Published in final edited form as:
J Occup Environ Med. 2011 October ; 53(10): 1146–1154. doi:10.1097/JOM.0b013e31822b8356.

Fitness for duty: A 3 minute version of the Psychomotor Vigilance Test predicts fatigue related declines in luggage screening performance

Mathias Basner, MD, PhD, MSc[1] and **Joshua Rubinstein, PhD**[2]

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

[2]Human Factors Program, Transportation Security Laboratory, Science and Technology Directorate, U.S. Department of Homeland Security, Atlantic City, New Jersey, USA

Abstract

Objective—To evaluate the ability of a 3-min Psychomotor Vigilance Test (PVT) to predict fatigue related performance decrements on a simulated luggage screening task (SLST).

Methods—Thirty-six healthy non-professional subjects (mean age 30.8 years, 20 female) participated in a 4 day laboratory protocol including a 34 hour period of total sleep deprivation with PVT and SLST testing every 2 hours.

Results—Eleven and 20 lapses (355 ms threshold) on the PVT optimally divided SLST performance into high, medium, and low performance bouts with significantly decreasing threat detection performance A′. Assignment to the different SLST performance groups replicated homeostatic and circadian patterns during total sleep deprivation.

Conclusions—The 3 min PVT was able to predict performance on a simulated luggage screening task. Fitness-for-duty feasibility should now be tested in professional screeners and operational environments.

Keywords

readiness to perform; fatigue; sleep; safety; accident; alertness

Introduction

It was shown that fatigue from night work and sleep loss adversely affects threat detection performance on a task that simulates threat detection demands of airport screeners (*Simulated Luggage Screening Task* – SLST) (1). Thus, fatigue in luggage screening personnel may pose a threat for air traffic safety unless countermeasures for fatigue are deployed.

In this context, it would be highly desirable to predict threat detection performance based on a simple fitness-for-duty test or objective monitoring of screener alertness during the screening task, in an effort to assure high levels of vigilance and detection performance.

Correspondence address: Mathias Basner, MD, MS, MSc, Division of Sleep and Chronobiology, Department of Psychiatry, University of Pennsylvania School of Medicine, 1013 Blockley Hall, 423 Guardian Drive, Philadelphia, PA 19104-6021, USA, phone (215) 573-5866, fax (215) 573-6410, basner@mail.med.upenn.edu.

A fitness-for-duty test needs to fulfill certain scientific and practical properties to be useful in operational environments (2, 3): It needs to be both operationally and conceptually valid (Does it measure what it purports to measure?). It needs to be reliable (Does it measure the same thing consistently?) and generalizable (Does it measure the same event in everyone?). It needs to be both sensitive (i.e., detect all relevantly impaired individuals) and specific (i.e., it should not pick up alert or only moderately impaired individuals). It needs to be easy to use (Can nearly everyone use it correctly?), unobtrusive, robust, and acceptable for the target population. It needs to be brief and easy to administer. In this context, portable handheld devices are optimal. Finally, it needs to give some form of feedback to inform the operator about her or his alertness-level, i.e. whether she or he is fit to perform the task. As Gilliland and Schlegel (3) point out, a fitness-for-duty measure with criterion validity for both risk factors (e.g. fatigue) and job performance is optimal, especially when job performance could result in serious property loss or threats to public or personal safety.

The psychomotor vigilance test (PVT) is a high-signal-load test based on simple reaction time (RT) designed to evaluate the ability to sustain attention and respond in a timely manner to salient signals (4, 5, 6). The PVT has been shown to be a valid tool for assessing neurocognitive performance in a number of experimental, clinical, and operational paradigms (7, 8, 9, 10, 11, 12, 13, 14). It has been shown to be sensitive to both total sleep deprivation (15, 16) and chronic partial sleep deprivation (17, 18). Balkin et al. (19) assessed the utility of a variety of instruments for monitoring sleepiness-related performance decrements and concluded that the PVT “was among the most sensitive to sleep restriction, was among the most reliable with no evidence of learning over repeated administrations, and possesses characteristics that make it among the most practical for use in the operational environment”.

The standard test duration of the PVT is 10 min with inter-stimulus intervals (ISI) of 2–10 s. A certain test duration is required as even severely sleep deprived subjects may be able to perform normally for a short time by increasing compensatory effort. To establish whether state instability, as seen by degradation in performance on the SLST, can be detected using a predictive performance measure, we developed a 3-min handheld version of the PVT, which we call the fitness-for-duty PVT or FD-PVT. The duration of the test was shortened to 3 min and the signal rate was increased (ISIs 1–4 s). This FD-PVT was validated against the standard 10-min PVT in both total and partial sleep deprivation paradigms (20). It was shown to achieve similar levels of sensitivity and specificity as the standard 10-min PVT (see Figure 1).

In this study, once every 2 hours during a period of 34 h of total sleep deprivation the 3-min PVT was performed immediately before the SLST. We investigated the coherence of FD-PVT performance and SLST performance. The co-variation of FD-PVT performance and SLST performance is a prerequisite for the applicability of the FD-PVT in predicting SLST performance. We also developed a method for finding suitable decision thresholds for the FD-PVT. This method identified two optimal FD-PVT lapse thresholds separating high performance bouts from medium performance bouts and medium performance bouts from low performance bouts on the SLST. If the FD-PVT was applied in the field, high performers could carry out the task, medium performers could be warned about their impaired performance, but nevertheless perform the task, and low performers could be asked not to proceed with the task. Both medium and low performers may be required to take a break or apply other appropriate countermeasures (caffeine, etc.).

Methods

Subjects and protocol

An experiment was designed to systematically evaluate threat detection performance during an SLST (1). This analysis is based on data gathered in a pilot study on N=12 subjects (mean age 32.4 ± 8.2 years, 7 female) and on data sampled within the final protocol on N=24 subjects (mean age 29.9 ± 6.5 years, 13 female). In total data on N=36 subjects (mean age 30.8 ± 7.1 years, 20 female) were obtained and analyzed. The study was approved by the Institutional Review Board of the University of Pennsylvania. Participants were informed about potential risks of the study, and a written informed consent and HIPAA consent were obtained prior to the start of the study. Sleep diaries and actigraphy (Mini Mitter Co. Inc., Bend, OR, model AW64) were used to estimate subjects' sleep time one week prior to the study.

Simulated luggage Screening Task (SLST)

For the SLST, we developed an electronic luggage database that included a large number of X-ray source images of bags, clothing, etc., and threats (guns and knives only) (1). We used this system to produce more than 5,800 unique simulated X-ray images of luggage, organized into 31 stimulus sets of 200 bags each, with 50 bags of each set containing single threats that varied in type (gun or knife) and target difficulty (high or low). Four typical examples are shown in Figure 2A-D.

Subjects were oriented on two days. On both days, examples of separate clutter and threat images as well as complete bags with and without threats were shown to them first. Then, an SLST with 30 bags (orientation day 1) and 200 bags (orientation day 2) was simulated and discussed with the subjects. Study participants stayed in the research lab for 5 consecutive days, which included a 34h period of wakefulness (i.e., night work and day work after a night without sleep). The study started at 8:00 on day 1 and ended at 8:00 on day 5. During 1 of every 2 hours awake, subjects performed a computerized neurobehavioral test battery (NTB) that lasted approximately 25–30 min, followed by an SLST. On the first day of the study, all subjects performed seven training bouts of the SLST (see below). A 34h period of total sleep deprivation started either on the next day (N=24 subjects, including the N=12 subjects of the pilot study) or on the day after that (N=12 subjects), following an 8h sleep period (the latter condition was added to the final protocol due to a time-in-study effect in SLST performance that was found in the pilot study). Altogether, subjects performed 31 SLST bouts (7 training bouts, 24 work bouts) during the study. The composition of each SLST bout differed according to type and target difficulty of threats. During the pilot study all 12 subjects received the same unique SLST work bout at the same time of the day. These data showed that SLST bouts differed in difficulty, and thus, the 24 unique SLST work bouts were block randomized in a Latin square design in the final protocol (i.e., each SLST appeared in each position exactly once).

The 200-bag stimuli sets were run on software that simulates an X-ray screening system. Subjects had to press the space bar (colored green) for safe bags and the letter "D" (colored red) for threat bags. Except for the first training session, the threat-detection task timed out after 7 seconds, and threat bags were considered a miss, while safe bags were considered a correct rejection. A blank screen was shown for 1 s between presentations of two consecutive bags. During three of the seven training sessions detailed feedback was given to incorrect answers, i.e. if subjects missed a threat ("ERROR: weapon was present") or wrongly classified a safe bag as a threat bag ("ERROR: NO weapon present"). Otherwise, subjects were only informed about their overall percentage of hits and false alarms at the end of each 200-bag trial, but no detailed feedback was given. A text message also reminded

them that the main goal of the task was to keep the threat detection rate high, while the secondary goal was to keep the rate of false alarms low, and that they should keep trying to attain a perfect score.

Hit rate (HR, true positive rate, proportion of threats detected) and false alarm rate (FAR, false positive rate, proportion of safe bags wrongly classified as threats) were used to compute A′ and B''_D, non-parametric signal detection theory measures of sensitivity and response bias (21, 22). Sensitivity A′ reflects detection accuracy and reveals the extent to which subjects are able to differentiate signal (threat bags) from noise (safe bags). A′ varies between 0.5 (signals cannot be distinguished from noise, performance at chance level) to 1.0 (complete separation of signal and noise, perfect accuracy). A′ can also be interpreted as the proportion of times subjects would correctly identify the signal, if signal and noise stimuli were presented simultaneously (21). A′ is unaffected by response bias, i.e. a subject's general willingness for responding "threat bag" versus "safe bag". B''_D is a measure of this response bias and ranges from −1 (liberal bias, all bags are classified as threats) to +1 (conservative bias, all bags are classified as safe), with 0 indicating no response bias.

Psychomotor Vigilance Test (PVT)

The 3 min PVT was performed on the PVT-192 (Ambulatory Monitoring Inc., Ardsley, NY), a handheld device measuring 21 × 11 × 6 cm and weighing ca. 650 g. The visual RT stimulus and performance feedback were presented on the device's 2.5 × 1 cm four-digit LED display. The inter-stimulus intervals varied randomly from 1–4 s. Subjects were instructed to press the response button as soon as each stimulus appeared, in order to keep RT as low as possible, but not to press the button too soon (which yielded a false start warning on the display). An auditory signal was given after a 30 s period without response. Because of their high operational validity, we chose lapses as the primary PVT outcome metric. In a systematic comparison of PVT outcome measures, lapses were shown to score high effect sizes (23). In our validation of the 3-min PVT we observed that subjects were generally faster on the 3-min compared to the 10-min PVT (20). In order to achieve a comparable number of lapses on both tests, lapse definition for the 3-min PVT was lowered from the standard ≥500 ms definition to ≥355 ms.

Coherence of 3-min PVT and SLST performance

We computed simple measures of coherence in the time domain using Pearson and Spearman rank correlations for each set of N=17 bivariate PVT/SLST data pairs. Although *coherence* is a term typically used in frequency domain analyses, it can be operationalized here in the time domain - as the Pearson correlation over time points within each subject between the FD-PVT and the SLST. Spearman rank correlations were also computed in order to assess whether the Pearson correlations were robust - that is, whether their values were highly influenced by a few extreme values. With rare exceptions, it was found that subject-specific Pearson and Spearman values were consistently similar. Thus, the primary analyses were based on Pearson correlation coefficients.

Decision threshold determination for 3-min PVT ≥355 ms lapses

The basic idea of the PVT fitness-for-duty test is to predict SLST performance based on ≥355 ms lapses on the PVT. We decided to divide SLST performance into three groups (high, medium, and low). We used the following method to find two PVT lapse thresholds that optimally differentiated between high, medium, and low SLST performance bouts: Within each subject, signal detection performance A′ on the 17 SLST bouts was rank ordered from high to low, and bouts were categorized into three groups (see Figure 3); high performance (ordered bouts 1–5), medium performance (ordered bouts 7–11), and low performance (ordered bouts 13–17). Bouts 6 and 12 separated the groups and were not used

for further analysis[1] (Figure 3 shows the rank ordered 17 test bouts of one subject by way of example). Each SLST A′ value was associated with a certain number of ≥355 ms lapses on the 3-min PVT (also plotted in Figure 3).

In the next step, data from all subjects were pooled. Every subject contributed 15 data points, each having the two attributes "signal detection performance category" (high, medium, or low) and "number of ≥355 ms lapses on the 3-min PVT". Two 3-min PVT lapse thresholds were used to divide the pooled data set into three groups. The two lapse thresholds were systematically varied until the percentage of correct SLST performance classifications was maximal. The lower lapse threshold differentiated high and medium SLST performance bouts, whereas the higher lapse threshold differentiated medium and low SLST performance bouts.

The procedure described above has two favorable properties. First, classification of SLST performance is based on rank ordering within subjects, eliminating influences of inter-individual differences in SLST performance on group classification. Second, subjects insensitive to sleep loss always score high, while subjects with prior sleep debt, non-compliant subjects, or de-motivated subjects always score low on the PVT. Therefore, these subjects will not contribute substantially to the determination of decision thresholds.

HR, FAR, A′, B''_D, test bout duration, and hours awake since wake-up time were compared between high, medium, and low SLST performance bouts (as determined by the 3-min PVT) with separate random subject effect regression models (PROC MIXED in SAS, Version 9.1, SAS Institute). As we found significant differences in HR ($P<0.001$), A′ ($P=0.034$), and B''_D ($P=0.044$) between the pilot study and the final protocol, these models were adjusted for pilot study membership.

Results

Effects of night work and sleep loss on SLST performance

Table 1 summarizes the effects of night work, sleep loss and time in study on SLST performance, response bias B''_D and bout duration as reported earlier (1). SLST performance A′ was affected by sleep deprivation. It decreased significantly during both night work and sleep loss. Both misses (errors of omission) and false alarms (errors of commission) increased during night work and sleep loss. Response bias B''_D decreased non-significantly, and subjects took less time to scan for threats while sleep deprived, as they completed the 200 bag set earlier. Detection accuracy A′ increased only marginally and non-significantly with time in study. However, there was a significant shift in response bias with time in study to more liberal criteria (i.e., both threat bags and safe bags were classified more often as threats), leading to significantly increased hit rates and false alarm rates at the end compared to the beginning of the study. Additionally, subjects took significantly less time to scan for threats at the end compared to the beginning of the study.

Coherence of 3-min PVT and SLST performance

The coherence of SLST performance and ≥355 ms lapses on the 3-min PVT is shown in Figure 4. Both measures were stable until 11 pm. After 11 pm, threat detection performance decreased while the number of lapses on the 3-min PVT increased. Extreme values were reached between 5 am and 7 am. Coherences based on averages across subjects was −0.913 ($p<.001$, Table 2). Average within subject coherence was −0.286 (SD 0.302). Within subject

[1]These bouts were used in subjects where data loss led to 15 or 16 instead of 17 valid SLST/PVT value pairs. One subject with fewer than 15 valid SLST/PVT value pairs was excluded from the data set that was used for finding optimal decision thresholds.

coherence ranged from −0.897 ($p<.001$) to 0.309 ($p=0.227$). SLST performance A′ and lapses on the 3-min PVT were significantly ($p<0.05$) negatively correlated in 9 subjects, non-significantly negatively correlated in 19 subjects and non-significantly positively correlated in 8 subjects.

Decision threshold determination for 3-min PVT ≥355 ms lapses

Figure 5 shows the percentage of correct classifications between high and medium SLST performance bouts and medium and low SLST performance bouts depending on the number of ≥355 ms lapses on the 3-min PVT. The highest percentage of correct classifications was reached at decision thresholds of 11 and 20 lapses, constituting three SLST performance groups:

- High SLST performance: 0–11 PVT lapses
- Medium SLST performance: 12–20 PVT lapses
- Low SLST performance: >20 PVT lapses

The number of ≥355 ms lapses per bout during 34 h of sleep deprivation is shown for each subject in Figure 6. Based on lapses, single bouts were categorized into the three SLST performance groups (high performance=white, medium performance=gray, and low performance=black background in Figure 6). Homeostatic and circadian influences on alertness were replicated by the PVT classification. Between 9:00 and 23:00, less than 10% of the subjects were categorized as low SLST performers, with the exception of 15:00 (at the afternoon dip), where 11% (5 out of 36) were classified as low SLST performers. At 19:00 no one was categorized as a low SLST performer, and 83% were categorized as high SLST performers. 31 of the 36 subjects were never classified as low performers between 9:00 and 23:00. After 23:00, fewer subjects were classified as high performers in favor of medium and low performance categories. The low performance group reached its maximum size at 5:00 (44%), the medium performance group at 7:00 (36%). Overall, 61.2% of 3-min PVT bouts were classified as being associated with high, 19.0% with medium, and 19.8% with low SLST performance during the 34-h period of total sleep deprivation.

The three PVT based SLST performance categories were compared according to A′, HR, FAR, B''_D, bout duration and hours awake (since wake-up time) in Figure 7. These analyses were adjusted for pilot study membership. Threat detection performance A′ and HR were both significantly lower in medium compared to high performers, in low compared to medium performers, and in low compared to high performers. HR in high performers was on average 7.0% higher compared to low performers. At the same time, FAR was higher in medium compared to high performers, in low compared to medium performers, and significantly higher in low compared to high performers. Here, FAR in low performers was on average 1.7% higher compared to high performers. Response bias increased non-significantly from high to medium to low performance groups. High performers took significantly more time to screen bags for threats. The same holds for medium versus low and high versus low performers. High performers needed on average 15.6 min to complete a 200-bag set, and therefore 2.0 min longer ($p<0.001$) than subjects in the low performance group. Test bouts in the high performance group (on average 14.4 h awake) were significantly closer to wake-up time than bouts in the medium (19.1 h awake) and low performance groups (21.9 h awake), and those in the medium performance group were significantly closer to wake-up time than bouts in the low performance group.

Discussion

Screening luggage for threats requires high and sustained levels of vigilant attention due to the continuous requirement for detecting weak and infrequent signals among high levels of

background clutter. It was shown that threat detection performance on a simulated luggage screening task deteriorated during night work and sleep loss, and that SLST and PVT performance covaried over a 34-h period of total sleep deprivation, a prerequisite for the ability of the 3-min PVT to predict SLST performance. Coherence of average SLST performance across subjects and average ≥355 ms lapse frequency on the 3-min PVT across subjects was high and 83% in the variability of SLST performance were explained by 3-min PVT outcomes. Within subject coherence was lower, probably due to some individuals insensitive to sleep loss and with low variability in both SLST and PVT performance. However, SLST and PVT performance were still negatively correlated within the majority of subjects (28 out of 36).

The purpose of fitness-for-duty testing is to detect relevantly impaired individuals unfit for the job. The relevance of impairment needs to be defined within the context of each predicted task, i.e. a certain level of impairment may be tolerated in one task, but it may be considered detrimental in another, especially in safety sensitive jobs (e.g. truck driving or luggage screening). In any case, some sort of feedback has to be given by the fitness-for-duty test, and this feedback has to lead to consequences. The type of feedback can range from a continuous measure of impairment (e.g. the percentage level of performance relative to a standard ranging from 0%–100%) to a dichotomous outcome (fit/unfit). A continuous measure of impairment in itself is of little use. One or more decision thresholds are needed to assign ranges of fitness-for-duty test outcomes to levels of impairment that are connected with specific consequences (in its easiest form whether or not the subject is allowed to perform the task). In our view, one threshold dividing fitness-for-duty test outcomes in “fit” and “unfit” may not be enough, because it is questionable whether subjects performing just above or below the single decision threshold are really fit or unfit to perform the task.

Here, we presented a method to find optimal decision thresholds to assign ≥355 ms lapses on the 3-min PVT to three SLST performance categories (high, medium, and low), although a higher number of categories could be used, too. 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 (24). If subjects falling in the medium performance category are not allowed to perform the task, employers need to be aware that the increase in performance on the task comes with the cost of excluding greater numbers of employees from doing their job.

The mixed models used for analyzing differences between high, medium, and low performance groups showed that 40.5% of the variance in SLST performance A′ were attributable to inter-individual differences, demonstrating that SLST performance varied considerably between subjects. However, the PVT is only able to predict relative decreases in SLST performance caused by fatigue or other influences both affecting PVT and SLST performance (e.g. alcohol and drugs, although this was not explicitly tested in this study) within a given subject. This is an important distinction, as the PVT is no surrogate or measure of absolute SLST performance between subjects (i.e. high performance on the PVT in a given individual does not guarantee high performance on the SLST). Psychomotor vigilant attention is an important but not the only factor determining SLST performance, e.g., screeners differ in their ability to fixate and recognize threat objects among high levels of background clutter (25). Thus, the determination of decision thresholds based on the rank order of SLST performance within subjects seemed a natural and reasonable approach.

After pooling the data of all subjects, decision thresholds of ≤11 ≥355 ms lapses and >20 ≥355 ms lapses on the 3-min PVT led to the highest percentage of correct assignments to the three performance groups. These two thresholds were optimal for the whole group, which does not imply that they were also optimal for each individual. In fact, prediction of job performance could be improved if thresholds were determined for each subject individually. However, this would require each subject running through an experimental protocol involving sleep deprivation and both FD-PVT and SLST testing, which does not seem a practical approach.

The categorization of all bouts of all subjects into the three SLST performance categories based on ≥355 ms lapses on the 3-min PVT was shown in Figure 6. Homeostatic and circadian influences of sleep deprivation were well replicated by the size of the three performance groups. The categorization was sensitive to sleep loss, i.e. more subjects showed up in the medium and low SLST performance groups after 16 hours of wakefulness, going well along with recent findings of a chronic sleep restriction experiment reported by Van Dongen et al. (17), who observed performance decrements only after wakefulness was chronically extended over 15.8 hours per day. At the same time, the categorization was specific in the sense that only a minority of subjects was assigned to the low SLST performance group during the first 15 hours of sleep deprivation (79.5% of the bouts were classified as high, 14.2% as medium, and only 6.3% as low SLST performance bouts).

There were five subjects that were classified at least once as low performers at or before 23:00. Overall, the average number of lapses was exceptionally high in these five subjects, ranking at 31, 32, 34, 35, and 36 relative to all 36 subjects. A prior sleep debt is one possible reason for the low PVT performance of these five subjects during the day, where one night with 8 hours time in bed in an unfamiliar environment may not have been enough to recover from this sleep debt. Sleep time prior to the study was determined with diaries and actigraphy. It averaged 8.06 h across days and subjects. Subject #3 had the fifth shortest average TST with 7.04 h and slept only 5 h in the night before the study began. Subject #8 slept above average (his average TST was 8.32 h), but only 4.75 h on the night preceding the study. Subject #35 had the second shortest individual average TST with 6.82 h, but slept 8 h in the night before the study began. Low TST prior to the study could not explain the low PVT performance of subjects #16 and #30, but subject #16 reported muscular aches 5 of 7 days, backache 4 of 7 days, joint pain 3 of 7 days, feeling too cold 5 of 7 days, and tiredness 2 of 7 days prior to the study, and subject #30 drank many caffeinated teas and colas before the study, which were not allowed during the study. As the PVT is not specific for fatigue, other reasons could have contributed to the overall low PVT performance levels of these five subjects, like alcohol, drugs, illness, and motivational factors, although alcohol, drugs, and illness are unlikely to have played a role because of screening tests and the controlled environment of the laboratory. This was an intent to treat analysis, and therefore we did not exclude any of the five subjects with low levels of PVT performance during the day.

Comparisons of high, medium, and low SLST performance groups as categorized by ≥355 ms lapse frequency on the 3-min PVT showed that threat detection performance differed significantly between the three groups confirming the ability of FD-PVT to predict SLST performance. HR was 7.0% and 4.0% higher in the high and the medium performance groups compared to the low performance group. Therefore, by allowing subjects with >20 lapses on the 3-min PVT to continue screening between 4 and 7 out of 100 threats would potentially be missed because of fatigue. It has to be borne in mind that these estimates of HR are conservative as they are confounded by time in study (there was a prominent effect of time in study on HR, and high, medium, and low performance groups differed significantly according to time in study). By controlling for hours awake (additional to pilot

study membership) differences in HR between high and medium performance groups to the low performance group increased from 7.0% to 8.0% and from 4.0% to 4.3%., respectively

At the same time, differences between high, medium, and low performance groups in FAR became insignificant by controlling for hours awake. Controlling for hours awake decreased the differences between performance groups in A′ marginally, but they remained significant between all groups (all $p<0.01$). Response bias increased from high to medium to low performance groups (although non-significantly), meaning that the willingness to classify both safe bags and threat bags as threats increased with the number of lapses on the 3-min PVT. By controlling for hours awake this trend was amplified, and low and high performance groups now differed significantly from each other ($p=.030$). Time used to complete a 200-bag set significantly decreased from high to medium to low performance groups. The difference between groups decreased by controlling for hours awake, but the low performance group still differed significantly from the medium (−0.7 min, $p=.007$) and the high (−0.9 min, $p<.001$) performance groups.

In summary, low ≥355 ms lapse frequency on the FD-PVT was associated with high threat detection performance, with high HRs, with conservative decision criteria (subjects less willing to classify both safe bags and threat bags as threats), and with long per bag response latencies. FD-PVT ≥355 ms lapse frequencies were not related to FAR, especially after controlling for hours awake.

Limitations

Our subject group consisted of healthy, young to middle aged, non-professional volunteers. It is therefore unclear how the results generalize to a group of professional luggage screeners who receive a special training that is refreshed on a regular basis. Also, in contrast to our subjects professional luggage screeners do not receive detection performance feedback at the end of their shift, which may restrict generalizability of our findings. Furthermore, it is unclear how the effects of sleep loss in the laboratory can be generalized to the operational environment. In other professions that have been studied (e.g., pilots, physicians in training, professional truck drivers), this was the case, although the magnitude of the effects can be smaller in a well trained professional cohort than in unselected laboratory subjects (26, 27, 28). Another reason our results may not readily generalize to the airport security environment is because the 25% threat prevalence we used was higher than found in security operations, at least for guns and knives. If all prohibited items (e.g., bottles, pocket knives, nail files/clippers, lighters, etc.) are taken into account, 25% threat prevalence is, in fact, not unusual at airport checkpoints, and there are times when the combined rate of different classes of prohibited items can exceed 25%. To determine the extent to which our findings have ecological validity, studies would have to be conducted on trained professional screeners and in operational environments.

Conclusion

A simulated luggage screening task was shown to be sensitive to night work and sleep loss and to co-vary with the number of ≥355 ms lapses on the 3-min PVT. A method was proposed to identify optimal 3-min PVT lapse decision thresholds for dividing SLST bouts into high, medium, and low performance bouts. Group classification was both specific and sensitive, i.e. only a minority of subjects was classified as low performers during the first 15 hours of sleeplessness, while most of the subjects transitioned from high to medium and low performance groups after 17 or more hours of wakefulness. It was shown that assignment to the different performance groups replicated homeostatic and circadian patterns during total sleep deprivation and that threat detection performance A′ and HR decreased significantly from high to medium to low performance groups. In conclusion, the 3-min PVT was shown

to be sensitive to sleep loss and to predict performance in a simulated luggage screening task, and should therefore be further validated as a fitness-for-duty test for luggage screeners. Future studies need to proof its feasibility and usefulness in professional screeners and operational environments.

Acknowledgments

This investigation was sponsored by support to XXX (Principal Investigator) from the Human Factors Program of the Transportation Security Laboratory, Science and Technology Directorate, U.S. Department of Homeland Security (FAA #04-G-010), by NIH grants R01 NR004281 and UL1 RR024134, and in part by the National Space Biomedical Research Institute through NASA NCC 9–58 and by NASA through NNX08AY09G. We would like to thank faculty and staff at XXX for gathering the data and the subjects for participating in our studies.

Funding: This investigation was sponsored by the Human Factors Program of the Transportation Security Laboratory, Science and Technology Directorate, U.S. Department of Homeland Security (FAA #04-G-010), by NIH grants R01 NR004281 and UL1 RR024134, and in part by the National Space Biomedical Research Institute through NASA NCC 9–58.

Literature

1. 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; 31:1251–1259. [PubMed: 18788650]
2. Dinges, DF.; Mallis, M. Managing fatigue by drowsiness detection: can technological promises be realized?. Hartley, L., editor. Pergamon; 1998. p. 209-229.
3. Gilliland, K.; Schlegel, RE. Readiness to perform: A critical analysis of the concept and current practices. Office of Aviation Medicine, Federal Aviation Administration; 1993.
4. Dinges DF, Powell JW. Microcomputer analysis of performance on a portable, simple visual RT task during sustained operations. BehavResMethods InstrumComput. 1985; 6:652–655.
5. Lim, J.; Dinges, DF. Molecular and Biophysical Mechanisms of Arousal, Alertness, and Attention Annals of the New York Academy of Sciences. Oxford: Blackwell Publishing; 2008. Sleep deprivation and vigilant attention; p. 305-322.
6. Goel N, Rao H, Durmer JS, Dinges DF. Neurocognitive consequences of sleep deprivation. SeminNeurol. 2009; 29:320–339.
7. Wyatt JK, Ritz-De Cecco A, Czeisler CA, Dijk DJ. Circadian temperature and melatonin rhythms, sleep, and neurobehavioral function in humans living on a 20-h day. Am J Physiol. 1999; 277:R1152–1163. [PubMed: 10516257]
8. Graw P, Krauchi K, Knoblauch V, Wirz-Justice A, Cajochen C. Circadian and wake-dependent modulation of fastest and slowest reaction times during the psychomotor vigilance task. Physiol Behav. 2004; 80:695–701. [PubMed: 14984804]
9. Van Dongen HP, Baynard MD, Maislin G, Dinges DF. Systematic interindividual differences in neurobehavioral impairment from sleep loss: evidence of trait-like differential vulnerability. Sleep. 2004; 27:423–433. [PubMed: 15164894]
10. Neri DF, Oyung RL, Colletti LM, Mallis MM, Tam PY, Dinges DF. Controlled Breaks as a Fatigue Countermeasure on the Flight Deck. Aviation Space and Environmental Medicine. 2002; 73:654–664.
11. Kribbs NB, Pack AI, Kline LR, et al. Effects of one night without nasal CPAP treatment on sleep and sleepiness in patients with obstructive sleep apnea. Am Rev Respir Dis. 1993; 147:1162–1168. [PubMed: 8484626]
12. Wesensten NJ, Belenky G, Thorne DR, Kautz MA, Balkin TJ. Modafinil vs. caffeine: effects on fatigue during sleep deprivation. Aviat Space Environ Med. 2004; 75:520–525. [PubMed: 15198278]
13. Wyatt JK, Cajochen C, Ritz-De Cecco A, Czeisler CA, Dijk DJ. Low-dose repeated caffeine administration for circadian-phase-dependent performance degradation during extended wakefulness. Sleep. 2004; 27:374–381. [PubMed: 15164887]
14. Signal TL, Gander PH, Anderson H, Brash S. Scheduled napping as a countermeasure to sleepiness in air traffic controllers. JSleep Res. 2009; 18:11–19. [PubMed: 19250171]

15. Doran SM, Van Dongen HP, Dinges DF. Sustained attention performance during sleep deprivation: Evidence of state instability. Archives Italiennes de Biologie: A Journal of Neuroscience. 2001; 139:1–15.

16. Jewett ME, Dijk DJ, Kronauer RE, Dinges DF. Dose-response relationship between sleep duration and human psychomotor vigilance and subjective alertness. Sleep. 1999; 22:171–179. [PubMed: 10201061]

17. Van Dongen HP, Maislin G, Mullington JM, Dinges DF. The cumulative cost of additional wakefulness: dose-response effects on neurobehavioral functions and sleep physiology from chronic sleep restriction and total sleep deprivation. Sleep. 2003; 26:117–126. [PubMed: 12683469]

18. Belenky G, Wesensten NJ, Thorne DR, et al. Patterns of performance degradation and restoration during sleep restriction and subsequent recovery: a sleep dose-response study. JSleep Res. 2003; 12:1–12. [PubMed: 12603781]

19. Balkin TJ, Bliese PD, Belenky G, et al. Comparative utility of instruments for monitoring sleepiness-related performance decrements in the operational environment. J Sleep Res. 2004; 13:219–227. [PubMed: 15339257]

20. Basner M, Dinges DF. Validity and sensitivity of a brief Psychomotor Vigilance Test (PVT-B) to total and partial sleep deprivation. Acta Astronautica. in revision.

21. Stanislaw H, Todorov N. Calculation of signal detection theory measures. BehavResMethods InstrumComput. 1999; 31:137–149.

22. Donaldson W. Measuring recognition memory. JExpPsycholGen. 1992; 121:275–277.

23. Basner M, Dinges DF. Maximizing sensitivity of the Psychomotor Vigilance Test (PVT) to sleep loss. Sleep. 2011; 34:581–591. [PubMed: 21532951]

24. Dinges DF, Maislin G, Brewster RM, Krueger GP, Carroll RJ. Pilot test of fatigue management technologies. Transportation Research Record: Journal of the Transportation Research Board. 2005; 1922:175–182.

25. McCarley JS, Kramer AF, Wickens CD, Vidoni ED, Boot WR. Visual skills in airport-security screening. PsycholSci. 2004; 15:302–306.

26. Rosekind MR, Boyd JN, Gregory KB, Glotzbach SF, Blank RC. Alertness management in 24/7 settings: lessons from aviation. OccupMed. 2002; 17:247–259. iv.

27. Philibert I. Sleep loss and performance in residents and nonphysicians: a meta-analytic examination. Sleep. 2005; 28:1392–1402. [PubMed: 16335329]

28. Philip P, Akerstedt T. Transport and industrial safety, how are they affected by sleepiness and sleep restriction? Sleep Medicine Reviews. 2006; 10:347–356. [PubMed: 16920370]

29. Efron, B.; Tibshirani, RJ. An introduction to the bootstrap. New York, NY: Chapman & Hall; 1993.

30. Curran-Everett D. Multiple comparisons: philosophies and illustrations. AmJ Physiol RegulIntegrComp Physiol. 2000; 279:R1–R8.

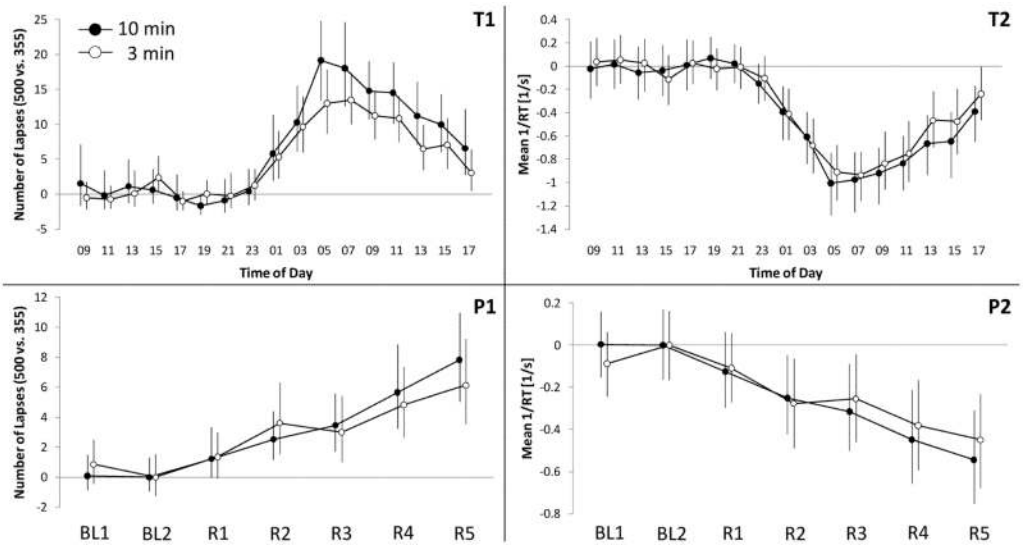

Figure 1.

Coherence of the 3-min (open circles) and the standard 10-min PVT (black circles) is shown for total (graphs T1 and T2, N=31 subjects staying awake for 34 h) and partial sleep deprivation (graphs P1 and P2, N=43 subjects restricted to 4 h TIB for 5 consecutive nights during R1 to R5 after 2 baseline nights BL1 and BL2 with 10 h TIB) for the outcome metrics mean reciprocal response time (mean 1/RT) and lapses (lapse thresholds 500 ms for the 10 min PVT and 355 ms for the 3 min PVT). Results were centered around daytime performance (bouts 1 to 7 for T1 and T2) and baseline performance on day 2 (BL2 for P1 and P2), respectively. Error bars represent bias-corrected and accelerated (BCa) 95% confidence intervals based on a bootstrap sample with 1,000,000 replications (29). Paired t-tests (two-sided and adjusted for multiple testing with the false discovery rate method (30)) indicated that there were no statistically significant differences between the 3 min and the 10 min PVT for bouts 8 to 17 (T1 and T2) and R1 to R5 (P1 and P2) at alpha 0.05.

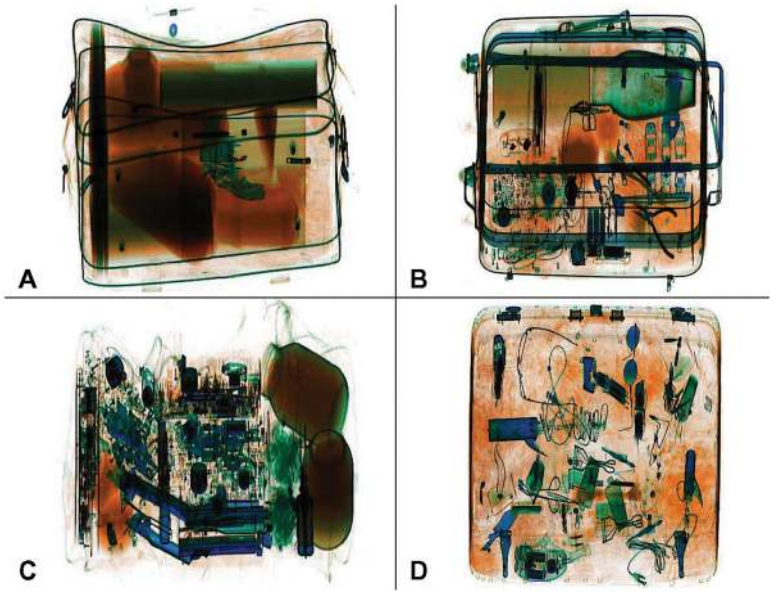

Figure 2.
Examples of simulated X-ray images of threat bags with typical hit rates (HR). A: gun with low target difficulty in the center (HR was 75%), B: knife with low target difficulty in upper right corner (HR was 56.5%), C: gun with high target difficulty in lower right corner (HR was 50%), D: knife with high target difficulty in lower left corner (HR was 32.5%). Reproduced from 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(9):1251–1259 (with permission)

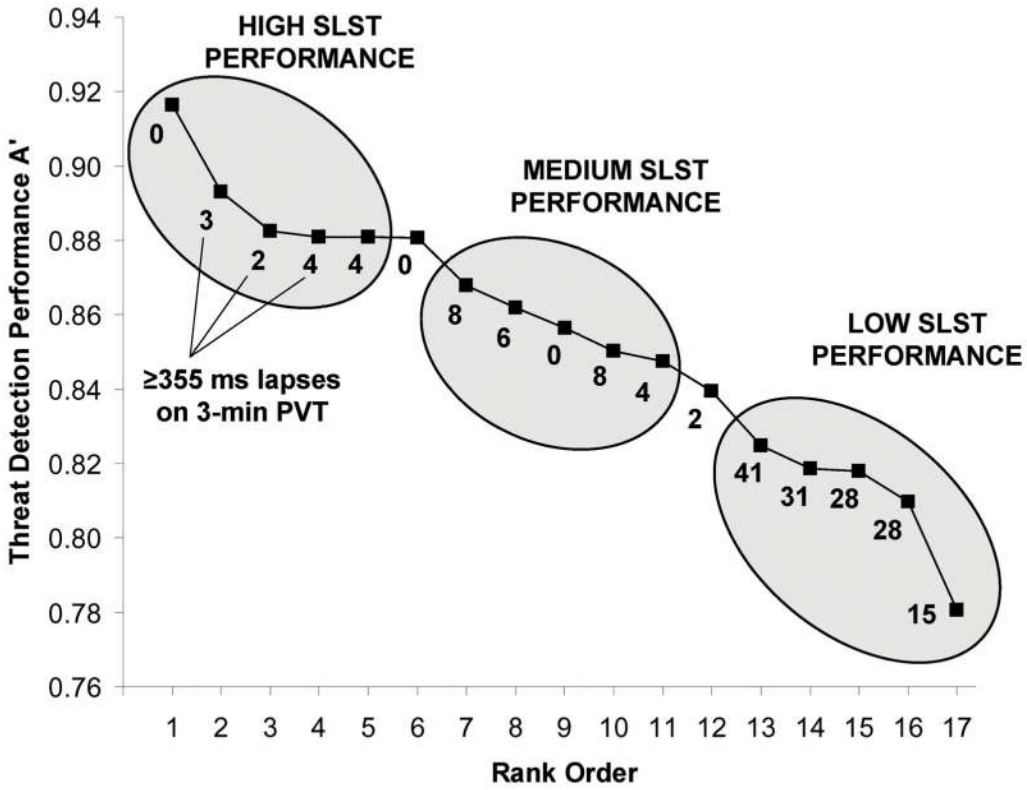

Figure 3.

Visualization of the method for determination optimal lapse thresholds. Bouts were rank ordered within subjects from highest to lowest threat detection performance A′, and then categorized into high (ordered bouts 1–5), medium (ordered bouts 7–11), and low (ordered bouts (12–17) performance bouts (data of one subject are shown by way of example). Each SLST bout was associated with a certain ≥355 ms lapse frequency on the 3-min PVT (number of lapses shown below each square representing an SLST bout). Data of all subjects were then pooled and lapse frequencies with the highest percentage of correct classifications according to SLST performance group were determined.

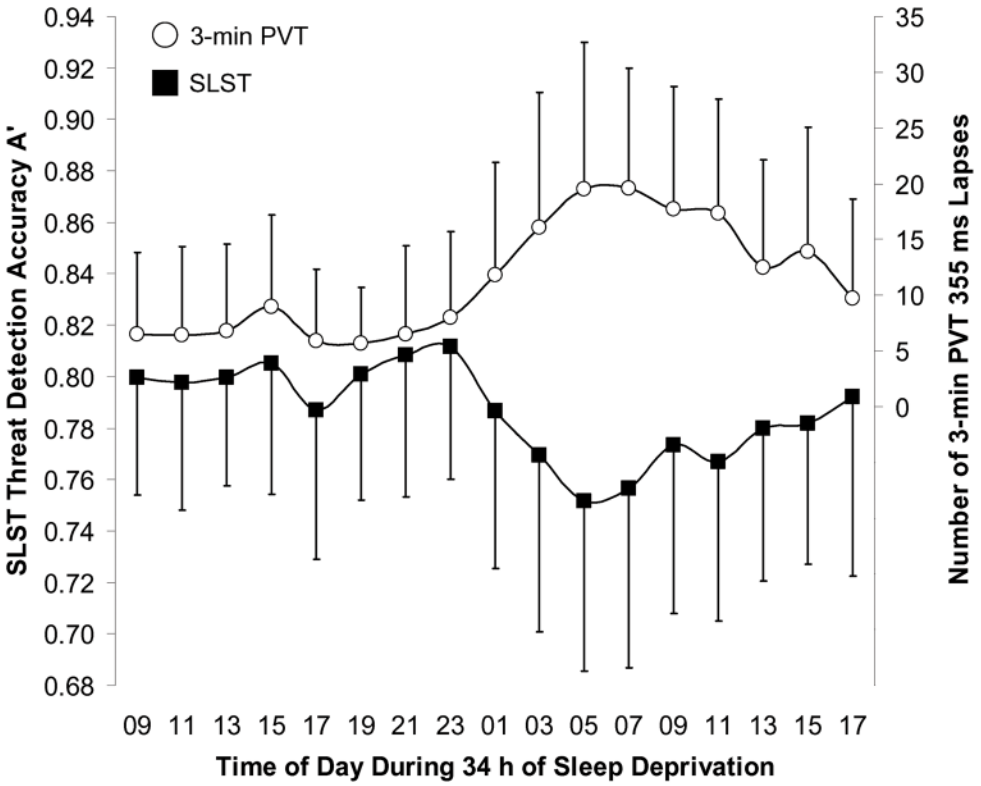

Figure 4.
Co-variation of mean (including standard deviation) SLST threat detection accuracy A′ (black squares, left ordinate) and average number of ≥355 ms lapses on the 3-min PVT (white circles, right ordinate) during a 34-h period of total sleep deprivation.

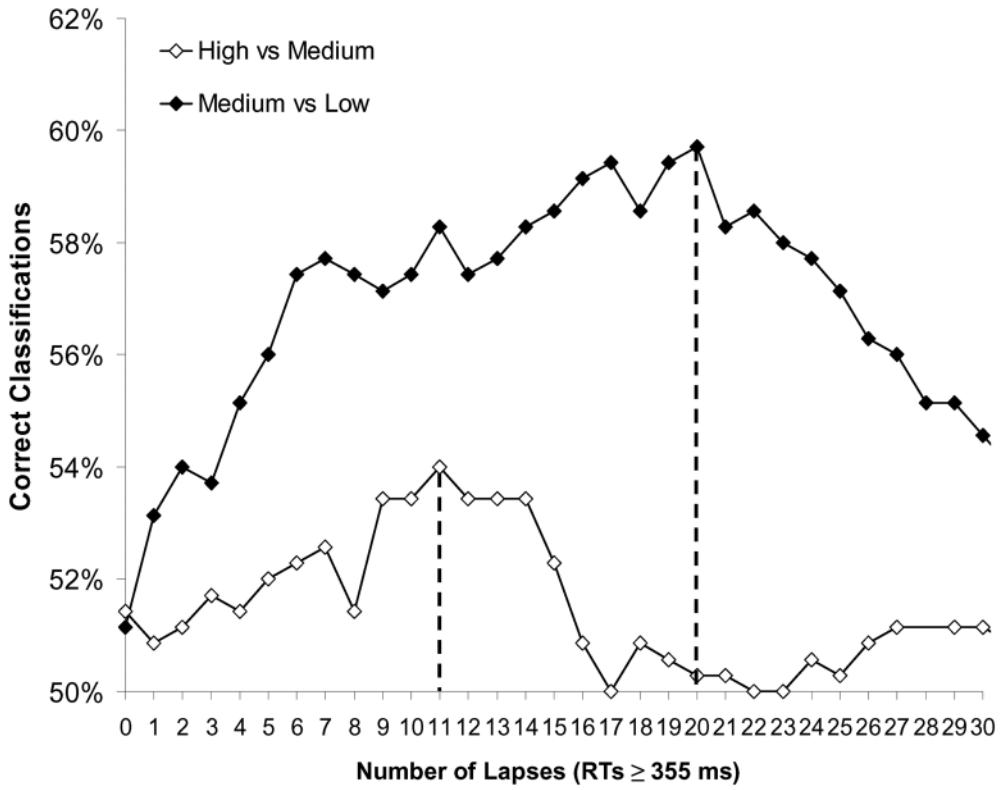

Figure 5.
The number of correct classifications between high and medium (white diamonds) and medium and low (black diamonds) SLST performance bouts based on the number of ≥355 ms lapses on the 3-min PVT is shown. The two optimal decision thresholds divide SLST performance into high (≤11 lapses), medium (12–20 lapses), and low (>20 lapses) groups.

	Time of Day During 34 Hours of Sleep Deprivation																	
Subject #	09	11	13	15	17	19	21	23	01	03	05	07	09	11	13	15	17	
1	1	0	0	1	0	0	2	4	3	12	31	27	26	22	21	9	33	
2	3	6	6	4	11	14	19	18	35	35	30	16	25	19	21	11	12	
3	20	21	33	33	13	13	13	20	32	33	32	34	26	31	26	28	24	
4	1	6	2	3	4	1	1	1	3	6	10	23	22	15	9	7	14	
5	2	1	0	2	0	7	3	3	1	20	20	22	5	3	1	3	2	
6	3	4	2	8	3	2	3	12	15	9	23	18	12	11	18	1	7	
7	8	2	4	6	12	7	4	8	24	16	34	8	14					
8	11	9	18	22	9	10	33	10	26	35	30	26	20	18	30	30	13	
9	5	4	7	1	0	1	1	0	11	21	47	45	27	22	26	23	0	
10	4	2	4	6	2	3	1	8	16	21	16	11	10	21	7	8	3	
11	0	1	1	0	1	1	1	2	2	9	10	20	7	6	8	0	0	
12	2	1	1	1	3	0	7	9	8	30	19	11	3	9	5	2	6	
13	3	6	4	5	2	1	1	1	2	3	1	2	8	10	5	3	2	
14	1	1	5	4	2	8	4	0	3	8	15	16	1	1	6	1	1	
15	2	4	4	7	2	6	5	16	23	21	5	6	16	22	1	12	9	
16	21	20	23	29	22	13	23	24	20	18	33	33	46	35	33	38	13	
17	0	2	0	4	4	6	0	4	8	41	28	31	15	28	8	3	2	
18	1	1	1	0	1	0	1	1	1	1	3	8	20	5	3	2	4	
19	1	6	4	8	3	3	2	5	11	11	44	14	11	18	15	17	9	
20	4	2	6	8	11	2	5	14	9	12	17	12	19	14	14	21	11	
21	3	4	4	3	3	6	0	3	11	20	36	18	9	14	5	16	6	
22	8	5	3	13	8	5	6	16	18	10	13	13	23	13	2	21	6	
23	3	2	1	14	0	3	2	1	3	5	10	31	34	29	14	21	3	
24	11	5	8	13	6	7	6	2	11	9	6	16	17	14	14	17	11	
25	0	0	0	1	2	1	0	0	0	1	1	5	9	5	3	2	2	
26	12	7	8	15	10	15	11	19	22	33	33	35	36	23	16	12	14	
27	0	3	3	8	3	4	3	1	1	5	3	5	1	1	1	1	2	
28	9	4	7	11	5	8	5	15	6	4	13	17	22	26	7	29	1	
29	14	12	6	20	3	10	5	7	7	7	8	8	11	23	8	12	5	
30	18	41	26	14	30	7	10	13	14	44	30	36	30	35	19	30	25	
31	4	6	13	11	8	14	7	9	19	21	38	37	40	37	31	39	22	
32	10	15	13	11	10	19	12	2	25	10	8	16	19	27	25	16	24	
33	1	0	1	2	0	1	1	1	0	1	4	20	4	5	5	12	8	
34	1	2	0	2	1	1	1	0	0	0	3	21	7	2	2	3	1	
35	29	16	14	24	5	2	29	29	29	26	24	19	23	11	6	15	27	
36	19	9	12	8	12	3	6	9	6	20	24	27	20	20	22	22	19	
Average	6.5	6.4	6.8	8.9	5.9	5.7	6.5	8.0	11.8	16.1	19.5	19.6	17.7	17.0	12.5	13.9	9.7	
High	81%	83%	78%	72%	86%	83%	83%	69%	61%	47%	36%	25%	36%	34%	54%	43%	66%	
Medium	14%	11%	14%	17%	8%	17%	8%	25%	17%	19%	19%	36%	28%	26%	20%	26%	17%	
Low	6%	6%	8%	11%	6%	0%	8%	6%	22%	33%	44%	39%	36%	40%	26%	31%	17%	

Figure 6.

The number of ≥355 ms lapses on the 3-min PVT is shown for all 36 subjects and all 17 bouts during 34-h of total sleep deprivation. Classification based on these lapses is indicated by white (high SLST performance group, ≤11 lapses), gray (medium SLST performance group, 12–20 lapses), and black (low SLST performance group, >21 lapses) backgrounds. The average number of ≥355 ms lapses and the size of the three performance categories is shown depending on time of day during total sleep deprivation at the bottom of the figure.

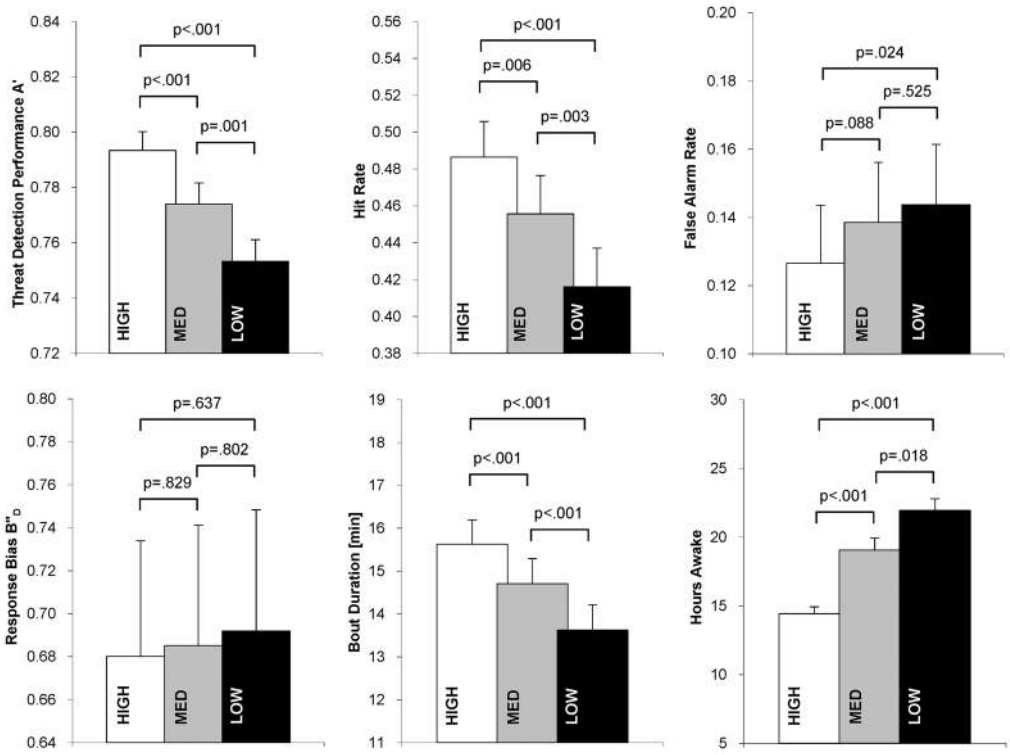

Figure 7.

Expected means and standard errors of threat detection performance A′, hit rate, false alarm rate, response bias B''_D, bout duration, and hours awake since wake-up time are compared between high, medium, and low SLST performance groups (group classification based on the number of ≥355 ms lapses on the 3-min PVT).

Table 1

Pearson's moment correlation coefficients representing coherence of threat detection accuracy A′ on the simulated luggage screening task (SLST) and ≥355 ms lapses on the 3 min PVT are given.

Subject	Pearson's rho
1	−0.457
2	−0.536*
3	−0.229
4	0.128
5	−0.409
6	0.309
7	−0.442
8	0.086
9	−0.416
10	−0.221
11	−0.528*
12	−0.632**
13	0.195
14	−0.548*
15	−0.274
16	−0.432
17	−0.701**
18	0.129
19	−0.750**
20	−0.142
21	−0.427
22	−0.627**
23	−0.897***
24	0.003
25	−0.067
26	−0.451
27	−0.097
28	−0.126
29	−0.193
30	−0.318
31	−0.377
32	0.272
33	−0.192
34	−0.263
35	0.007
36	−0.670**

Subject	Pearson's rho
Average of within subject coherences	−0.286
Coherence of between subject averages	−0.913***

*** P<.001,

** P<.01,

* P<.05

Table 2

Effects of night work, sleep loss, and time in study on threat detection performance in a simulated luggage screening task on N=24 healthy non-professional subjects (1). Point estimates of absolute changes and associated 95% confidence limits (in brackets) are given.

	Night Work (Night Work–Day Work)	Sleep Loss (Deprived - Rested)	Time in Study (Last Bouts - First Bouts)
Hit Rate	**−0.017** (−0.040, +0.006)	**−0.035**** (−0.061, −0.009)	**+0.076***** (+0.058, +0.094)
False Alarm Rate	**+0.025***** (+0.010, +0.039)	**+0.009** (−0.009, +0.026)	**+0.043***** (+0.030, +0.057)
Accuracy A′	**−0.023***** (−0.034, −0.012)	**−0.019**** (−0.030, −0.008)	**+0.003** (−0.006, +0.012)
Response Bias B''_D	**−0.049** (−0.100, +0.001)	**−0.005** (−0.067, +0.057)	**−0.179***** (−0.223, −0.135)
Bout Duration [s]	**−54.8***** (−73.2, −36.4)	**−19.1** (−38.4, +0.2)	**−60.3***** (−78.6, −42.1)

Night work: Day work after 8 h sleep (9 am to 7 pm) was compared to night work immediately following day work (9 pm to 7 am). Sleep loss: Day work (9 am to 5 pm) after 8 h sleep was compared to day work (9 am to 5 pm) after a night without sleep. Time in study: Day work after 8 h sleep (9 am to 9 pm) at the beginning of the study was compared to day work after 8 h sleep (9 am to 9 pm) at the end of the study. Mixed effects regression models with random intercepts and random slopes for bout number within each condition with unstructured covariance were used for comparisons between conditions (Proc Mixed, SAS version 9.1, SAS Institute Inc.). A variable indicating work bout number was included in the model, adjusting for time in study between conditions. For details see Basner et al. (1).

* $p<0.05$,

** $p<0.01$,

*** $p<0.001$ (H0: no difference between groups).