






Original Article

Quantifying the effects of sleep loss: relative effect sizes of the psychomotor vigilance test, multiple sleep latency test, and maintenance of wakefulness test

Thitaporn Chaisilprungraung^{1,†}, Emily K. Stekl¹, Connie L. Thomas^{1,2}, Margaux E. Blanchard¹, John D. Hughes¹, Thomas J. Balkin^{1,*} and Tracy J. Doty¹

¹Walter Reed Army Institute of Research, Silver Spring, MD, USA and

²Walter Reed National Military Medical Center, Bethesda, MD, USA

[†]The author is currently a researcher at the Learning Institute, King Mongkut's University of Technology Thonburi, Bangkok Thailand.

*Corresponding author. Thomas J. Balkin, Walter Reed Army Institute of Research, 503 Robert Grant Avenue, Silver Spring, MD 20910, USA. Email: thomas.j.balkin.ctr@health.mil.

Abstract

The psychomotor vigilance test (PVT) is a widely-used, minimally invasive, inexpensive, portable, and easy to administer behavioral measure of vigilance that is sensitive to sleep loss. We conducted analyses to determine the relative sensitivity of the PVT vs. the multiple sleep latency test (MSLT) and the maintenance of wakefulness test (MWT) during acute total sleep deprivation (TSD) and multiple days of sleep restriction (SR) in studies of healthy adults. Twenty-four studies met the criteria for inclusion. Since sleepiness countermeasures were administered in some of these studies, the relative sensitivity of the three measures to these interventions was also assessed. The difference in weighted effect size (eta-squared) was computed for each pair of sleepiness measures based on available raw test data (such as average PVT reaction time). Analyses revealed that the sleep measures were differentially sensitive to various types of sleep loss over time, with MSLT and MWT more sensitive to TSD than the PVT. However, sensitivity to SR was comparable for all three measures. The PVT and MSLT were found to be differentially sensitive to the administration of sleepiness countermeasures (drugs, sleep loss, etc.), but PVT and MWT were found to be comparably sensitive to these interventions. These findings suggest the potential utility of the PVT as a component of next-generation fatigue risk management systems.

Key words: Psychomotor Vigilance Test; Multiple Sleep Latency Test; Maintenance of Wakefulness Test; objective sleepiness; meta-analysis; relative sensitivity

Statement of Significance

The potential utility of the PVT as a component of a comprehensive fatigue risk management system depends, in part, on its sensitivity to sleep loss. In the present analyses, it was found that the MSLT and MWT are more sensitive to total sleep deprivation than the PVT, but that sensitivity to multiple nights of sleep restriction (a common problem in operational environments) was comparable for all three measures. It was also found that the PVT and MWT were comparably sensitive to sleepiness interventions (e.g. caffeine). These findings suggest the potential utility of the PVT as a component of next-generation fatigue risk management systems.

Introduction

Sleepiness in the operational environment constitutes a threat to operator performance [1–3] and safety [4, 5] in the short term, and operator health [6–8] in the longer term. Sleep disorders such as obstructive sleep apnea (OSA) have a high incidence among working-age adults with a worldwide estimate of 936 million people aged 30–69 years [9], and the danger to public safety posed by operators with OSA-associated chronic sleepiness is well known

[10]. Of course, sleepiness constitutes a danger in operational environments regardless of the cause of that sleepiness, and regardless of whether it is chronic or acute sleepiness. Because sleepiness resulting from random, situational causes (working a double shift, jet lag, etc.) is a universal experience, there is a longstanding and ongoing need to accurately identify individual operators who are significantly, even if only transiently, impaired due to sleepiness.

Prior to the introduction of the psychomotor vigilance test (PVT) by Dinges and Powell [11], those of us engaged in efforts to understand, measure, and mitigate the effects of inadequate sleep in the operational environment were in a quandary: Recognizing that it is impossible to manage, much less investigate, that which cannot be measured (G. Belenky, personal communication), the problem was that there existed no standard, sensitive, and logistically feasible way to objectively assess and quantify sleepiness in the operational environment.

Although the easiest way to assess sleepiness is to simply ask operators to rate their current level of sleepiness on a validated subjective rating scale (e.g. the Stanford Sleepiness Scale [12] or the Karolinska Sleepiness Scale [13]), it has been found that the sensitivity (and thus the accuracy and operational utility) of subjective sleepiness scales wane under conditions of chronic sleep restriction [14, 15]. Over time, operators can become subjectively inured to a chronically elevated level of sleepiness [14]. In addition, for a variety of reasons, operators may not be willing to accurately report their subjective sleepiness level [16, 17]. Therefore, while self reports of excessive sleepiness should always be taken seriously in the operational environment, subjective ratings that fall within the normal range cannot always be trusted to reflect objective reality. Therefore, objective measurement of sleepiness is a *de facto* requirement.

Since its introduction in 1977, the “gold standard” for objectively measuring sleepiness has been the multiple sleep latency test (MSLT) [18]. In this test, the latency to initiate sleep during multiple naps administered across the day under sleep-conducive conditions is assessed. Later, because of possible floor effect with the MSLT, the maintenance of wakefulness test (MWT) was introduced [19]. The MWT is similar to the MSLT, except that the individual being tested is instructed to remain awake (i.e. resist sleepiness)—typically while reclining in a comfortable chair in a dimly lit room for 20–40 min. Recognizing that the ability to resist sleep onset is more relevant to operational performance than the ability to initiate sleep—and consistent with the finding that the MWT is sometimes more sensitive to improvements in alertness following treatment of a sleep disorder [20]—it has been suggested that the MWT should be used to test and verify that an individual’s sleepiness level is not so high that his/her ability to function safely in operational environments is impacted [20, 21]. Although it has been argued that this is not an appropriate use of the MWT [22], it is currently one of the (several) measures used by the US Air Force [23] and US Navy [24] to recertify pilots who have been diagnosed with, and are being treated for, obstructive sleep apnea. Likewise, although it “remains controversial” the FAA has also sometimes requested MWT testing in commercial pilots [25].

The American Academy of Sleep Medicine (AASM) does not currently endorse using the MWT to help determine whether operators can safely perform their duties, taking the position that more research on the relationship between MWT results and actual operational performance is needed [26]. Indeed, the debate regarding whether or not the MWT should be used to inform decisions regarding the ability of individuals to function safely in operational environments reflects, in part, differing opinions regarding the extent to which the relevant scientific literature justifies this use of the MWT (see [21, 22]). Unfortunately for those presently tasked with determining whether pilots (or other operators) with sleep apnea are too sleepy to return to duty despite being treated with continuous positive airway pressure, the AASM does not currently endorse *any* measure for such an assessment. And for those presently charged with determining whether a

sleep apneic pilot being treated with CPAP should be allowed to resume flying, waiting for more research is not an option.

There have long been, and continue to be, many efforts to develop new measures of sleepiness, however none have been sufficiently validated as stand-alone predictors of operational performance and safety (for a review see Baiardi and Mondini [27]). Given this state of the science (i.e. a dearth of good options), and given that there is some scientific literature that supports using it for this purpose, it is not surprising that MWT data continue to be utilized to inform return-to-duty decisions, and can currently be considered a “less-than-gold standard” for this purpose.

However, measurement of SOL requires polysomnography and continuous monitoring by technical personnel who are proficient at identifying polysomnographic indicators of sleep onset in real time. This, along with the fact that these assessments are often, but not always, performed in conjunction with a prior full night of polysomnography [28], renders their widespread use to screen for sleepiness in operational environments both logistically and financially infeasible. Such assessments are therefore typically performed only when a sleep disorder has been identified or is suspected. And, obviously, another limitation is that MWT findings do not provide insight into the operationally meaningful day-to-day (or hour-to-hour) transient variations in sleepiness that can result from the occasional poor night of sleep (e.g. due to the need to care for a sick child). MWT findings only allow inferences about chronic (i.e. trait-like) sleepiness. Therefore, as stated in the AASM position paper [26]: “Ultimately, more time-efficient and cost-effective tests are needed to assess sleepiness and wakefulness. Especially valuable would be the development of fast and reliable field tests for sleepiness and alertness.”

Although it is not a measure of sleepiness *per se*, the PVT may fill this gap. It is relatively inexpensive, portable, minimally invasive (especially the 3-min version of the PVT [29]), objective, and highly sensitive to sleep loss [30]. And, although it does not generally correlate well with other performance measures [31], it is nonetheless an excellent tool for reflecting the effects of sleepiness on operational performance, for the following reasons:

Like the MWT, the PVT reflects the ability to sustain wakefulness (i.e. fight sleepiness). However, an added advantage (in terms of relevance to operational performance) is that it also reflects the extent to which an individual is able to perform a task while fighting sleepiness. The ability to maintain focus and engage in the performance of a task while resisting sleep is essentially what is required in many operational scenarios.

In addition, although virtually all operational tasks require some measure of sustained attention, few tasks require the same level of continuous, focused attention that the PVT requires. For example, it has been shown that experienced adult drivers take their eyes off the road to check the rearview mirror more often than inexperienced teen drivers—a practice that actually enhances situation awareness and safety [32]. In contrast, in order to maintain normal performance on the PVT (i.e. keep response times shorter than 0.5 s) it is necessary to continuously monitor the screen. The ability to continuously concentrate (maintain focus) for an extended period is especially sensitive to the detrimental effects of sleep loss [33]. This may be why PVT performance is more sensitive to sleep loss than most other behavioral measures [30], and why it is therefore logically a good potential predictor of sleep loss-induced performance deficits in a variety of operational environments. Pizza et al. [34] have shown that the driving performance (on a driving simulator) of sleep apneics accurately reflects sleepiness as measured by both the MSLT and

MWT, but especially by the MWT (as would be expected based on the previous discussion in which it was pointed out that the MWT is more closely aligned with operational task performance than the MSLT, because the former requires “fighting sleepiness” whereas the latter involves “surrendering to sleepiness”). Because performance on the PVT, like driving performance (or any other operationally-relevant task) requires that the effects of sleepiness be resisted, it is reasonable to hypothesize that PVT performance more closely reflects SOL on the MWT than on the MSLT. Consistent with this hypothesis, the primary objective of the present study was to determine the relative effect size of the PVT vs. the MSLT and MWT during sleep loss and in response to various experimental manipulations (see [Tables 1, 2 and 3](#)).

Methods

We identified studies in the literature in which both the PVT and a sleep latency test (MWT or MSLT) were administered to healthy adult participants, and for each study, computed the *difference in effect sizes* between the pairs of tests. The goal of the analysis was to determine whether the difference in effect size computed from two of these measures obtained in the same study, when averaged across all studies in the review, differed significantly from zero. Significant effect size differences would indicate that the PVT and sleep latency test were differentially sensitive to study manipulations. The following sections describe the methodologies used for acquiring the articles and for comparing effect sizes of test outcomes. This review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement guidelines (see [Supplementary Table SM 1](#) for PRISMA checklist). The review protocol was not registered.

Search strategy and selection

Article citations were generated via PubMed using keywords that directly contained the names of PVT and sleep latency tests: (“*maintenance wakefulness test*” OR “*MWT*” OR “*sleep latency test*” OR “*MSLT*” OR “*objective sleepiness*” OR “*psychomotor vigilance*”) AND “*sleep*”. The following criteria were used for selecting the articles: (1) peer-reviewed article published in English, (2) all participants were healthy adults (age > 18 years), (3) participants’ sleep (including short naps) was monitored in a sleep laboratory at one point throughout the course of the study, and (4) both PVT and MWT (or PVT and MSLT) were performed and test results (means and SD) were reported. Abstracts, reviews, book chapters, and case reports were excluded and no limit on publication time was imposed. The search was performed between March 23, 2021 and February 23, 2022 and generated 1900 citations. One reviewer screened the titles and abstracts of all articles against the inclusion criteria. If the abstracts contained insufficient information to determine eligibility, the reviewer performed full-text screening. For each article that was selected from the screening, a second reviewer verified the eligibility by performing another round of full-text screening. Any differences in the reviewers’ decisions were resolved by a third reviewer. All reviewers worked independently. After article screening, a total of 18 articles were selected for the review.

In addition to the search performed using test-specific keywords, we also utilized an additional database that contained a large corpus of studies where sleep manipulations (sleep deprivation, sleep restriction, nap) were performed in healthy adult participants. The additional database was compiled by entering the following keywords on PubMed: “*sleep loss*” OR “*sleep deprivation*” OR “*sleep restriction*” OR (“*nap*” AND (“*performance*” OR

“*emotion*”). We supplemented the original search with data from the additional database because studies are only identified on PubMed if the matching keywords are contained in the title or the abstract. As there are many studies where the PVT/MWT/MSLT are mentioned only in the paper’s body, a large number of studies were potentially missed in the original search. An article search for the additional database was carried out between June 2020 and February 2022 and generated 15 600 citations. To create the additional database, 12 research assistants performed full-text screening based on the following criteria: (1) the publication was peer-reviewed and published in English, (2) all participants were healthy adults, (3) participants’ sleep was monitored in the sleep research laboratory. For each article, at least two research assistants read through the entire article and verified whether all inclusion criteria were met. Any conflicts were resolved by a third reviewer. The screening process resulted in 1286 publications in the additional database. In addition to screening, the research assistants also extracted a variety of types of information from each paper, some of which were not relevant to the purpose of the current review (authors, publication year, PMID, number of participants, age, types of sleep manipulations, types of tests or interventions performed, screening methods, participants eligibility). Data concerning whether the PVT, MWT, and MSLT were performed were also collected in the additional database and were used as a basis for selecting articles for the present review. [Figure 1](#) depicts the selection process.

Of 1286 studies in the additional database, nine studies were found in which both the PVT and a sleep latency test (MWT or MSLT) were performed and raw test outcomes (means and SD) were available. Together with the original search results, 27 studies were selected for the review.

Post-hoc study screening

To ensure that the study data were obtained under comparable conditions, we imposed additional screening criteria to studies in our database. For the first analysis of the review, studies were further screened according to whether the PVT and MWT/MSLT data were obtained at similar times of day, and whether all subjects within a study contributed to all of the extracted datasets. For example, Vgontzas et al. [59] reported that their PVT results were based on a subset of 18 participants, whereas the MSLT results were based on the entire sample of 25 subjects, data from this study were therefore excluded from the analyses. In addition, we verified that all studies selected for the review involved independent samples of participants. As a result of this post-hoc screening, three studies were excluded from the review [59–61], resulting in 24 studies in the final analysis (see [Figure 1](#) and [Tables 1 and 2](#)).

In the second analysis of the review, studies were further screened and selected for the analysis if (1) either sleep restriction or total sleep deprivation was performed, and (2) the average test scores (and SD) for PVT and MWT and/or MSLT were available for different days or at different times-of-day during sleep loss. Of 24 studies selected for the review, 13 met both criteria and were included in the final analysis (see [Table 3](#)).

Data extraction

The following types of information were extracted from each study: (1) publication information (authors, publication year, PMID); (2) participants’ characteristics (age, N); (3) study design; (4) primary study intervention and experimental conditions; (5) type, length and days of sleep loss (if any); (6) MWT/MSLT sleep onset latency data (means and SD); and (7) raw outcome of PVT

Table 1. Table displaying characteristics of all studies where both PVT and MWT were performed

Study	Study design	Primary independent variable(s)	Experimental condition (N per cond)	Reported raw outcome	
				Sleep latency	PVT
Banks et al. [35]	Between-subject	Amount of recovery sleep	0 h-TIB (13); 2 h-TIB (27); 4 h-TIB (29); 6 h-TIB (25); 8 h-TIB (21); 10 h-TIB (27)	MWT latency to Microsleep	Lapse (>500 ms); Fastest 10% RT
Doty et al. [36]	Between	Caffeine	Caffeine 200 mg (24); Placebo (24)	MWT latency to Stage 1	Lapse (>500 ms); Speed (1/ RT × 1000)
Gasior et al. [37]	Between	Drugs (Lisdexamfetamine)	LDX 20 mg (27); LDX 50 mg (27); LDX 70 mg (27); Armodafinil 250 mg; (27); Placebo (27)	MWT latency to Stage 1	Median RT (least SQRT transformed)
Goel et al. [38]	Between	Genotype (PER3)	PER3 ^{4/4} (52); PER3 ^{4/5} (63); PER3 ^{5/5} (14)	MWT latency to Microsleep	Lapse (>500 ms)
Goel et al. [39]	Between	Cognitive workload and Sleep restriction	Moderate WL + SR (18); Moderate WL + NSR (11); High WL + SR (18); High WL + NSR (16)	MWT latency to Microsleep	Lapse; Mean 1/RT
Rupp et al. [40]	Between	Sleep extension	Extended (12); Habitual (12)	MWT latency to Microsleep	Lapse (>500 ms); Mean 1/RT
Rupp et al. [41]	Between	Social condition and personality type	Extravert/enriched (11); Introvert/enriched (13); Extravert/impoverished (12); Introvert/impoverished (12)	MWT latency to Stage 1	Transformed lapse (>500 ms); Mean 1/RT
Rupp et al. [42]	Between	Genotype (PRE3 and ADORA2A)	PER3 ^{4/4} (7); PER3 ^{4/5} (10); ADORA2A ^{CT/CT} (9); ADORA2A ^{TT/TT} (9)	MWT latency to Stage 1	Lapse (>500 ms); Mean 1/RT
Schweitzer et al. [43]	Between	Nap and caffeine	Nap (17); Caffeine (17); Nap and caffeine (17); Placebo (16)	MWT latency to Stage 1	Lapse (>500 ms, SQRT transformed)
Wright et al. [44]	Between	Caffeine and Light	Bright light caffeine (10); Bright light placebo (10); Dim light caffeine (9); Dim light placebo (9)	MWT latency to Stage 1	Mean RT
Wright et al. [45]	Between	Menstrual phase	Luteal (9); Follicular (8); Oral contraceptive (8)	MWT latency to Stage 2	Mean RT (difference from baseline)

(means and SD of speed and number of lapses). As raw test outcomes for PVT, MWT, and MSLT were frequently published only in the visual form (i.e. in a plot or a graph), an online webtool (WebPlotDigitizer V.4.5; <https://automeris.io/WebPlotDigitizer/>) [62] was used to extract means and SD of test results from the papers not reporting raw test outcomes in numeric form. The use of web-based tools to estimate data from graphs is common for systematic reviews [63, 64], and is recommended for research based on evidence concerning the reliability and validity of the extracted data [65]. One reviewer extracted each item of data from the paper. The accuracy of the extracted data was verified by a second reviewer. Any conflicts were resolved by a third reviewer.

Effect size computation

The present analyses were performed to determine whether the difference in effect size of two measures obtained from the same study, when averaged across all studies in the review, was

significantly different from zero. We first computed the effect sizes (eta-squared; η^2) of each test administered in the study (PVT and MWT or PVT and MSLT) using raw test data. For the PVT, raw test data corresponded to means and standard deviation (SD) of one or several outcome metrics (lapse, mean RT, median RT, mean 1/RT, fastest 10% RT, slowest 10% RT, PVT error) chosen for the study. For the MWT or MSLT, raw test data corresponded to means and SD of sleep onset latency. R package “rpsychi” was used to compute effect size estimates [66]. The effect size computation involved running a one-way ANOVA model to analyze the effect of the study interventions on each test outcome. Eta-squared was computed as the proportion of the effect’s sum of square to the total sum of square based on the output information reported in ANOVA.

$$\eta^2 = \frac{SS_{\text{Effect}}}{SS_{\text{Total}}}$$

The calculated effect size in each study reflects the proportion of variance in the data that is explained with the membership of

Table 2. Table displaying characteristics of all studies where both PVT and MSLT were performed

Study	Study design	Primary independent variable(s)	Experimental condition (N per cond)	Reported raw outcome	
				Sleep latency	PVT
Arnal et al. [46]	Within-subject	Sleep extension	Sleep extension (14); Habitual sleep (14)	MSLT latency to Stage1	Lapse (>500 ms); Speed (1/RT × 1000)
Belenky et al. [47]	Between	Sleep dose	3 h-TIB (18); 5 h-TIB (16); 7 h-TIB (16); 9 h-TIB (16)	MSLT latency to Stage1	Lapse (>500 ms); Speed (1/RT × 1000); Fastest 10% RT
Drake et al. [48]	Within	Sleep loss speed	Control, 8 h-TIB (12); Slow, 6 h-TIB (12); Intermediate, 4 h-TIB (12), Rapid, 0 h-TIB (12)	MSLT latency to Stage1	Lapse (>500 ms); Median RT;
Franzen et al. [49]	Between	Total sleep deprivation	SD (15); Non SD (13)	MSLT latency to Stage1	Lapse (>500 ms); Mean RT
Guilleminault et al. [50]	Within	Auditory stimulation	Stim without arousal (6); Stim with arousal (6)	MSLT (not reported)	Lapse (not reported); Mean RT;
Ikeda et al. [51]	Within	Type of awakening	Forced-awakening (11); Self-awakening (11)	MSLT latency to Stage1	Lapse (>500 ms); Mean RT; Fastest 10% RT
Lamond et al. [52]	Between	Combination of TSD and recovery	24 h SD + 9 h REC (10); 24 h SD+6 h REC (10); 48 h SD + 9h REC (10)	MSLT latency to Stage1	Lapse (>500 ms); Speed (1/RT × 1000); Fastest 10% RT
McBean et al. [53]	Within	Sleep fragmentation	Baseline in-lab (11); Post-fragmentation (11)	MSLT latency to Stage1	Lapse (not reported); Mean RT; Slowest 10% RT
Pejovic et al. [54]	Within	Period during sleep restriction	Baseline (30); Restriction (30); Recovery (30)	MSLT latency to Stage1	Lapse (not reported); Median RT; Fastest 10% RT; Slowest 10% RT
Roehrs et al. [55]	Mixed	Sleep restriction and ethanol	Ethanol 0.0 g/kg (20); Ethanol 0.3 g/kg (20); Ethanol 0.6 g/kg (20); Ethanol 0.9 g/kg (20); 8 h-TIB (12); 6 h-TIB (12); 4 h-TIB (12); 0 h-TIB (12)	MSLT latency to Stage1	Lapse (not reported); fastest 10% RT
Sauvet et al. [56]	Within	Exercise training	Pretraining (16); Posttraining (16)	MSLT latency to Stage1	Speed (1/RT × 1000); Errors ("RT <80 ms of >500 ms")
Walsh et al. [57]	Between	Modafinil	Modafinil (16); Placebo (16)	MSLT latency to Stage1	Lapse (>500 ms)
Walsh et al. [58]	Between	Gaboxadol	Gaboxadol (20); Placebo (21)	MSLT latency	Mean RT; Slowest 10% RT

different groups defined by the study's intervention. Eta-squared was preferred over standardized mean differences (Cohen's *d*) because the primary independent variable of a study was often comprised of multiple experimental conditions (e.g. 3 h vs. 5 h vs. 7 h vs. 9 h-TIB for sleep dose [47]). Whereas η^2 and other effect sizes in the correlation family (ω^2 , Pearson's *r*, etc.) allow for measuring the association strength among multiple experimental conditions in the study, standardized mean differences (Cohen's *d*) allow for comparison of average means only between two conditions. In some studies where multiple independent variables were present (e.g. light and caffeine), we constructed a separate ANOVA for each independent variable and reported all η^2 estimates for the study.

For the first analysis of the review, all raw data available in the article were used in the analysis except those data from baseline or pretreatment days, because these data were neutral with respect to our primary hypothesis. Some studies also included

a multiple night sleep-satiation phase prior to sleep loss ("sleep banking" studies) but there were too few of these studies to assess the effects of sleep satiation. In the second analysis, only data from control or placebo groups were used. This was in order to control for the effects of other manipulations (pharmacological or other interventions), and to ensure that the computed effect size most closely reflects the pure effect of sleep loss. For studies that included no control condition, data were selected from the experimental group that was hypothesized to most closely resemble the baseline condition.¹

¹ In particular, data for the Wright and Badia study [45] were selected from the 'follicular group' as it was hypothesized that the performance of this group was not affected by increased alertness due to rising body temperature (unlike the luteal and the oral contraceptive group). Similarly, PER3 4/5 and ADORA2AT/T groups were included from the Rupp et al. study [40] because these genotypes were hypothesized to be less resilient to sleep restriction.

Table 3. Table displaying characteristics of all studies included for sleep loss effect size analysis

Study	Study design	Primary independent variable(s)	Experimental condition (N per cond)	Reported raw outcome	
				Sleep latency	PVT
Doty et al. [36]	Sleep restriction (5 nights; 5 h-TIB)	SR1, SR2, SR3, SR4, SR5	Placebo (24)	MWT Stage1	Lapse (>500 ms); Speed (1/RT × 1000)
Goel et al. [39]	Sleep restriction (5 nights; 4 h-TIB)	SR1, SR4, SR5*	Moderate WL + SR (18)	MWT microsleep	Lapse; Mean 1/RT
Rupp et al. [42]	Sleep restriction (7 nights; 3 h-TIB)	SR1, SR2, SR3, SR4, SR5, SR6, SR7	PER3 ^{4/5} (10) and ADORA2A ^{T/T} (9)	MWT Stage1	Lapse (>500 ms); Mean 1/RT
Rupp et al. [40]	Sleep restriction (7 nights; 3 h-TIB)	SR1, SR2, SR3, SR4, SR5, SR6, SR7	Habitual (12)	MWT microsleep	Lapse (>500 ms); Mean 1/RT
Rupp et al. [41]	Total sleep deprivation (36 h)	2200, 0000, 0200, 0400, 0600, 0800, 1000, 1200, 1400, 1600	Introvert/impoverished (12)	MWT Stage1	Transformed lapse (>500 ms); Mean 1/RT
Wright and Badia [45]	Total sleep deprivation (24 h)	2130, 0030, 0330, 0630 [†]	Follicular (8)	MWT Stage2	Mean RT (difference from baseline)
Wright et al. [44]	Total sleep deprivation (45.5 h)	Night1: 2130, 0030, 0330, 0630 and night2: 2130, 0030, 0330, 0630 [†]	Dim light placebo (9)	MWT Stage1	Mean RT
Drake et al. [48]	Sleep restriction (4n, 6 h-TIB)	SR1, SR2, SR3, SR4	Slow, 6h-TIB (12) [§]	MSLT Stage1	Lapse (>500 ms); Median RT;
Ikeda et al. [51]	Sleep restriction (4n, 5 h-TIB)	1000, 1100, 1200, 1300, 1400, 1500, 1600	Self-awakening (11)	MSLT Stage1	Lapse (>500 ms); Mean RT; Fastest 10% RT
Walsh et al. [58]	Sleep restriction (4n, 5 h-TIB)	SR3, SR4 ^f	Placebo (21)	MSLT	Mean RT; Slowest 10% RT
Belenky et al. [47]	Sleep restriction (7n, 3,5-TIB)	SR1, SR2, SR4, SR4, SR5, SR6, SR7	3h-TIB (18) and 5h-TIB (16)	MSLT Stage1	Lapse (>500 ms); Speed (1/RT × 1000); Fastest 10% RT
Arnal et al. [46]	Total sleep deprivation (33h)	0000, 0300, 0600, 0930, 1300, 1600	Habitual (14)	MSLT Stage1	Lapse (>500 ms); Speed (1/RT × 1000)
Sauvet et al. [56]	Total sleep deprivation (33 h)	0200, 0600, 0930, 1300, 1600	Pre-training (16)	MSLT Stage1	Speed (1/RT × 1000); Errors ("RT <80 ms of >500 ms")

*MWT raw data not available for Day 2 and 3 (and was thus excluded from the analysis).

† For both studies, MWT (but not PVT) was also conducted at 0915 h.

§ Data from intermediate (4 h-TIB) and rapid (0 h-TIB) groups were not included due to limited time points available for PVT/MSLT.

f Raw data not available for SR1 and SR2.⁷

Difference in weighted effect size (ΔES_w)

To compare the sensitivity of test outcomes, we computed the difference in weighted effect size for each pair of test measures (PVT lapse vs. MWT sleep latency, PVT RT vs. MWT latency, PVT lapse vs. MSLT latency, PVT RT vs. MSLT latency). The weighting was performed to account for the variability in sample size and study design. Because the studies in our review recruited different numbers of participants and administered different numbers of experimental conditions, the η^2 estimate was weighed by *N per condition*. *N per condition* was derived by dividing the total number of subjects by the number of experimental conditions in between-subjects studies. For studies with a repeated-measures design, *N per condition* is the same as the total *N*. To preserve the original scaling, *N per condition* was further normalized to the scale of 0–1, with values closer to 1 denoting a higher sample size per experimental condition. The difference in weighted effect size (ΔES_w) was derived by subtracting the weighted effect size of the sleep latency test from the weighted effect size of the corresponding PVT in the study. The formula for computing ΔES_w for each study can be summarized as follow:

$$\Delta ES_w = (\eta_{PVT}^2 - \eta_{MWT,MSLT}^2) \times N / \text{condition}_{\text{scaled}}$$

All analyses and effect size computations were performed using R [67].

Risk of bias analysis

Two reviewers independently assessed the risk of bias using a revised Cochrane Risk of Bias assessment tool for randomized controlled trials (RCTs) [68]. The risk of bias was assessed at the level of individual studies within five domains: randomization process; deviations from the intended interventions; missing outcome data; measurement of the outcome; and selection of the reported result. Each domain was classified as low, some concerns, or high risk of bias based on the judgment of the reviewer and an overall risk of bias for each study was determined. Reviewer classifications were compared and discrepancies were resolved in agreement.

Results

Effects of primary independent variables

Of 24 studies included in the final analysis of the review, 11 studies utilized both PVT and MWT and 13 utilized both PVT and MSLT (see study characteristics in Tables 1 and 2). Studies utilizing data from PVT lapse (defined as RT > 500 ms; 18 of 24 studies, 75.0%) constitute the majority in our review. The second most common PVT metric was speed (mean 1/RT × 1000; 10 studies, 41.7%), followed by mean RT (6 studies, 25.0%), fastest

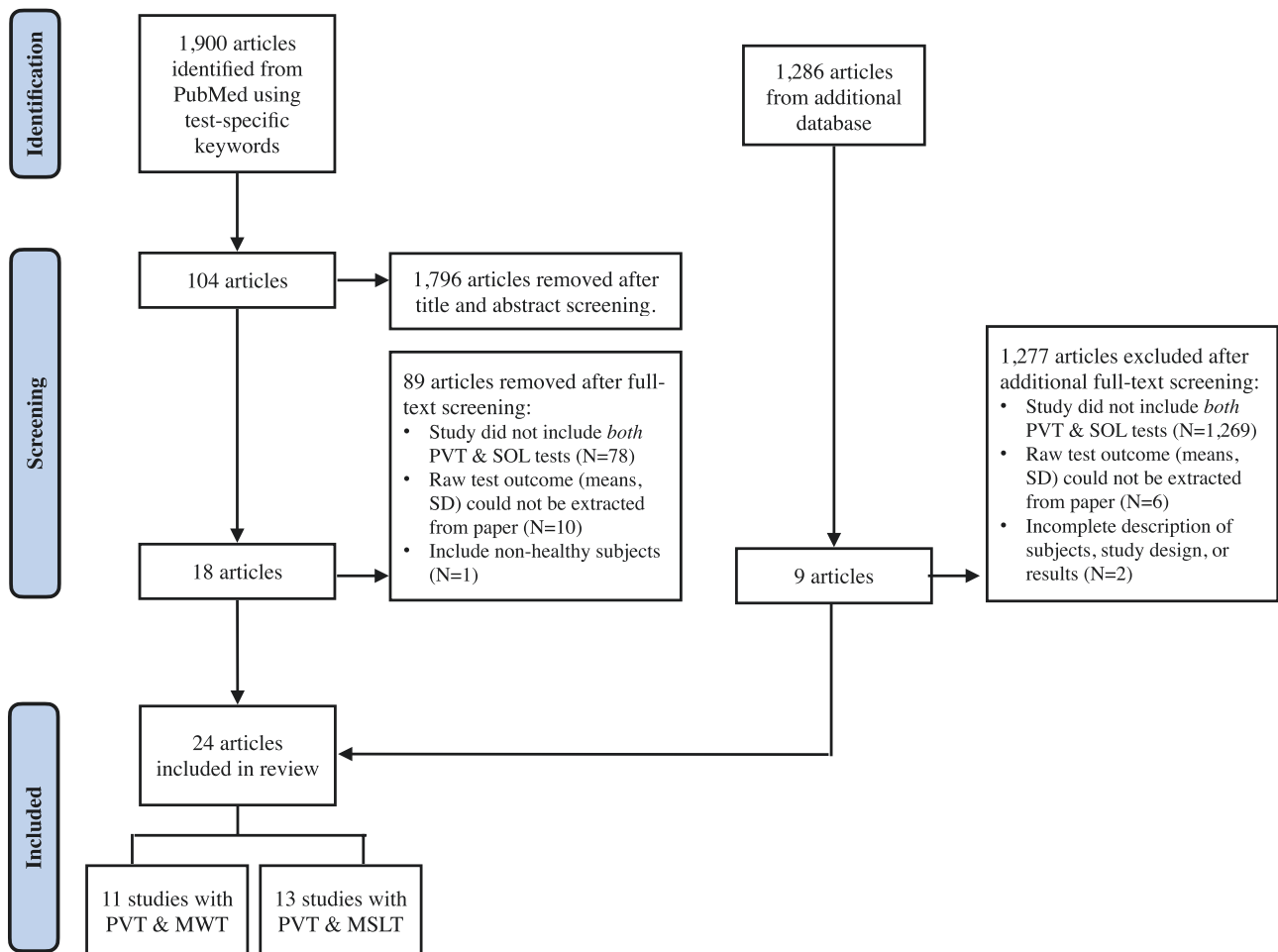


Figure 1. Flow chart illustrating the article selection process.

10% RT (6 studies, 25.0%), median RT (3 studies, 12.5%), slowest 10% RT (3 studies, 12.5%), and PVT error (defined in the paper as “RT < 80 ms of > 500 ms”; 1 study, 4.2%). Table 4 below shows the weighted effect size (η^2) for each test outcome, as well as the difference in weighted effect size calculated for each pair PVT vs. sleep latency test (for unweighted effect size data, see Supplementary Table SM 2).

To determine the relative sensitivity of PVT vs. sleep latency tests, we performed one-sample *t*-tests to analyze whether the average effect size difference (ΔES_w) was significantly different from zero. Due to the variety of PVT metrics available and the limited number of studies in our review, the analysis grouped together PVT data according to whether the metric corresponded to number of lapse or was derived from reaction time (i.e. mean RT, mean 1/RT, fastest 10% RT, slowest 10% RT, median RT; henceforth: “RT-based PVT”) to increase the power of the analysis. For this analysis, data from PVT error was excluded because it fit neither category and was available in only one study.

Figure 2 shows the bar plot of ΔES_w across different pairs of test outcomes. On the *x*-axis are the various comparisons (PVT lapse vs. MWT sleep latency, RT-based PVT vs. MWT latency, PVT lapse vs. MSLT latency, RT-based PVT vs. MSLT latency). The *y*-axis represents the average weighted effect size difference, with positive values denoting greater effect sizes for the PVT, and negative values denoting greater effect sizes for the sleep latency test. Across studies, the average (SD) of the weighted effect size difference is 0.006 (0.012) for PVT lapse vs. MWT, -0.005 (0.038) for

PVT RT vs. MWT, -0.023 (0.028) for PVT lapse vs. MSLT, and -0.024 (0.029) for PVT RT vs. MSLT.

One-sample *t*-tests revealed that ΔES_w was not significantly different from zero between PVT lapse vs. MWT ($t(11) = 1.60, p = .1$), or between PVT RT vs. MWT ($t(12) = -0.49, p = .6$). However, a significant difference in ΔES_w was found between PVT lapse and MSLT scores ($t(10) = -2.77, p = .02$), and between PVT RT and MSLT scores ($t(20) = -3.79, p = .001$). The finding therefore suggested that there was a greater similarity in outcome sensitivity between the PVT and MWT measures. The finding also revealed higher overall MSLT effect size compared to PVT, suggesting that the MSLT is more sensitive to study manipulations.

Effects of day or time-of-day during sleep loss

The present review included studies in which subjects were exposed to either total sleep deprivation or multiple nights of sleep restriction. A second set of analyses was performed to assess the relative effects of these procedures, and time of day effects during sleep loss—on MSLT, MWT, and PVT performance. It is well established that both the PVT and sleep latency tests are sensitive to sleep loss [69–72]. Less well understood, however, is whether, and the extent to which, different types of sleep loss differentially impact performance on the PVT, MWT, and MSLT—information that could be useful for determining the condition(s) under which each of these measures may be most useful. To

Table 4. Table showing the weighted effect size calculated for each outcome and each study, and the weighted effect size difference (ΔES_w) calculated between PVT and sleep latency test

Study	Primary IV	N per cond	Sleep latency test		PVT		Weighted ES difference (PVT-sleep latency)
			Type	Weighted ES	Outcome metric	Weighted ES	
Banks et al. [35]	Amount of recovery sleep	23.7	MWT	0.134	Lapse	0.129	-0.004
					Fastest 10%RT	0.136	0.002
Doty et al. [36]	Caffeine*	24	MWT	0.148	Lapse	0.145	-0.003
					1/RT × 1000	0.008	0.005
Gasior et al. [37]	Drugs (Lisdexamfetamine)*	27	MWT	0.148	Median RT	0.024	-0.124
Goel et al. [38]	Genotype (PER3)	43	MWT	0.005	Lapse	0.010	0.005
Goel et al. [39]	Sleep restriction	15.8	MWT	0.026	Lapse	0.014	-0.012
					1/RT × 1000	0.007	-0.020
	Cognitive workload	15.8	MWT	0.006	Lapse	0.001	-0.006
Rupp et al. [40]	Sleep extension*	12	MWT	0.000	Lapse	0.004	0.004
					1/RT × 1000	0.000	0.000
Rupp et al. [41]	Personality type	12	MWT	0.000	Lapse	0.001	0.001
					1/RT × 1000	0.001	0.001
Rupp et al. [41]	Social condition	12	MWT	0.000	Lapse	0.002	0.002
					1/RT × 1000	0.005	0.005
Rupp et al. [42]	Genotype (ADORA2A)	9	MWT	0.000	Lapse	0.016	0.016
					1/RT × 1000	0.018	0.018
	Genotype (PER3)	8.5	MWT	0.000	Lapse	0.031	0.031
Schweitzer et al. [43]	Caffeine*	16.5	MWT	0.006	Lapse	0.021	0.015
	Nap*	16.5	MWT	0.002	Lapse	0.023	0.021
Wright et al. [44]	Caffeine*	9.5	MWT	0.016	Mean RT	0.010	-0.006
	Light*	9.5	MWT	0.001	Mean RT	0.007	0.006
Wright and Badia [45]	Menstrual phase	8.3	MWT	0.004	1/RT × 1000	0.002	-0.003
Arnal et al. [46]	Sleep extension*	14	MSLT	0.014	Lapse	0.008	-0.006
					1/RT × 1000	0.004	-0.010
Belenky et al. [47]	Sleep dose	16.5	MSLT	0.042	Lapse	0.053	0.011
					1/RT × 1000	0.012	-0.030
					Fastest 10%RT	0.046	0.004
Drake et al. [48]	Sleep loss speed	12	MSLT	0.032	Median RT	0.004	-0.028
Franzen et al. [49]	Total sleep deprivation	14	MSLT	0.100	Lapse	0.041	-0.059
					Mean RT	0.035	-0.065
Guilleminault et al. [50]	Auditory simulation	6	MSLT	0.000	Mean RT	0.000	0.000
Ikeda et al. [51]	Type of awakening	11	MSLT	0.002	Lapse	0.003	0.002
					Mean RT	0.004	0.002
Lamond et al. [52]	Recovery sleep	10	MSLT	0.036	Lapse	0.000	-0.036
					1/RT × 1000	0.004	-0.032
					Fastest 10%RT	0.004	-0.032
Lamond et al. [52]	Total sleep deprivation	10	MSLT	0.000	Lapse	0.002	0.002
					1/RT × 1000	0.012	0.012
					Fastest 10%RT	0.006	0.006
McBean et al. [53]	Sleep fragmentation	11	MSLT	0.025	Lapse	0.004	-0.021
					Mean RT	0.003	-0.022
					Slowest 10%RT	0.000	-0.025

Table 4. Continued

Study	Primary IV	N per cond	Sleep latency test		PVT		Weighted ES difference (PVT-sleep latency)
			Type	Weighted ES	Outcome metric	Weighted ES	
Pejovic et al. [54]	Sleep restriction	30	MSLT	0.096	Lapse	0.026	-0.070
					Fastest 10%RT	0.024	-0.072
					Slowest 10%RT	0.014	-0.082
					Median RT	0.016	-0.080
Roehrs et al. [55]	Ethanol	20	MSLT	0.068	Lapse	0.015	-0.053
					Fastest 10%RT	0.041	-0.026
	Amount of total sleep deprivation	12	MSLT	0.059	Lapse	0.034	-0.025
					Fastest 10%RT	0.037	-0.022
Sauvet et al. [56]	Exercise training*	16	MSLT	0.001	1/RT × 1000	0.012	0.012
					Error	0.008	0.007
Walsh et al. [57]	Modafinil*	16	MSLT	0.022	Lapse	0.022	-0.001
Walsh et al. [58]	Gaboxadol	20.5	MSLT	0.008	Mean RT	0.000	-0.008
					Slowest 10%RT	0.000	-0.008

*Sleepiness countermeasures.

assess this, we analyzed whether the average weighted effect size difference ($\overline{\Delta ES_w}$) differed depending on whether total sleep deprivation (TSD) or chronic sleep restriction (SR) were performed in the study.

Of 13 studies included for the sleep loss analysis, seven studies utilized both PVT and MWT and six utilized both PVT and MSLT (see study characteristics in Table 3). Table 5 shows the weighted effect size for each dependent outcome, and the weighted effect size difference computed for each pair of outcomes (for numerical data corresponding to unweighted effect sizes, see Supplementary Tables SM 3).

A two-sample t-test was performed to determine whether the $\overline{\Delta ES_w}$ differed across TSD vs. SR studies. Due to the small sample size available for this analysis ($N = 28$), the weighted effect size differences were averaged types of PVT metrics (lapse, speed, meanRT, etc.) and types of sleep latency test (MWT and MSLT). The test revealed a significant difference in $\overline{\Delta ES_w}$ between TSD vs. SR studies ($t(8) = -3.55, p < .01, \overline{\Delta ES_w} = -0.06$ and 0.002 for TSD and SR studies respectively), suggesting that different types of sleep loss exerted differential impacts on the outcome sensitivity of PVT and sleep latency test (Figure 3).

One-sample t-tests were also performed to assess whether the $\overline{\Delta ES_w}$ significantly differed from zero. The tests revealed that the $\overline{\Delta ES_w}$ was not significantly different from zero for SR studies ($t(19) = 0.36, p = .7, \overline{\Delta ES_w} = 0.002$), suggesting that the PVT and sleep latency tests are equally sensitive to the effects of multiple days of sleep restriction. On the other hand, the $\overline{\Delta ES_w}$ obtained from TSD studies were significantly different from zero ($t(6) = -3.67, p = .01, \overline{\Delta ES_w} = -0.06$), suggesting that sleep latency tests are more sensitive than the PVT to acute total sleep deprivation (Figure 3).

We also tried repeating the same analyses on the data separated by types of sleep latency test. The test revealed no significant difference for either MWT or MSLT for SR studies (PVT vs. MWT: $t(7) = 1.36, p = .2, \overline{\Delta ES_w} = 0.001$; PVT vs. MSLT: $t(11) = -0.36, p = .7, \overline{\Delta ES_w} = -0.003$). For TSD studies, the test revealed a trend towards significance for PVT vs. MSLT ($t(3) = -2.5, p = .09, \overline{\Delta ES_w} = -0.07$) but a non-significance result for PVT vs. MWT ($t(3) = -2.3,$

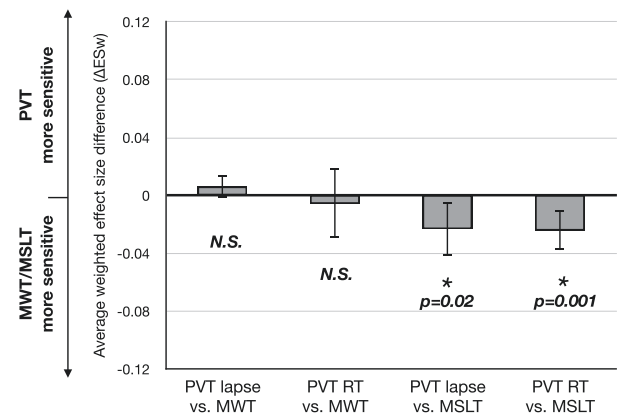


Figure 2. Bar plot showing the average weighted effect size difference ($\overline{\Delta ES_w}$) across studies for the PVT and sleep latency tests. Error bars indicate 95% confidence intervals for the mean. These findings suggest that in response to various study manipulations (drugs, sleep loss, etc.), the PVT and MWT's effect sizes are not significantly different from one another (i.e. $\overline{\Delta ES_w}$ not significantly different from zero). However, the MSLT effect sizes were significantly larger than the PVT effect sizes (i.e. $\overline{\Delta ES_w} > 0$), indicating that the MSLT is generally more sensitive to study manipulations.

$p = .1, \overline{\Delta ES_w} = -0.03$). It is noteworthy, however, that only four observations were available for analysis in the latter two tests, thus the absence of significance results was likely due to insufficient statistical power.

Risk of bias within studies

Supplementary Figure summarizes the risk of bias for all studies included. The proportion of studies with low risk of bias for each domain were as follows: randomization process (58%); deviations from the intended interventions (50%); missing outcome data (100%); measurement of the outcome (100%); and selection of the reported result (100%). Three studies [68, 69, 71] classified as high risk of bias for randomization were pre-post studies, where

Table 5. Table showing the weighted effect size calculated for each outcome and each study, and the weighted effect size difference (ΔES_w) calculated between PVT and sleep latency test

Study	Sleep loss type	N per cond	Sleep latency test		PVT		Weighted ES difference (PVT-sleep latency)
			Type	Weighted ES	Outcome metric	Weighted ES	
Doty et al. [36]	Sleep restriction (5n, 4 h-TIB)	24	MWT	0.019	Lapse	0.014	-0.005
					1/RT × 1000	0.041	0.022
Goel et al. [39]	Sleep restriction (5n, 4 h-TIB)	18	MWT	0.038	Lapse	0.031	-0.008
					1/RT × 1000	0.016	-0.022
Rupp et al. [40]	Sleep restriction (7n, 3 h-TIB)	12	MWT	0.007	Lapse	0.055	0.047
					1/RT × 1000	0.010	0.003
Rupp et al. [42]	Sleep restriction (7n, 3 h-TIB)	9	MWT	0.001	Lapse	0.026	0.025
					1/RT × 1000	0.025	0.024
Belenky et al. [47]	Sleep restriction (7n, 3 h-TIB)	18	MSLT	0.032	Lapse	0.121	0.089
					1/RT × 1000	0.007	-0.025
					Fastest 10%RT	0.034	0.002
	Sleep restriction (7n, 5 h-TIB)	16	MSLT	0.052	Lapse	0.016	-0.036
					1/RT × 1000	0.024	-0.028
					Fastest 10%RT	0.025	-0.028
Drake et al. [48]	Sleep restriction (4n; 6 h-TIB)	12	MSLT	0.001	Median RT	0.001	0.000
Ikeda et al. [51]	Sleep restriction (4n; 5 h-TIB)	11	MSLT	0.007	Lapse	0.008	0.001
					Mean RT	0.009	0.002
					Fastest 10%RT	0.002	-0.005
Walsh et al. [24]	Sleep restriction (4n; 5 h-TIB)	21	MSLT	0.010	Mean RT	0.004	-0.006
					Slowest 10%RT	0.003	-0.007
Rupp et al. [41]	Total sleep deprivation (22 h)	12	MWT	0.079	Lapse	0.022	-0.057
					1/RT × 1000	0.026	-0.053
Wright et al. [44]	Total sleep deprivation (45.5 h)	9	MWT	0.036	Mean RT	0.017	-0.019
Wright et al. [45]	Total sleep deprivation (24 h)	8	MWT	0.000	1/RT × 1000	0.000	0.000
Arnal et al. [46]	Total sleep deprivation (24 h)	14	MSLT	0.120	Lapse	0.035	-0.086
					1/RT × 1000	0.045	-0.076
Sauvet et al. [56]	Total sleep deprivation (40 h)	16	MSLT	0.166	1/RT × 1000	0.036	-0.130
					Error	0.173	0.007

random sequence generation was not necessary. Additionally, two subcomponents of the deviations from the intended interventions domain included blinding of the assigned intervention and appropriate analysis used to estimate the effect of assignment to intervention, which contributed to the three pre-post studies being classified as either high risk of bias or some concerns. For the 13 studies included in the second analysis, the proportion of studies classified as low risk of bias in each domain were: randomization process (46%); deviations from the intended interventions (77%); missing outcome data (100%); measurement of the outcome (100%); and selection of the reported result (100%).

General Discussion

The primary objective of this review was to compare the sensitivity to sleep loss (as reflected by effect size analyses) of the PVT vs. the MWT and MSLT. To accomplish this, we assessed the extent to which the difference in effect sizes of two tests (the PVT and either the MSLT or the MWT) obtained in the same study, when averaged across all studies included in the review, were

significantly different from zero. These analyses revealed that the relative sensitivity of the outcome measures varied as a function of the type of sleep loss, with measures of SOL (MSLT and MWT) more sensitive to acute total sleep deprivation than PVT measures. However, we also found that all of the measures were comparably (albeit somewhat less) sensitive to sleep restriction. These findings are consistent with those of Balkin et al. [30] who directly compared the sensitivities (using an effect size-derived statistic) of various tests of cognitive and psychomotor performance (including simulated driving, Stanford Sleepiness Scale, serial addition/subtraction, Stroop task, etc.) and measures of sleepiness (MSLT and PVT) across seven days of sleep restriction. In that study, it was found that of all tests administered, the PVT and the MSLT showed the highest, and comparable, sensitivities to sleep loss (sensitivity index;² 0.954 for PVT and 0.961 for MSLT).

A secondary aim of the present study was to compare the sensitivities (again, as reflected by effect size) of the PVT, MSLT,

² Defined in that study as the ratio of effect size to the 95% confidence interval.

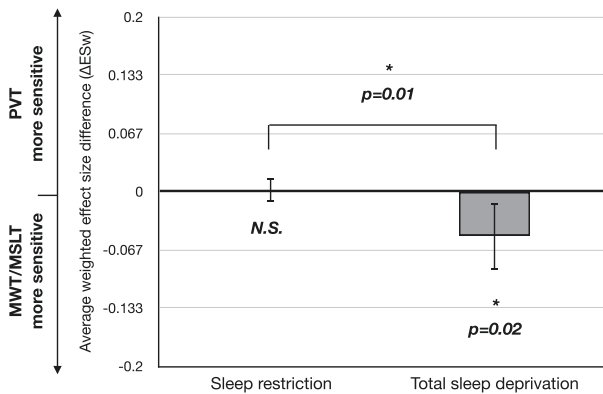


Figure 3. Bar plot showing the average weighted effect size difference (ΔES_w) across studies of sleep restriction and total sleep deprivation. Error bars denote 95% confidence intervals for the mean. This finding suggests that test sensitivity varies as a function of sleep loss type (acute total sleep deprivation vs. sleep restriction), with acute TSD affecting sleep latency tests more severely than the PVT ($t(9.3) = 3.06$, $p = 0.01$), whereas SR affects all tests equally (i.e. showing relatively limited sensitivity).

and MWT to various experimental interventions (i.e. sleepiness countermeasures like caffeine). These analyses revealed that the PVT and MWT (both of which involve resisting sleepiness) were more comparably sensitive to the administration of sleepiness countermeasures than were the PVT and MSLT. This may, at least in part, be because both the MWT and PVT require “resistance to sleepiness,” whereas the MSLT measures the effect of “withdrawal of resistance to sleepiness”. In other words, this finding suggests the possibility that the countermeasures administered in the studies included in the present review differentially affect the ability to resist sleepiness vs. the ability to initiate sleep.

The implications of the present findings include the following: The PVT and MWT are comparably sensitive to both sleep loss and to the application of sleepiness countermeasures, perhaps because both reflect the ability to resist the effects of sleepiness. Logically, the ability to resist sleepiness is important for maintaining performance and safety in operational environments. That being the case, it has been deemed necessary to assess this ability in some operators under some circumstances (e.g. sleep apneic military pilots who are being treated with continuous positive airway pressure). Despite some controversy regarding this use of the MWT [21, 22, 25], the MWT has, and continues to be, used for this purpose (albeit, not as the sole determinant, consistent with the recommendation by Littner et al. [28]).

Regardless of the scientific concerns regarding the advisability of using the MWT to inform decisions regarding the likelihood that individuals can safely perform their duties in operational environments (based to a significant extent on an opinion that more validation studies are needed [22]), it is widely recognized that the MWT is generally not well-suited to this purpose for other, practical reasons [26]. First, the MWT is logistically cumbersome and expensive (requiring at least one whole day of testing in a fully-equipped and staffed sleep laboratory or clinic). Second, when used to inform return-to-duty decisions, MWT findings (like MSLT findings) are presumed to reflect a relatively stable, trait-like characteristic of the individual being tested. But the extent to which findings from the MWT can or should be applied to predict operational performance over the ensuing

weeks, months, or years is unknown. Third, because of its aforementioned logistical requirements and expense, the large-scale studies that would be needed to validate the MWT as a predictor of safety within specific operational environments are difficult to conduct.

In sharp contrast, the PVT is portable and relatively easy to administer (with PVT apps for smartphones and other portable devices currently available [73–75]). Second, the PVT is only minimally invasive, with a 3-min version having been validated [73]. This means that it can be used to monitor state-like changes in sleepiness-mediated performance capability multiple times during each work shift. Lastly, because of its logistical advantages and low cost, the ability to perform large-scale validation studies in operational environments is relatively enhanced.

As indicated by Ferris et al. [76]: “The Psychomotor Vigilance Test (PVT) is considered the gold standard for detecting sleep loss and circadian misalignment related changes in performance in laboratory and field settings.” It is therefore not surprising that PVT performance is the primary outcome variable upon which several currently-available mathematical performance prediction models are based, such as the Unified Model of Performance [77]. One of the attributes of the PVT that makes it especially useful to this modeling effort is its considerable sensitivity to sleep loss—a level of sensitivity that most likely exceeds that of most measures of actual operational performance. performance prediction modeling is the likelihood that, for example, PVT speed (1/RT) has been found to be more sensitive to sleep restriction than measures of “lane deviation” or “lane position” on a driving simulator [30]. This constitutes a significant advantage of the PVT because in order for a measure to be a useful predictor of the deleterious effects of sleepiness on operational performance, it is logically necessary for that measure be more sensitive to sleep loss than the actual operational performance (and thus capable of revealing trends that provide an “early warning” of possible sleepiness-related deficits in operational performance).

The logistical considerations, combined with the present finding that the PVT’s sensitivity to sleep restriction and sleepiness countermeasures is comparable to that of the MWT, suggest that the PVT has potential as an acceptable measure of operator ability to maintain vigilance and thus perform safely in operational environments.

Our analysis also revealed that, unlike either of the SOL tests, the PVT was differentially sensitive to total sleep deprivation vs. sleep restriction. No significant differences between the MWT and MSLT were observed, perhaps because of insufficient statistical power due to small sample size. Another reason might have to do with the fact that effect sizes (η^2) computed in the first analysis reflected variations in test scores across TIB groups, whereas the effect sizes computed in the second analysis reflected variation in performance across days or times-of-day during sleep loss. It is possible that the day-to-day changes were too subtle for the effect size differences to be discernible. Nevertheless, the present findings do suggest that the outcomes from the PVT and sleep latency tests were more closely aligned during chronic sleep restriction than during acute total sleep deprivation (TSD).

Caveats to, and limitations of, the present review include the following: First, it should be noted that the calculated effect sizes (η^2) do not necessarily correspond to the reported outcomes of the selected studies (i.e. confirmation or rejection of the hypotheses tested in these studies). This is because our analysis methods involved the recalculation of η^2 from available raw PVT or sleep latency data. This recalculation necessitated assignment of a primary independent variable(s) for each study, and our assignments

did not necessarily comport with the original aims of the study. For instance, the goal of the Doty et al. [36] study was to identify whether, and the extent to which, the rate at which performance decline during SR differs between caffeine vs. placebo groups. We defined the primary independent variable for that study simply as “caffeine” (this was done to ensure that the way effect sizes were computed was consistent across studies). Thus, the computed effect sizes simply reflected variations in test scores across the caffeine vs. placebo conditions, and not whether any significant interaction between caffeine and days of SR was evident. Second, as previously mentioned, our review consisted of studies that differed considerably in experimental design and type of intervention applied. The heterogeneity of the dataset ($I^2 = 75.5\%$ calculated for the SOL measure of Analysis 1), though necessary as a part of our research question, made it difficult to obtain a meaningful direct comparison of effect sizes (η^2). There are also suggestions that the calculated η^2 is larger for within-subjects design than between-subjects design in studies where two groups of observations are positively correlated [78]. Our findings are therefore best interpreted as the differences in weighted effect size (ΔES_w), rather than η^2 , *per se*. Third, for many of the selected studies, raw test data for the PVT, MWT, MSLT were not available (either as numerical values or in graphic form). This limited the sample size of our review and the scope of possible analyses (number of publications excluded due to this problem = 16; 37.2%), and suggests that the relevant body of research is somewhat underrepresented in our dataset. However, a strength of the present review is that it only included studies in which PVT and MSLT or MWT data were collected within subjects, thus reducing potential confounding due to individual differences in sleep need, study differences in the amount of induced sleep pressure, and test environments.

Conclusion

The present review of 24 studies reveals that the PVT is just as sensitive to sleep restriction as both the MWT and the MSLT. It was also found that the PVT and the MWT are comparably sensitive to sleepiness countermeasures such as caffeine. On the basis of these findings, it is suggested that the PVT may constitute a sensitive and appropriate means of assessing and monitoring sleepiness-related variations in performance capacity—a capability that could improve performance and safety in a variety of operational environments. It is suggested that future research be focused on assessment of the predictive value of PVT measures for specific industries and occupations, and (depending on the findings from these studies) integration of the PVT into fatigue risk management systems.

Supplementary material

Supplementary material is available at *SLEEP Advances* online.

Acknowledgments

This material has been reviewed by the Walter Reed Army Institute of Research, and there is no objection to its presentation and/or publication. The opinions or assertions contained herein are the private views of the authors and are not to be construed as official or as reflecting the position of the Department of the Army or the Department of Defense. The authors wish to thank Sidney C. Allotey-Addo, Avery J. Denby, Maddison C. Pinner, Grace

R. Klosterman, Victoria R. Garriques, Claire E. Chanatry, Trevor J. Flick, Matthew E. Bohn, Jamari R. McCoy, and Anik Zingariello, for their research assistance in acquiring studies for this review as well as for their help in screening and selecting the studies. We also thank Dr. Paul Bliese, University of South Carolina, for the statistical advice and guidance he provided in support of this study.

Disclosure Statement

Financial Disclosure: TJB serves as a paid consultant to Echelon Med Tech LLC and Neteera. **Nonfinancial Disclosure:** None. The authors declare no competing interests.

References

1. Durmer JS, et al. Neurocognitive consequences of sleep deprivation. *Semin Neurol*. 2005;**25**(1):117–129.
2. Ochab JK, et al. Observing changes in human functioning during induced sleep deficiency and recovery periods. *PLoS One*. 2021;**16**(9):e0255771.
3. Smith MG, et al. Effects of six weeks of chronic sleep restriction with weekend recovery on cognitive performance and wellbeing in high-performing adults. *Sleep* 2021;**44**(8):zsab051.
4. Sadeghniiat-Haghighi K, et al. Fatigue management in the workplace. *Ind Psychiatry J*. 2015;**24**(1):12.
5. Thomas MJ, et al. Prior sleep, prior wake, and crew performance during normal flight operations. *Aviat Space Environ Med*. 2010;**81**(7):665–670.
6. Banks S, et al. Behavioral and physiological consequences of sleep restriction. *J Clin Sleep Med*. 2007;**3**(5):519–528.
7. Irwin MR. Why sleep is important for health: a psychoneuroimmunology perspective. *Annu Rev Psychol*. 2015;**66**:143–172.
8. Khan S, et al. Shiftwork-mediated disruptions of circadian rhythms and sleep homeostasis cause serious health problems. *Int J Genomics*. 2018;**2018**.
9. Benjafield AV, et al. Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis. *Lancet Respir Med*. 2019;**7**(8):687–698.
10. Balkin TJ. Assessment of the occupational sleep medicine field. In: Kryger MH, Roth T, Goldstein CA, eds. *Principles and Practice of Sleep Medicine-E-Book*. Philadelphia: Elsevier Health Sciences; 2021.
11. Dinges DF, et al. Microcomputer analyses of performance on a portable, simple visual RT task during sustained operations. *Behav Res Meth Instrum Comput*. 1985;**17**(6):652–655.
12. Hoddes E, et al. The development and use of the Stanford Sleepiness Scale (SSS). *Psychophysiol*. 1972;**9**:150.
13. Akerstedt T, et al. Subjective and objective sleepiness in the active individual. *Int J Neurosci*. 1990;**52**(1-2):29–37.
14. Van Dongen HPA, 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.
15. Silva EJ, et al. Circadian and wake-dependent influences on subjective sleepiness, cognitive throughput, and reaction time performance in older and young adults. *Sleep*. 2010;**33**(4):481–490.
16. Saletin JM, et al. Short daytime naps briefly attenuate objectively measured sleepiness under chronic sleep restriction. *Sleep*. 2017;**40**(9). doi:10.1093/sleep/zsx118
17. Tremaine R, et al. The relationship between subjective and objective sleepiness and performance during a simulated night-shift with a nap countermeasure. *Appl Ergon*. 2010;**42**(1):52–61.

18. Carskadon MA, et al. Sleep tendency: an objective measure of sleep loss. *Sleep Res.* 1977;**6**(200):940.
19. Mitler MM, et al. Maintenance of wakefulness test: a polysomnographic technique for evaluation treatment efficacy in patients with excessive somnolence. *Electroencephalogr Clin Neurophysiol.* 1982;**53**(6):658–661.
20. Sangal RB, et al. Maintenance of wakefulness test and multiple sleep latency test. Measurement of different abilities in patients with sleep disorders. *Chest* 1992;**101**(4):898–902.
21. Arand DL. The MSLT/MWT should be used for the assessment of workplace safety. *J Clin Sleep Med.* 2006;**2**(2):124–127.
22. Bonnet MH. The MSLT and MWT should not be used for the assessment of workplace safety. *J Clin Sleep Med.* 2006;**2**(2):128–131.
23. Force UA. *Air Force Waiver Guide.* US Air Force; 2016: 682–688.
24. Navy US. *US Navy Aeromedical Reference and Waiver Guide.* Washington, DC: US Navy. Published online 2008: 178–182.
25. Goldstein CA, Chervin RD. Use of clinical tools and tests in sleep medicine. In: *Sleep and Breathing Disorders E-Book.* Amsterdam: Elsevier Health Sciences. Published online 2016: 35–45.
26. Krahn LE, et al. Recommended protocols for the Multiple Sleep Latency Test and Maintenance of Wakefulness Test in adults: guidance from the American Academy of Sleep Medicine. *J Clin Sleep Med.* 2021;**17**(12):2489–2498. doi:10.5664/jcsm.9620.
27. Baiardi S, et al. Inside the clinical evaluation of sleepiness: subjective and objective tools. *Sleep Breath.* 2020;**24**(1):369–377.
28. Littner MR, et al. Practice parameters for clinical use of the multiple sleep latency test and the maintenance of wakefulness test. *Sleep.* 2005;**28**(1):113–121.
29. 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.
30. 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.
31. Van Dongen HPA, et al. Systematic interindividual differences in neurobehavioral impairment from sleep loss: evidence of trait-like differential vulnerability. *Sleep.* 2004;**27**(3):423–433.
32. Lee SE, et al. Eyeglance behavior of novice teen and experienced adult drivers. *Transp Res Rec.* 2006;**1980**(1):57–64.
33. Stawarczyk D, et al. Conjoint influence of mind-wandering and sleepiness on task performance. *J Exp Psychol Hum Percept Perform.* 2016;**42**(10):1587–1600.
34. Pizza F, et al. Daytime sleepiness and driving performance in patients with obstructive sleep apnea: comparison of the MSLT, the MWT, and a simulated driving task. *Sleep.* 2009;**32**(3):382–391.
35. Banks S, et al. Neurobehavioral dynamics following chronic sleep restriction: dose-response effects of one night for recovery. *Sleep.* 2010;**33**(8):1013–1026.
36. Doty TJ, et al. Limited efficacy of caffeine and recovery costs during and following 5 days of chronic sleep restriction. *Sleep.* 2017;**40**(12). doi:10.1093/sleep/zsx171
37. Gasior M, et al. Maintenance of wakefulness with lisdexamfetamine dimesylate, compared with placebo and armodafinil in healthy adult males undergoing acute sleep loss. *J Clin Psychopharmacol.* 2014;**34**(6):690–696.
38. Goel N, et al. PER3 polymorphism predicts cumulative sleep homeostatic but not neurobehavioral changes to chronic partial sleep deprivation. *PLoS One.* 2009;**4**(6):e5874.
39. Goel N, et al. Cognitive workload and sleep restriction interact to influence sleep homeostatic responses. *Sleep.* 2014;**37**(11):1745–1756.
40. Rupp TL, et al. Banking sleep: realization of benefits during subsequent sleep restriction and recovery. *Sleep.* 2009;**32**(3):311–321.
41. Rupp TL, et al. Socializing by day may affect performance by night: vulnerability to sleep deprivation is differentially mediated by social exposure in extraverts vs introverts. *Sleep.* 2010;**33**(11):1475–1485.
42. Rupp TL, et al. PER3 and ADORA2A polymorphisms impact neurobehavioral performance during sleep restriction. *J Sleep Res.* 2013;**22**(2):160–165.
43. Schweitzer PK, et al. Laboratory and field studies of naps and caffeine as practical countermeasures for sleep-wake problems associated with night work. *Sleep.* 2006;**29**(1):39–50.
44. Wright KP, et al. Combination of bright light and caffeine as a countermeasure for impaired alertness and performance during extended sleep deprivation. *J Sleep Res.* 1997;**6**(1):26–35. doi:10.1046/j.1365-2869.1997.00022.x.
45. Wright KP, et al. Effects of menstrual cycle phase and oral contraceptives on alertness, cognitive performance, and circadian rhythms during sleep deprivation. *Behav Brain Res.* 1999;**103**(2):185–194.
46. Arnal PJ, et al. Benefits of sleep extension on sustained attention and sleep pressure before and during total sleep deprivation and recovery. *Sleep.* 2015;**38**(12):1935–1943.
47. Belenky G, et al. Patterns of performance degradation and restoration during sleep restriction and subsequent recovery: a sleep dose-response study. *J Sleep Res.* 2003;**12**(1):1–12.
48. Drake CL, et al. Effects of rapid versus slow accumulation of eight hours of sleep loss. *Psychophysiology.* 2001;**38**(6):979–987.
49. Franzen PL, et al. Relationships between affect, vigilance, and sleepiness following sleep deprivation. *J Sleep Res.* 2008;**17**(1):34–41.
50. Guilleminault C, et al. The effect of CNS activation versus EEG arousal during sleep on heart rate response and daytime tests. *Clin Neurophysiol.* 2006;**117**(4):731–739.
51. Ikeda H, et al. Self-awakening improves alertness in the morning and during the day after partial sleep deprivation. *J Sleep Res.* 2014;**23**(6):673–680.
52. Lamond N, et al. The dynamics of neurobehavioural recovery following sleep loss. *J Sleep Res.* 2007;**16**(1):33–41.
53. McBean AL, et al. Effects of a single night of postpartum sleep on childless women's daytime functioning. *Physiol Behav.* 2016;**156**:137–147.
54. Pejovic S, et al. Effects of recovery sleep after one work week of mild sleep restriction on interleukin-6 and cortisol secretion and daytime sleepiness and performance. *Am J Physiol Endocrinol Metab.* 2013;**305**(7):E890–E896.
55. Roehrs T, et al. Ethanol and sleep loss: a “dose” comparison of impairing effects. *Sleep* 2003;**26**(8):981–985.
56. Sauvet F, et al. Beneficial effects of exercise training on cognitive performances during total sleep deprivation in healthy subjects. *Sleep Med.* 2020;**65**:26–35.
57. Walsh JK, et al. Modafinil improves alertness, vigilance, and executive function during simulated night shifts. *Sleep* 2004;**27**(3):434–439.
58. Walsh JK, et al. Slow wave sleep enhancement with gaboxadol reduces daytime sleepiness during sleep restriction. *Sleep* 2008;**31**(5):659–672.
59. Vgontzas AN, et al. Adverse effects of modest sleep restriction on sleepiness, performance, and inflammatory cytokines. *J Clin Endocrinol Metab.* 2004;**89**(5):2119–2126.
60. Horne J, et al. Sleep extension versus nap or coffee, within the context of “sleep debt.”. *J Sleep Res.* 2008;**17**(4):432–436.

61. Dinges DF, et al. Cumulative sleepiness, mood disturbance, and psychomotor vigilance performance decrements during a week of sleep restricted to 4-5 hours per night. *Sleep*. 1997;**20**(4):267–277.
62. Rohatgi A. *Webplotdigitizer: version 4.5*, 2021. <https://automeris.io/WebPlotDigitizer>. Accessed August 10, 2022.
63. Almasri J, et al. Outcomes of vascular access for hemodialysis: a systematic review and meta-analysis. *J Vasc Surg*. 2016;**64**(1):236–243.
64. Whited N, et al. Antibodies against SARS-CoV-2 in human breast milk after vaccination: a systematic review and meta-analysis. *Breastfeed Med*. 2022;**17**(6):475–483.
65. Drevon D, et al. Intercoder reliability and validity of WebPlotDigitizer in extracting graphed data. *Behav Modif*. 2017;**41**(2):323–339.
66. Okumura Y, Okumura MY. Package 'rpsychi.' Published online 2012. <http://www2.uaem.mx/r-mirror/web/packages/rpsychi/rpsychi.pdf>
67. Team RC. R: A Language and Environment for Statistical Computing. 2013. <https://cran.microsoft.com/snapshot/2014-09-08/web/packages/dplr/vignettes/xdate-dplr.pdf>. Accessed October 20, 2022.
68. Sterne JA, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ*. 2019;**366**.
69. Arand D, et al. The clinical use of the MSLT and MWT. *Sleep*. 2005;**28**(1):123–144.
70. Dorrian JF, et al. *Psychomotor Vigilance Performance: Neurocognitive Assay Sensitive to Sleep Loss*. In: Kushida CA, (ed.) *Sleep Deprivation: Clinical Issues, Pharmacology, and Sleep Loss Effects*. Boca Raton, FL: CRC Press; 2004.
71. Lim J, et al. Sleep deprivation and vigilant attention. *Ann N Y Acad Sci*. 2008;**1129**:305–322.
72. Rosenthal L, et al. Level of sleepiness and total sleep time following various time in bed conditions. *Sleep*. 1993;**16**(3):226–232.
73. Grant DA, et al. 3-minute smartphone-based and tablet-based psychomotor vigilance tests for the assessment of reduced alertness due to sleep deprivation. *Behav Res Methods*. 2017;**49**(3):1020–1029.
74. Matsangas P, et al. Hand-held and wrist-worn field-based PVT devices vs. the standardized laptop PVT. *Aerosp Med Hum Perform*. 2020;**91**(5):409–415.
75. Price E, et al. Validation of a smartphone-based approach to in situ cognitive fatigue assessment. *JMIR Mhealth Uhealth*. 2017;**5**(8):e125.
76. Ferris M, et al. The impact of shift work schedules on PVT performance in naturalistic settings: a systematic review. *Int Arch Occup Environ Health*. 2021;**94**(7):1475–1494.
77. Ramakrishnan S, et al. A unified model of performance: validation of its predictions across different sleep/wake schedules. *Sleep*. 2016;**39**(1):249–262.
78. Lakens D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Front Psychol*. Published online 2013;**863**.