpFS) were correctly classified. Tests with 7 or 8 LpFS were frequently misclassified as HIGH instead of MEDIUM performance, while tests with 14-16 lapses were sometimes misclassified as LOW instead of MEDIUM performance. Tests with 17–19 LpFS were sometimes misclassified as MEDIUM instead of LOW performance. No test with ≥22 LpFS was misclassified. There was a single instance of a test in the validation data set with 21 LpFS that was incorrectly classified as HIGH instead of LOW performance (see text for discussion).

Obviously, the choice of the decision threshold (i.e. the posterior probability at which PVT-BA stops sampling data) affects both test performance and duration. Based on the data of the training data set, I chose a decision threshold of 99.619% that lead to a correct classification in >95% of test bouts and no misclassifications across two categories (i.e. HIGH classified as LOW or LOW classified as HIGH) in the training data set. This threshold decreased average test duration to 103.2 s in the training data set. Another sensible choice for the decision threshold would have been one that just lead to no misclassifications across two categories. In the training data set, this decision threshold (98.9593%) was associated with an average test duration of 91.4 s and 90.1% correct decisions. The assumption of conditional independence of consecutive tests was tested and confirmed previously for PVT-A [35]. It was not tested here again.

According to results shown in Figure 4, the MEDIUM performance category can best be described as a transitional category for all versions of the PVT investigated here. With increasing sleep pressure driven by homeostatic and circadian influences, participant performance transitioned from the HIGH to the MEDIUM and from the MEDIUM to the LOW performance category, causing a decrease in HIGH performance classifications, an increase in LOW performance classifications, but relatively stable MEDIUM performance classifications around 20–40% across all administrations.

None of the performance classifications differed between PVT, PVT-B and PVT-BA in the training data set. PVT-BA overestimated HIGH performance and underestimated MEDIUM performance in the validation data set relative to PVT-B, but LOW performance classifications did not differ. Interestingly, these observed classification differences made the PVT-BA more similar to classifications based on the standard 10 min PVT.

Limitations

Several limitations apply to this study. First, categorizing the continuous outcome metric LpFS on PVT-B into three discrete categories (HIGH, MEDIUM, and LOW performance) constitutes a loss of information. However, as explained above, I believe that it will be sufficient for many applied and research contexts to know whether a test bout qualifies as HIGH, MEDIUM, or LOW performance, and that the number of LpFS would add little relevant information. Also, it would be easy to report LpFS and other common PVT outcome metrics at PVT-BA termination and even project the expected LpFS for the full 3 min PVT-B in addition to the performance category. On the same note, it would be possible to introduce more than three performance categories, but PVT-BA test performance would have to be reestablished.

Second, PVT-B is a work-paced task (i.e. the behavior of the tested individual does not influence task duration). In contrast, PVT-BA 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 LpFS. While the first may be unproblematic as it is impossible to fake high performance, the latter is a more severe threat to the validity of the test, as non-compliant or poorly motivated subjects may choose to lapse frequently or bias towards false starts in order 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 non-compliance. For example, multiple consecutive false starts or lapses are rare events in motivated subjects.

Third, in our analyses the PVT-BA algorithm was applied post-hoc to data collected with PVT-B. However, I do not see major obstacles in implementing the PVT-BA algorithm in an online, real time fashion.

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

Finally, performance impairment on PVT-BA indicates reduced vigilant attention due to fatigue or other reasons. As 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-min PVT for a simulated luggage screening task [25]). However, it is unknown how PVT-BA specifically relates to many other tasks, and how well it predicts performance on these tasks.

Future studies should prospectively investigate how PVT-BA performance classified as HIGH, MEDIUM and LOW is associated with other cognitive and operational test outcomes both under controlled laboratory settings and in naturalistic field studies. Also, studies should investigate the utility of PVT-BA as a fit-for-duty tool, including its acceptability to employers and employees. Finally, it will be important to investigate PVT-BA performance in relation to clinical sleep disorders and assess its potential as a clinical screening tool.

Conclusions

I developed and validated an accurate, sensitive, and specific adaptive duration version of the 3 min PVT-B. Test duration of the

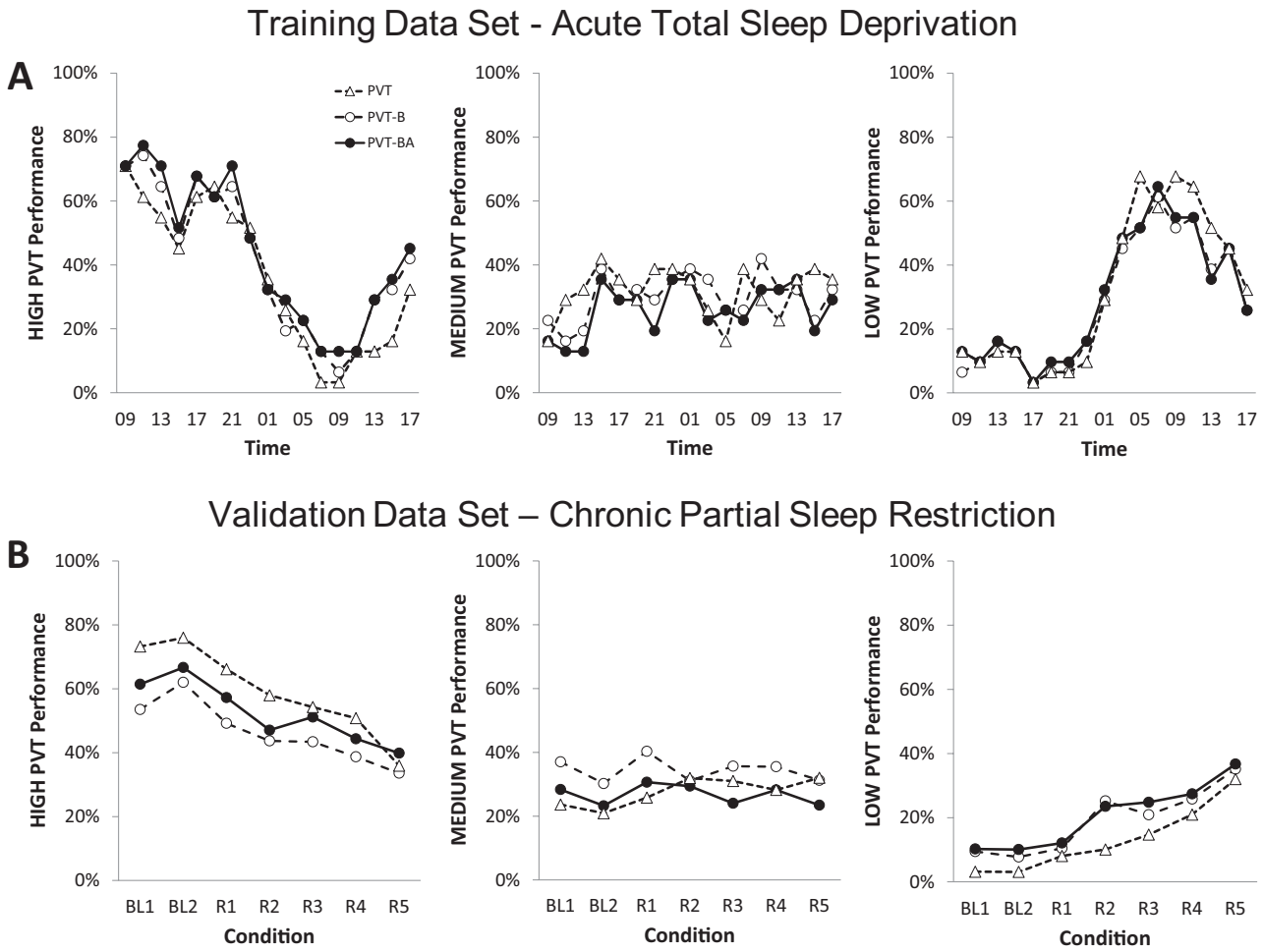

Figure 4. The proportion of performance bouts classified as HIGH, MEDIUM, or LOW is shown for the training data set (A, across 33 h of acute total sleep deprivation) and for the validation data set (B, chronic partial sleep deprivation with two baseline nights BL1 and BL2 and 5 nights of sleep restricted to 4 h time in bed R1 to R5). Classifications based on the full 3 min PVT-B are shown as open circles, while classifications based on the adaptive duration version of the PVT-B (PVT-BA) are shown as black circles. The figure also shows classifications based on the standard 10 min PVT, which was taken either shortly before or after the PVT-B. In general, agreement between the three versions of the test was high. In the training data set, the PVT-BA did not differ from either the PVT-B or full 10-min PVT in any of the performance categories (all $p > .09$). In the validation data set, PVT-BA significantly overestimated HIGH performance bouts relative to the PVT-B (mean difference +6.2%; $p = .004$) and significantly underestimated HIGH performance bouts relative to the 10 min PVT (mean difference –6.7%; $p < .001$). PVT-BA significantly underestimated MEDIUM performance bouts relative to the PVT-B (mean difference –7.7%; $p < .001$) but did not differ from the 10 min PVT (mean difference –0.8%; $p = .366$). PVT-BA significantly overestimated LOW performance bouts relative to the 10 min PVT (mean difference +7.1%; $p < .001$) but did not differ from PVT-B (mean difference +0.8%; $p = .391$).

adaptive PVT-BA averaged 1 min 43 s, with some tests terminating after less than 20 s, making it the briefest valid assessment of vigilant attention to date and increasing practicability of the PVT in operational and clinical settings. The adaptive duration strategy may be superior to a simple reduction of PVT duration where the fixed test duration may be too short to identify subjects with moderate impairment (showing deficits only later during the test), but unnecessary long for those who are either fully alert or severely impaired.

Acknowledgments

I published the manuscript that introduced the first adaptive version of the standard 10 min PVT with my long-term mentor and friend David F. Dinges, who originally invented the PVT. This manuscript is dedicated to David, who is not only an outstanding scientist but also a terrific mentor and human being. Like so many other trainees and mentees, I owe him a lot and would not be where I am today without him. Thank you, David!

This paper is based on and repeats many of the analyses presented in the manuscript first introducing the adaptive duration version of the standard 10 min PVT and several text passages are replicated unedited from this original publication [35]. I thank the subjects participating in the experiments and the faculty and staff who were involved in all aspects of the studies that provided the data for the analyses presented here, specifically Siobhan Banks, Hans van Dongen, Namni Goel as well as David F. Dinges who served as PI on these studies but was not included as a co-author due to the nature of this Festschrift. This investigation was sponsored by the Human Factors Program of the Transportation Security Laboratory, Science and Technology Directorate, U.S. Department of Homeland Security (FAA #04-G-010), by National Institutes of Health grants R01 NR004281 and UL1 RR024134, and in part by the National Space Biomedical Research Institute through NASA NCC 9-58.

Disclosure Statement

The author has no financial or non-financial conflicts of interest to disclose related to the work presented in this manuscript.

Data Availability Statement

The data underlying this article will be shared with personal identifiers removed on reasonable request to the corresponding author.

References

1. Dinges DF, *et al*. Microcomputer analysis of performance on a portable, simple visual RT task during sustained operations. *Behav Res Methods Instrum Comput*. 1985;**6**(17):652–655.
2. Basner M, *et al*. Maximizing sensitivity of the Psychomotor Vigilance Test (PVT) to sleep loss. *Sleep*. 2011;**34**(5):581–591.
3. Dorrian J, *et al*. Psychomotor vigilance performance: neurocognitive assay sensitive to sleep loss. In: *Sleep Deprivation: Clinical Issues, Pharmacology and Sleep Loss Effects*. New York, NY: Marcel Dekker, Inc.;2005:39–70.
4. Dinges DF, *et al*. Performing while sleepy: effects of experimentally-induced sleepiness. In: Monk TH, ed. *Sleep, Sleepiness and Performance*. Chichester, United Kingdom: John Wiley and Sons, Ltd.;1991:97–128.
5. Warm JS, *et al*. Vigilance requires hard mental work and is stressful. *Hum Factors*. 2008;**50**(3):433–441.
6. Dinges DF, *et al*. Managing fatigue by drowsiness detection: can technological promises be realized?. In: Hartley, L, ed. *Managing Fatigue in Transportation*. Proceedings of the 3rd Fatigue in Transportation conference, Fremantle, Australia; 1988.
7. Van Dongen HP, *et al*. 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**(2):117–126.
8. Gunzelmann G, *et al*. Fatigue in sustained attention: generalizing mechanisms for time awake to time on task. In: Ackerman PL, ed. *Cognitive Fatigue: Multidisciplinary Perspectives on Current Research and Future Applications*. Washington, D.C.: American Psychological Association;2010:83–101.
9. Chee MW, *et al*. Lapsing during sleep deprivation is associated with distributed changes in brain activation. *J Neurosci*. 2008;**28**(21):5519–5528.
10. Drummond SP, *et al*. The neural basis of the psychomotor vigilance task. *Sleep*. 2005;**28**(9):1059–1068.
11. Tomasi D, *et al*. Impairment of attentional networks after 1 night of sleep deprivation. *Cereb Cortex*. 2009;**19**(1):233–240.
12. Lim J, *et al*. Sleep deprivation impairs object-selective attention: a view from the ventral visual cortex. *PLoS One*. 2010;**5**(2):e9087.
13. Basner M, *et al*. Repeated administration effects on psychomotor vigilance test performance. *Sleep*. 2018;**41**(1):181–186.
14. Basner M, *et al*. Cognition test battery: adjusting for practice and stimulus set effects for varying administration intervals in high performing individuals. *J Clin Exp Neuropsychol*. 2020;**42**(5):516–529.
15. Goel N, *et al*. Neurocognitive consequences of sleep deprivation. *Semin Neurol*. 2009;**29**(4):320–339.
16. Lim J, *et al*. A meta-analysis of the impact of short-term sleep deprivation on cognitive variables. *Psychol Bull*. 2010;**136**(3):375–389.
17. Balkin TJ, *et al*. Comparative utility of instruments for monitoring sleepiness-related performance decrements in the operational environment. *J Sleep Res*. 2004;**13**(3):219–227.
18. Chua EC, *et al*. Heart rate variability can be used to estimate sleepiness-related decrements in psychomotor vigilance during total sleep deprivation. *Sleep*. 2012;**35**(3):325–334.
19. Philip P, *et al*. Transport and industrial safety, how are they affected by sleepiness and sleep restriction? *Sleep Med Rev*. 2006;**10**(5):347–356.
20. Dinges DF. An overview of sleepiness and accidents. *J Sleep Res*. 1995;**4**(S2):4–14.
21. Van Dongen HP, *et al*. Sleep, circadian rhythms, and psychomotor vigilance. *Clin Sports Med*. 2005;**24**(2):237–249, vii–viii.
22. Gunzelmann G, *et al*. Individual differences in sustained vigilant attention: insights from computational cognitive modeling. In: Proceedings from the 30th Annual Meeting of the Cognitive Science Society;2008; Washington, DC.
23. Van Dongen HP, *et al*. The efficacy of a restart break for recycling with optimal performance depends critically on circadian timing. *Sleep*. 2011;**34**(7):917–929.
24. Zhou X, *et al*. Mismatch between subjective alertness and objective performance under sleep restriction is greatest during the biological night. *J Sleep Res*. 2012;**21**(1):40–49.
25. Basner M, *et al*. 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**(10):1146–1154.
26. Gilliland K, *et al*. Readiness to Perform: A Critical Analysis of the Concept and Current Practices. Office of Aviation Medicine, Federal Aviation Administration;1993. DOT FAA AM-93 13.
27. Loh S, *et al*. The validity of psychomotor vigilance tasks of less than 10-minute duration. *Behav Res Methods Instrum Comput*. 2004;**36**(2):339–346.
28. Roach GD, *et al*. Can a shorter psychomotor vigilance task be used as a reasonable substitute for the ten-minute psychomotor vigilance task? *Chronobiol Int*. 2006;**23**(6):1379–1387.
29. Lamond N, *et al*. Fatigue assessment in the field: validation of a hand-held electronic psychomotor vigilance task. *Aviat Space Environ Med*. 2005;**76**(5):486–489.
30. Lamond N, *et al*. The sensitivity of a palm-based psychomotor vigilance task to severe sleep loss. *Behav Res Methods*. 2008;**40**(1):347–352.
31. Thorne DR, *et al*. The Walter Reed palm-held psychomotor vigilance test. *Behav Res Methods*. 2005;**37**(1):111–118.
32. Basner M, *et al*. Validity and sensitivity of a brief Psychomotor Vigilance Test (PVT-B) to total and partial sleep deprivation. *Acta Astronaut*. 2011;**69**(11–12):949–959.
33. Benderoth S, *et al*. Reliability and validity of a 3-min psychomotor vigilance task in assessing sensitivity to sleep loss and alcohol: fitness for duty in aviation and transportation. *Sleep*. 2021;**44**(11).
34. Lim J, *et al*. Imaging brain fatigue from sustained mental workload: an ASL perfusion study of the time-on-task effect. *Neuroimage*. 2010;**49**(4):3426–3435.
35. Basner M, *et al*. An adaptive duration version of the PVT accurately tracks changes in psychomotor vigilance induced by sleep restriction. *Sleep*. 2012;**35**(2):193–202.
36. Banks S, *et al*. Neurobehavioral dynamics following chronic sleep restriction: dose-response effects of one night of recovery. *Sleep*. 2010;**33**(8):1013–1026.
37. Dinges DF, *et al*. Pilot test of fatigue management technologies. *Transp Res Rec*. 2005;**1922**:175–182.
38. Basner M, *et al*. An ECG-based algorithm for the automatic identification of autonomic activations associated with cortical arousal. *Sleep*. 2007;**30**(10):1349–1361.
39. Hunink M, *et al*. *Decision Making in Health and Medicine: Integrating Evidence and Values*. Cambridge, U.K.: University Press;2001.
40. Fleiss J, *et al*. *Statistical Methods for Rates and Proportions*. Hoboken, NJ: John Wiley & Sons, Inc.;2003.
41. Landis JR, *et al*. The measurement of observer agreement for categorical data. *Biometrics*. 1977;**33**(1):159–174.