

Comparative utility of instruments for monitoring sleepiness-related performance decrements in the operational environment

THOMAS J. BALKIN, PAUL D. BLIESE, GREGORY BELENKY, HELEN SING, DAVID R. THORNE, MARIA THOMAS, DANIEL P. REDMOND, MICHAEL RUSSO and NANCY J. WESENSTEN

Department of Behavioral Biology, Division of Neuroscience, Walter Reed Army Institute of Research, Silver Spring, MD, USA

Accepted in revised form 17 May 2004; received 23 December 2003

SUMMARY As both military and commercial operations increasingly become continuous, 24-h-per-day enterprises, the likelihood of operator errors or inefficiencies caused by sleep loss and/or circadian desynchrony also increases. Avoidance of such incidents requires the timely application of appropriate interventions – which, in turn, depend on the ability to measure and monitor the performance capacity of individuals in the operational environment. Several factors determine the potential suitability of candidate measures, including their relative sensitivity, reliability, content validity, intrusiveness and cumbersomeness/fieldability. In the present study, the relative sensitivity (defined as the ratio of effect size to 95% confidence interval) of several measures to the effects of sleep loss was compared in a sleep restriction experiment, in which groups were allowed 3, 5, 7, or 9 h time in bed (TIB) across seven consecutive nights. Of the measures compared, the Psychomotor Vigilance Test 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.

KEYWORDS effect size, fieldability, performance measures, sleep loss

INTRODUCTION

The burdens imposed by sleepiness-related accidents and sleepiness-mediated decrements in operational productivity are difficult to gauge (see Leger, 1995; Webb, 1995), but estimates suggest a staggering cumulative effect on both society and the economy (e.g. Leger, 1994). Occasionally, the effects have also been spectacularly catastrophic – with, for example, sleepiness associated with nighttime work implicated as a possible contributing factor in the accidents at Chernobyl, Bhopal and Three Mile Island (Rajaratnam and Arendt, 2001).

As both military and civilian industrial endeavors become increasingly continuous (24-h-per-day) operations, the potential for further sleepiness-related incidents – ranging from operational inefficiencies to serious accidents – increases accordingly. And with the burdens of human operators shifting away from the physical and towards the cognitive

(e.g. J. Biever, unpublished data), the need for effective monitoring of human cognitive performance is becoming increasingly important. But the task of determining how to exactly monitor human performance in the operational environment – especially the military operational environment for which few validated, operationally relevant performance metrics exist – is difficult and complex; a dilemma compounded by the fact that little is understood about the physiological underpinnings of sleep-loss-mediated deficits in performance capacity.

It has long been known that sleep loss negatively affects performance on a variety of tasks (with the first scientific investigation performed by Patrick and Gilbert, 1896), and those task characteristics that make them especially sensitive to sleep loss (e.g. long duration, inherently uninteresting, lack of feedback) have been delineated by Wilkinson (1965). Although it was at one time hypothesized that performance decrements following sleep loss were invariably ‘lapses’ in performance (i.e. the Walter Reed lapse hypothesis; Lubin, 1967), with these lapses perhaps signifying ‘microsleep’ events (brief episodes of sleep stage 1-like EEG activity; Guillemainault and Dement,

Correspondence: Thomas J. Balkin, Department of Behavioral Biology, Walter Reed Army Institute of Research, 503 Robert Grant Avenue, Room No. 2A26, Silver Spring, MD 20910, USA. Tel.: 301.319.9350; fax: 301.319.9979; e-mail: thomas.balkin@na.amedd.army.mil

1977), subsequent research clearly indicates that performance deficits are not invariably associated with such lapses – sleep loss-induced performance deficits often occur during objectively (polysomnographically) defined wakefulness (e.g. Balkin *et al.*, 2000; McCarthy and Waters, 1997).

As sleepiness is presumed to be the intervening variable accounting for sleep loss-related performance deficits, one approach for management of performance in the operational environment would be to objectively measure sleepiness, and then infer the impact of that level of sleepiness on performance in the operational environment. The multiple sleep latency test (MSLT; Carskadon *et al.*, 1986) is currently the ‘gold standard’ for measurement of sleepiness, but (a) the MSLT is too cumbersome to be of practical use in the operational environment (requiring a quiet, dark bedroom, the capability to conduct polysomnographic monitoring, etc.), and (b) it has been reported that a substantial proportion of normal research subjects – that is, individuals with no complaints or apparent difficulties with daytime functioning – nevertheless have average sleep onset latencies of less than 5 min (Harrison and Horne, 1996), which is in the ‘pathological’ range (Richardson *et al.*, 1978). This suggests that individuals vary with respect to ‘sleepability’, i.e. their ability to initiate sleep regardless of extant sleep debt (Harrison and Horne, 1996). Thus, to the extent that an individual’s sleepability quotient is high, the MSLT would be inappropriate for drawing conclusions about performance capacity.

The Maintenance of Wakefulness Test (MWT; Mitler *et al.*, 1982) might be better for this purpose because it measures the ability of an individual to resist sleep onset under sleep-conducive conditions – an ability with obvious implications for performance capacity in many operational environments. However, the MWT is just as cumbersome as the MSLT, and therefore just as impractical for use in the operational environment.

Other measures of sleepiness based on physiological changes with sleep loss (pupillometric, oculomotor, eyelid closure, EEG-based measures, etc.) would seem promising based on their possible imperviousness to competing factors such as motivation and the likelihood that many such measures could be obtained passively and unobtrusively, but these measures (a) are generally not validated adequately, (b) often require specialized and proprietary hardware/software, and (c) depending on the nature of the measure, some maintenance in the field may be required (e.g. adding gel to, or re-attachment of, electrodes).

Another approach would be to make predictions of operational performance based on objective measures of prior sleep. For example, sleep duration and timing can be measured in the home environment with wrist actigraphy. As this requires only that the individual wear a wristwatch-like device (the data from which could be telemetered for automatic analyses), the intrusiveness (at least with respect to interruptions of operations) would be minimal. However, (a) the parameters of the algorithm for predicting performance from actigraphically measured sleep history would need to be individualized to take

into account the considerable individual differences in susceptibility/resistance to the effects of sleep loss (which may change over time), and (b) because the device essentially entails monitoring of behavior outside the operational environment, compliance may be problematic in some (especially non-military) groups of operators.

One approach with considerable promise is that of ‘embedded performance measures’ (e.g. lane-tracking devices on trucks) because these could be used to monitor relevant performance in the actual operational environment, allowing automatic detection of performance deterioration (and thus facilitating intervention with effective countermeasures) well before catastrophic performance failure occurs. However, not all jobs in the operational environment are amenable to monitoring with embedded performance measures. In fact, in the military operational environment, only a subset of relevant tasks would be amenable to development of such embedded measures – and this would, in any event, be a daunting endeavor.

Therefore, as a first step toward determining which of the many available measures constitute a promising metric of general sleep-related performance capacity *for use in the operational environment*, we tested, compared, and judged several candidate measures across seven consecutive days that were preceded by 9, 7, 5 or 3 h in bed per night, using the following criteria (which, although perhaps not exhaustive nor completely orthogonal, were deemed critical for serious consideration).

Sensitivity

Of primary importance is that the selected measure be sensitive to the effects of sleep loss (with, for example, no restriction of range because of floor/ceiling effects). Optimal sensitivity means that the measure could, at least potentially, be ‘mapped onto’ specific, operationally relevant tasks that are themselves less sensitive to the effects of sleep loss. Thus, for example, if the selected performance measure was decremented by four ‘just noticeable differences’ (JNDs) with every hour of sleep loss, while the operationally relevant task was decremented by just one JND for every hour of sleep loss, there would be the potential to use the selected measure to predict performance on the operationally relevant task. (It should be noted, however, that the validity of this sort of cross-task mapping would also depend on other factors – such as the extent to which the sleep loss-induced performance degradation curves of the predictor and predicted tasks are similarly shaped.)

Reliability

This is the extent to which the selected measure is ‘repeatable’ and, for example, not subject to ‘learning’ effects (i.e. performance does not improve as a function of repeated administration). Thus, the ‘Tower of Hanoi’ problem would not be amenable for use in the operational environment, because even if this measure was exquisitely sensitive to the effects of sleep

loss initially, it would be expected to be relatively insensitive on subsequent administrations after the solution of the problem had been learned (Ahonniska *et al.*, 2001).

Content validity

In the present context, this is the specificity with which the selected measure reflects sleep history-related changes in performance capacity (and, conversely, the extent to which the selected measure is relatively impervious to the influence of other factors that, although affected by sleep loss, would not be expected to significantly impact general cognitive performance capacity). Although it is not possible to quantify the relative content validity of the various measures (i.e. the extent to which fluctuations in these measures reflect fluctuation in general cognitive performance capacity), it is possible to rate purely physiological measures such as blood pressure as having relatively low content validity in comparison with the psychomotor tasks in which, for example, response speed is measured – because ‘cognitive processing speed’ is a relevant determinant of performance in a wide array of operationally relevant tasks.

Intrusiveness

From a practical standpoint, it is essential that the selected measure be non-intrusive or only minimally intrusive (i.e. measurement procedures result in minimal/inconsequential interference in the performance of operationally relevant tasks). In this respect, purely psychophysiological measures such as ‘eyelid closure speed’ would be preferable to psychomotor performance tasks – especially if the measurements could be obtained passively and unobtrusively (i.e. like an embedded performance measure).

Fieldability

Measurement procedures must also be compatible with the operational environment (e.g. requiring minimal expenditures in terms of cost, time, individual effort required for data collection; and ease with which data can be compiled, processed and interpreted). Like intrusiveness, this factor represents a practical consideration that mediates the potential utility of a measure in the operational environment, and is akin to the ‘Ease of Use’ factors identified previously by Dinges and Mallis (1988).

Of these factors, only the first two (sensitivity and reliability) can be objectively quantified. The remaining factors (content validity, intrusiveness and cumbersomeness) can be subjectively rated, but the extent to which these factors might individually or synergistically impact the utility of a particular measure in the operational environment might best be determined empirically. Thus, the primary focus of the present paper is to compare the relative sensitivity and reliability (as defined previously) of an array of measures commonly used in laboratory-based sleep deprivation studies.

METHODS

General design and procedures

A complete description of the study subjects, design and procedures can be found in Balkin *et al.* (2000). Briefly, 66 commercial motor vehicle licensed drivers (16 women, 50 men; age range 24–62 years) participated. They spent 14 days in the Johns Hopkins Bayview General Clinical Research Center (Baltimore, MD, USA) (see Table 1). The first 2 days were adaptation/training (T1, T2) and the third served as baseline (BL); and subjects were in bed from 23:00–07:00 hours [8 h time in bed (TIB)] on these days. Beginning on the fourth day and continuing for a total of 7 days (E1–E7), subjects were placed on one of four schedules: 9 h TIB (22:00–07:00 hours; $n = 16$); 7 h TIB (24:00–07:00 hours; $n = 16$); 5 h TIB (02:00–07:00 hours; $n = 16$), or 3 h TIB (04:00–07:00 hours, $n = 18$). On the 11th day and continuing for a total of three ‘recovery’ days (R1–R3), subjects were again allowed to sleep from 23:00–07:00 hours (8 h TIB). Data from recovery days are not presented here, but a description and discussion of subsets of these data can be found in the reports by Belenky *et al.* (2003) and Russo *et al.* (2003).

Tests administered

Subjects performed a series of cognitive and alertness tests daily (see Table 2). The following tests were administered at least 4 times per day: Psychomotor Vigilance Task (Dinges and Powell, 1985) for 10 min per session; synthetic work task (Elsmore, 1994) for 15 min per session; simulated driving for *c.* 45 min per session (see Balkin *et al.*, 2000); and the Walter Reed Performance Assessment Battery (see Thorne *et al.*, 1985), which included the following tasks, in order of administration: Stanford Sleepiness Scale (SSS; Hoddes *et al.*, 1973) (one trial, *c.* 10-s duration); Profile of Mood States (65 items, *c.* 2-min duration); code substitution (54 trials; *c.* 2-min duration); serial addition/subtraction (60 trials; *c.* 3-min duration); grammatical (logical) reasoning (32 trials, *c.* 2-min duration); running memory (48 trials, *c.* 2-min duration); time estimation (interval reproduction) (one trial, *c.* 30-s duration); 10-Choice reaction time (RT) (60 trials, *c.* 2-min duration); Stroop color naming (48 trials, *c.* 2-min duration), and delayed recall (nine items, *c.* 1-min duration). It should be noted that the tasks comprising the PAB were terminated by trial count rather than duration, so the listed task durations are approximate. A 4-Choice RT (*c.* 8 min) test (Thorne *et al.*, 1985) and a modified sleep latency test (SLT) (Carskadon *et al.*, 1986) were administered twice daily [modifications were (a) that the nap trials were administered only twice per day (instead of the typical 4+ trials), and the criterion for ending the test (awakening the subject) was the first appearance of a sleep spindle or K-complex (i.e. subjects were not awakened after three consecutive 30-s epochs of stage 1 sleep)]. An oculomotor function test [fitness impairment tester (FIT)] was administered six times per day (*c.* 45 s per administration). Details of these tests can be found in Balkin *et al.* (2000).

Table 1 Study schedule

Day	Preceding hours of time in bed	TIB schedule (hours)
Training 1 (T1)	8	23:00–07:00
Training 2 (T2)	8	23:00–07:00
Baseline (BL)	8	23:00–07:00
Experimental 1 (E1)	3, 5, 7, or 9	04:00–07:00, 02:00–07:00, 00:00–07:00, or 22:00–07:00
Experimental 2 (E2)	3, 5, 7, or 9	04:00–07:00, 02:00–07:00, 00:00–07:00, or 22:00–07:00
Experimental 3 (E3)	3, 5, 7, or 9	04:00–07:00, 02:00–07:00, 00:00–07:00, or 22:00–07:00
Experimental 4 (E4)	3, 5, 7, or 9	04:00–07:00, 02:00–07:00, 00:00–07:00, or 22:00–07:00
Experimental 5 (E5)	3, 5, 7, or 9	04:00–07:00, 02:00–07:00, 00:00–07:00, or 22:00–07:00
Experimental 6 (E6)	3, 5, 7, or 9	04:00–07:00, 02:00–07:00, 00:00–07:00, or 22:00–07:00
Experimental 7 (E7)	3, 5, 7, or 9	04:00–07:00, 02:00–07:00, 00:00–07:00, or 22:00–07:00
Recovery 1 (R1)	8	23:00–07:00
Recovery 2 (R2)	8	23:00–07:00
Recovery 3 (R3)	8	23:00–07:00

Analytic strategy to determine relative sensitivity of measures

To determine the relative sensitivity to sleep restriction of each outcome measure, the effect-size of the relationship between total sleep time (TST) and each outcome measure [psychomotor vigilance test (PVT), RT, SLT, etc.] was computed. Effect-size estimates are the building blocks of meta-analyses, and allow comparisons of findings across a broad range of study conditions (Cohen, 1988; Rothstein *et al.*, 2002). In the present study, effect size estimates formed the basis for comparing the relative sensitivity (to sleep restriction) of a wide variety of measures.

The effect-size was estimated using the R^2 change between two models. The first model regressed mean performance during the experimental phase on (a) baseline performance and (b) mean TST during the baseline period. The second model regressed mean performance during the experimental phase on three variables: (a) baseline performance, (b) mean TST during the baseline period, and (c) mean TST during the experimental period. Thus, the difference between the two models is the inclusion of mean TST during the experimental phase in model 2. The R^2 difference between the two models represents the effect of sleep restriction on the outcome of interest after controlling for individual differences in both baseline performance and baseline sleep. The two separate models and the

resulting differences in R^2 values were computed for each of the outcome measures (PVT, SLT, etc).

As with a typical meta-analysis, it was critical to estimate the variability associated with the effect-size to ensure (among other things) that the computed effect size for each measure differed significantly from zero (Rothstein *et al.*, 2002). Although there are a number of ways to estimate effect size variability, one of the most versatile methods (as it requires few *a priori* assumptions about the data) is the bootstrap.¹ This procedure provides highly accurate estimates of confidence intervals (Efron and Tibshirani, 1993; Hall, 1988; Tibshirani, 1988), and because the procedure can be used with very few assumptions about the characteristics of the data, it is applicable to data that are non-normally distributed as well as to data that are normally distributed. The versatility of the bootstrap was important in the present study because the nature of the performance measures, and the shapes of the distributions of the outcome data from these measures, differed considerably.

Upper and lower 95% confidence intervals of the effect-size were based on (a) the variability of 25 bootstrapped samples replicated, (b) 1000 times for a total of 25 000 bootstrapped samples. It should be noted that although effect-size changes between models 1 and 2 were positive (adding new variables to a model cannot reduce the explained variance), the confidence interval limits could take on negative values because they were based on the variability of the R^2 values. As described in Results, determination of the relative sensitivities of the measure was dependent on both their effect sizes and their confidence intervals – because some outcomes with moderately large effect sizes had high variability (indicating that the effect sizes did not differ from chance levels). In contrast, other outcomes with relatively small effect sizes but with low variability did differ significantly from chance levels. Thus, sensitivity was operationally defined in the present study as the ratio of the effect size of an outcome measure to its 95% confidence interval (i.e. effect size divided by the confidence interval) – producing a unitless index with which outcome measures were compared.

¹Bootstrap estimates of variability are based on the following procedure. First, a sample is drawn with replacement from the original data. The number of observations drawn equals the number of observations in the original sample. For instance, if a sample contains data from 66 participants, the bootstrap will draw a sample of 66; however, since replacement is used, participant 1 may be selected 5 times, while participant 2 may not be selected at all. The statistic of interest (R^2 in the present case) is calculated on this sample and recorded. This procedure is done a small number of times (e.g. 25 times), and the variability of the statistic across the 25 samples is calculated and recorded. Ultimately this variability is used to calculate confidence intervals; however, to improve the estimates of the confidence intervals the small samples (e.g. 25) are drawn 1000 times or more and the final variance estimates are based upon a very large number of randomly drawn samples (25,000 in our example).

Table 2 Daily test administration schedule

Test [duration (min)]	TIB groups								
	All (9, 7, 5, 3)		7, 5, 3*		3*				
Vitals (5)	07:30	10:30	13:30	16:30	19:30	21:30	22:20	00:50	02:50
FIT (0.75)	07:30	10:30	13:30	16:30	19:30	21:30	22:20	00:50	02:50
STISIM (45)	07:40	10:40	13:40		19:40		22:30	01:00	03:00
PAB (15)		09:00	12:00	15:00		21:00			
SYN (15)		09:15	12:15	15:15		21:15			
PVT (10)		09:30	12:30	15:30		21:30			
SLT (20 max) [†]		09:40/10:05		15:40/16:05		21:40/22:20			
PAB 2 (10) [†]		10:05/09:50		16:05/15:50					
ORG (30 max)				16:45					
PAB 3 (10)								00:00	02:00
PVT (10)								00:10	02:10
Meals	08:30			12:40	17:30	23:15			
Shower					18:00				

*Experimental days 1–7 only.

[†]Vertical slash (/) indicates alternation between subject pairs.

FIT, Fitness Impairment Tester; STISIM, Systems Technologies Inc. Simulator; PAB, Performance Assessment Battery; SYN, Synthetic Work Task; TIB, Time in Bed; PVT, Psychomotor Vigilance Task; ORG, Organizational Task.

RESULTS

Nightly total sleep time

Night-time sleep data were analyzed using a two-way mixed ANOVA for sleep group (3, 5, 7, or 9-h TIB) and day (BL, E1, E2, E3, E4, E5, E6, E7, R1, R2, and R3), with repeated measures on the latter factor. Mean TST (sum of stages 1, 2, SWS and REM) increased significantly in the 9-h group and decreased significantly in the 3, 5, and 7-h groups across the sleep restriction/augmentation phase (E1–E7) compared with BL (group \times night, $F_{(30,610)} = 141.83$, $P = 0.0000$). TST significantly differed among all sleep groups on nights E1–E7 (Tukey HSD, $P < 0.05$), with the 3, 5, 7, and 9-h groups averaging 2.87, 4.66, 6.28, and 7.93 h of sleep, respectively. There were no group differences on the baseline night (Tukey HSD; $P > 0.05$).

Performance and alertness measures – sensitivity

Table 3 lists the various outcome measures from the present study, rank ordered as a function of sensitivity to sleep restriction. Listed for each measure are (a) the effect size, (b) the upper 95% confidence limit, (c) the lower 95% confidence limit, (d) the confidence interval (i.e. the upper minus the lower 95% confidence limit), (e) an indication of whether the effect size differs significantly from zero (i.e. whether there was an overall effect of sleep restriction on the measure of interest), and (f) the sensitivity index, defined here as the effect size divided by the confidence interval.

As shown in Table 3, the largest effect size was evident for the SLT at 0.45 – more than twice that of the next largest effect size (0.21) evident for the PVT. However, because variability (i.e. during baseline) on the SLT was also relatively large (as indicated by a confidence interval of 0.47) compared with that of the PVT (with a confidence interval of 0.22), the sensitivity

indexes for these two measures were comparable (at 0.957 and 0.955, respectively). The next highest sensitivity index was for lane deviation on the driving simulator (StiSim lane deviation) at 0.594. Statistically significant effect sizes were evident only for the top 9 (of 25) measures in the present study.

Performance and alertness measures – learning/practice effects

Performance and alertness data were also analyzed using mixed ANOVA with TIB group (3, 5, 7, or 9 h) as the first factor, day (BL, E1, E2, E3, E4, E5, E6, E7, R1, R2, and R3) as the second factor, and time of day (which varied across measures) as the third factor, and with repeated measures on the latter two factors. Table 4 lists the results of *post hoc* comparisons among specific days (separately for each sleep group) for those measures from Table 2 that showed both (a) a significant sleep group \times day interaction; and (b) at least one significant *post hoc* contrast among sleep groups. Learning effects (i.e. better performance on E7, R1, R2, or R3 compared with BL) were evident in all sleep groups for relative speed on 4-Choice RT, running memory, Stroop and serial addition/subtraction. Learning effects were evident in the 7-h group for relative accuracy on grammatical reasoning.

DISCUSSION

Of the various measures compared in the present study, the most sensitive – as reflected by the index produced by dividing effect size by confidence interval size – was the SLT and speed on the PVT (described by Dinges and Powell, 1985). The next most sensitive measures by this criterion were standard deviation of lane position on the StiSim driving simulator (described in Balkin *et al.*, 2000), Wilkinson 4-Choice RT, the SSS (Hoddes *et al.*, 1973), speed on serial addition/subtraction, speed on 10-Choice RT (both described in Thorne *et al.*,

Table 3 Outcome measures, rank ordered from highest to lowest sensitivity to sleep restriction, with effect size and confidence interval parameters listed for each

Outcome measure	Comparisons across measures				Sensitivity index (ratio of effect size to interval range)	Significant
	Effect size	Lower 95% confidence limit	Upper 95% confidence limit	Confidence interval range		
Sleep latency	0.447	0.214	0.679	0.465	0.961	Yes
PVT – speed (1/RT)	0.208	0.124	0.343	0.218	0.954	Yes
StiSim lane deviation	0.186	0.063	0.378	0.315	0.591	Yes
10-Choice reaction time – speed	0.032	0.012	0.074	0.062	0.510	Yes
StiSim lane position	0.075	0.024	0.173	0.149	0.502	Yes
Wilkinson 4-Choice reaction time – speed	0.131	0.012	0.275	0.263	0.496	Yes
Stanford Sleepiness Scale	0.095	0.030	0.227	0.197	0.482	Yes
Serial addition subtraction – speed	0.048	0.017	0.121	0.104	0.467	Yes
10-Choice reaction time – % correct	0.070	0.012	0.175	0.164	0.425	Yes
Saccadic velocity (FIT)	0.020	0.003	0.065	0.062	0.328	No
SynWork	0.055	-0.004	0.172	0.176	0.314	No
StiSim number of accidents	0.094	-0.195	0.142	0.337	0.280	No
Stroop color naming – speed	0.026	0.001	0.096	0.096	0.271	No
Stroop color naming – % correct	0.047	-0.007	0.178	0.185	0.252	No
Serial addition subtraction – % correct	0.028	-0.014	0.099	0.113	0.246	No
Time estimation	0.083	-0.024	0.329	0.353	0.235	No
Wilkinson 4-Choice reaction time – % correct	0.034	-0.012	0.134	0.147	0.231	No
Latency to pupil constriction (FIT)	0.012	-0.003	0.048	0.052	0.229	No
FIT impairment index	0.008	-0.017	0.021	0.038	0.212	No
Running memory – % correct	0.052	-0.029	0.221	0.250	0.210	No
Code substitution – speed	0.018	-0.014	0.087	0.101	0.177	No
Running memory – speed	0.029	-0.022	0.156	0.177	0.166	No
Logical reasoning – % correct	0.017	-0.013	0.091	0.104	0.164	No
Initial pupil diameter (FIT)	0.000	-0.002	0.001	0.003	0.139	No
Amplitude of pupil constriction (FIT)	0.000	-0.003	0.001	0.004	0.101	No
Logical reasoning – speed	0.001	-0.008	0.002	0.010	0.074	No

A 'yes' in the 'significant' column indicates that the effect size for the corresponding measure was significantly greater than zero ($P < 0.05$). FIT, Fitness Impairment Tester; PVT, psychomotor vigilance test; RT, reaction time.

1985), StiSim lane position (see Balkin *et al.*, 2000), and 10-Choice RT (Thorne *et al.*, 1985).

As an indicator of reliability – in the present context, the extent to which performance on a particular measure varies as a function of prior TST and *only* TST – the degree to which each of the various measures was subject to learning effects was gauged by comparing performance on the recovery days (i.e. following 8 h TIB on each night) versus performance on the baseline day (which also followed 8 h TIB). In the present study, significantly improved performance during recovery relative to baseline indicated that some performance-enhancing knowledge, skill, or strategy had been acquired over the course of the repeated test sessions – an effect that might not be apparent during the experimental phase of the study (in those groups whose sleep was restricted) because the effects of learning and accumulating sleep debt could offset each other. However, failure to find differences would not necessarily mean that no learning occurred – only that learning, if it occurred, was minimal – or had been offset by fatigue (e.g. a negative effect of the repeated testing paradigm). By these criteria, learning was evident on several of the performance measures, including two of the 'top nine' most sensitive measures: response speed on the serial addition/subtraction task, and on the Wilkinson 4-Choice RT test. Thus, although

reasonably sensitive to sleep restriction, these two measures might prove to be less useful indicators of performance capacity in some operational environments – unless training to asymptote is performed prior to fielding.

It is important to note that, to some extent, the criteria by which the various measures in the present study were compared (i.e. with indicators of sensitivity and reliability) – and the criteria by which it is suggested that measures might best be compared to, determine relative utility in the operational environment (i.e. content validity, cumbersomeness/fieldability, and intrusiveness) – constitute 'straw men'. The various criteria, and/or prioritization of the list of criteria, may vary as a function of the operational environment and/or the purpose of the testing. Thus, for example, among those factors listed by Manzey (2000) as being critical for assessment of performance during space flight was 'diagnosticity' (as well as sensitivity and reliability) – an appropriate consideration because in that environment, the ability to distinguish the deleterious effects of microgravity from those of general stress might impact the choices for intervention.

Likewise, several additional factors might be appropriate for inclusion in virtually any list of criteria for establishing/rating the utility of performance measures in the operational environment. For example, Dinges and Mallis (1988) suggest

Table 4 Significant *post hoc* contrasts among days (BL versus E7, R1, R2, and R3) for several tasks/dependent measures as a function of sleep group. Better performance on E7, R1, R2, or R3 compared with BL is indicative of learning/practice effects (see text). For most dependent variables, a cell entry of 'BL < ...' indicates possible learning effects

Task	Dependent measure	Post hoc contrast versus BL (3 h)			Post hoc contrast versus BL (5 h)			Post hoc contrast versus BL (7 h)			Post hoc contrast versus BL (9 h)						
		E7	R1	R2	R3	E7	R1	R2	R3	E7	R1	R2	R3	E7	R1	R2	R3
Total sleep time	Absolute minutes of sleep	BL < E7	NS	NS	NS	BL < E7	NS	NS	NS	BL < E7	NS	NS	NS	NS	NS	NS	NS
PVT	Relative speed	BL > E7	BL > R1	BL > R2	BL > R3	BL > E7	BL > R1	BL > R2	BL > R3	BL > E7	BL > R1	BL > R2	BL > R3	Day simple effect NS	NS	NS	NS
PVT	Relative speed – 2 times of day	BL > E7	BL > R1	BL > R2	BL > R3	BL > E7	NS	BL > R2	BL > R3	NS	NS	NS	NS	Day simple effect NS	NS	NS	NS
STISIM	Relative SD of lane tracking	BL < E7	BL < R1	BL < R2	BL < R3	BL < E7	NS	NS	NS	Day simple effect NS	NS	NS	NS	Day simple effect NS	NS	NS	NS
STISIM	Relative lane position	BL < E7	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	Day simple effect NS	NS	NS	NS
Stanford Sleepiness Scale	Rel. Sleepiness	BL > E7	NS	NS	NS	Day simple effect NS	NS	NS	NS	Day simple effect NS	NS	NS	NS	Day simple effect NS	NS	NS	NS
Wilkinson 4-Choice RT	Relative Speed	NS	NS	NS	BL < R3	BL < E7	BL < R1	BL < R2	BL < R3	BL < E7	BL < R1	BL < R2	BL < R3	BL < E7	BL < R1	BL < R2	BL < R3
Running memory Modified MSLT	Relative Speed	NS	BL < R1	BL < R2	BL < R3	BL < E7	BL < R1	BL < R2	BL < R3	BL < E7	BL < R1	BL < R2	BL < R3	BL < E7	BL < R1	BL < R2	BL < R3
Stroop	Abs. Latency to Sleep (min)	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	Day simple effect NS	NS	NS	NS
Serial addition/ subtraction	Relative Speed	NS	BL < R1	BL < R2	BL < R3	BL < E7	BL < R1	BL < R2	BL < R3	BL < E7	BL < R1	BL < R2	BL < R3	BL < E7	BL < R1	BL < R2	BL < R3
Running memory	Relative accuracy	BL > E7	NS	NS	NS	Day simple effect NS	NS	NS	NS	Day simple effect NS	NS	NS	NS	Day simple effect NS	NS	NS	NS
Serial addition/ subtraction	Relative accuracy	BL > E7	NS	NS	NS	NS	NS	NS	NS	Day simple effect NS	NS	NS	NS	Day simple effect NS	NS	NS	NS
Grammatical reasoning	Relative accuracy	Day simple effect NS	NS	NS	NS	Day simple effect NS	NS	NS	NS	BL < E7	BL < R1	BL < R2	BL < R3	Day simple effect NS	NS	NS	NS
Time estimation	Relative coefficient of variation	BL < E7	NS	NS	NS	Day simple effect NS	NS	NS	NS	Day simple effect NS	NS	NS	NS	Day simple effect NS	NS	NS	NS
Wilkinson 4-Choice RT	Relative accuracy	NS	NS	NS	NS	NS	BL > E7	NS	NS	Day simple effect NS	NS	NS	NS	Day simple effect NS	NS	NS	NS

that, in addition to sensitivity (the extent to which a measure reflects operationally relevant levels of variation in performance capacity) and validity (the extent to which a measure reflects a construct – such as *sleepiness* – that it is purported to reflect), measures should demonstrate concurrent validity (ability to predict the operationally relevant performance capacity at the same point in time); predictive validity (ability to predict the operationally relevant performance capacity at a future point in time); reliability (consistently measure the same ability); specificity (be relatively impervious to influence by factors that may not impact operationally relevant performance directly – thus, for example, avoiding ‘false alarms’); and generalizability (reflect the same abilities in all individuals tested). Indeed, notable deficiencies in any of these areas – including those in which there is little definitional overlap with the criteria outlined previously in the present paper, could vitiating the usefulness of a particular measure in a particular operational environment. In addition, as Dinges and Mallis (1988) point out, no less important than the practical considerations are the legal constraints that could apply to performance testing in actual operational environments.

It is important to note that there are many different types of measures that could have been selected for comparison, and several different aspects of the generated data that could have been analyzed and compared. The present results reflect only a small subset of all the possible data treatments that could have been applied. To complicate matters further, it is possible that individual differences in various capacities such as spatial, arithmetic, or linguistic abilities, manual dexterity, etc., could produce differential sensitivities to various types of tasks during sleep loss. And that, as a result, the relative positions in a rank ordering of measures (on the basis of sensitivity to sleep loss) could vary for individuals. This is a topic that warrants further investigation.

As in any study that includes multiple dependent variables (including studies in which more traditional statistical analyses are applied to determine differences between means), generalizability is determined (limited) by the study design. In the present study, it is possible that the relative sensitivity of the various measures was affected by (in no particular order): (a) *task duration/time on task*, Wilkinson (1965) has shown that long-duration tasks are generally more sensitive to the effects of sleep loss than are short duration tasks, so it is possible that sensitivity of a particular task could be increased by merely lengthening it. (b) *Task sequence*. It is also possible that the sensitivity of tasks administered toward the end of a block of tests (e.g. after the PAB) was inflated, with those later tasks ‘benefiting’ from residual fatigue, boredom, cognitive resource depletion, etc., produced by earlier tests. (c) *Timing/time of day*. Although it is known that performance on most tasks varies systematically across the day, differential circadian effects across tasks have seldom been examined. Some of the variance in sensitivity seen in the present study may be a function of the relative timing of task administrations. (d) *Number of tests administered*. It is possible that sensitivity to sleep loss varies as a function of the number of

times per day that a particular test is administered (i.e. the sampling rate). (e) *The type of sleep loss*. Belenky et al. (2003) suggest that there may be different physiological effects of total versus partial sleep loss, so it is conceivable that these two types of sleep loss might also produce different task sensitivity profiles.

Clearly, it is not feasible to conduct a study, or a series of studies, in which all of these possibilities are fully controlled, because it would mean counterbalancing in terms of sequence and time of day, and testing multiple versions of each task (differing by duration) while also varying the number of administrations per day – a process that could quickly create a staggering number of experimental cells for each measure of interest. Rather, it is suggested that further studies in which head-to-head comparisons of the sensitivity and reliability of multiple measures be conducted so that a Bayesian decision-making process can be applied to identify those measures (and measure parameters) that are likely to provide relatively good sensitivity. This information, along with practical/logistical considerations, will provide the basis for choosing performance measures for fielding in operational environments. Based on the present findings as well as the aforementioned practical considerations, the PVT should be regarded the leading candidate among the measures tested thus far.

ACKNOWLEDGEMENTS

This work was supported by the US Army Medical Research and Materiel Command (Military Operational Medicine Program, Project S15 Q), the US Department of Transportation (Contract/Grant No. DTFH61-94-Y-00090), and the GCRC/Johns Hopkins Bayview Medical Center (Grant No. M01RR02719), Baltimore, Maryland.

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. This study was approved by the Walter Reed Army Institute of Research Human Use Committee and the United States Army Medical Research and Materiel Command Human Subjects Review Board of the Army Surgeon General and was performed in accordance with the ethical standards laid down in the 1964 Declaration of Helsinki.

REFERENCES

- Ahonniska, J., Ahonen, T., Aro, T., Tolvanen, A. and Lyytinen, H. Practice effects on visuomotor and problem-solving tests by children. *Percept. Mot. Skills*, 2001, 92: 479–494.
- Balkin, T., Thorne, D., Sing, H., Thomas, M., Redmond, D., Wesensten, N., Williams, J., Hall, S. and Belenky, G. *Effects of Sleep Schedules on Commercial Driver Performance*. Report No. DOT-MC-00-133. US Department of Transportation, Federal Motor Carrier Safety Administration, Washington, DC, 2000.

- Belenky, G., Wesensten, N. J., Thorne, D. R., Thomas, M. L., Sing, H. C., Redmond, D. P., Russo, M. B. and Balkin, T. J. Patterns of performance degradation and restoration during sleep restriction and subsequent recovery: a sleep dose-response study. *J. Sleep Res.*, 2003, 12: 1–13.
- Carskadon, M. A., Dement, W. C., Mitler, M. M., Roth, T., Westbrook, P. R. and Keenan, S. Guidelines for the multiple sleep latency test (MSLT): a standard measure of sleepiness. *Sleep*, 1986, 9: 519–524.
- Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*, 2nd edn. Erlbaum, Hillsdale, NJ, 1988.
- Dinges, D. F. and Mallis, M. M. Managing fatigue by drowsiness detection: can technological promises be realised? In: L. R. Hartley (Ed.) *Managing Fatigue in Transportation. Proceedings of the Third International Conference on Fatigue and Transportation, Fremantle, Western Australia*. Elsevier Science Ltd, Oxford, 1988: 209–229.
- Dinges, D. F. and Powell, J. W. Microcomputer analyses of performance on a portable, simple visual RT task during sustained operations. *Behav. Res. Methods Instrum. Comput.*, 1985, 17: 652–655.
- Efron, B. and Tibshirani, R. *An Introduction to the Bootstrap*. Chapman and Hall, New York, 1993.
- Elsmore, T. F. SYNWORK1: a PC-based tool for assessment of performance in a simulated work environment. *Behav. Res. Methods Instrum. Comput.*, 1994, 26: 421–426.
- Guilleminault, C. and Dement, W. C. Amnesia and disorders of excessive sleepiness. In: R. R. Drucker-Colin and J. L. McGaugh (Eds) *Neurobiology of Sleep and Memory*. Academic Press, New York, 1977: 439–456.
- Hall, P. Theoretical comparison of bootstrap confidence intervals. *Ann. Stat.*, 1988, 16: 1–50.
- Harrison, Y. and Horne, J. A. “High sleepability without sleepiness”. The ability to fall asleep rapidly without other signs of sleepiness. *Neurophysiol. Clin.*, 1996, 26: 15–20.
- Hoddes, E., Zarcone, V., Smythe, H., Phillips, R. and Dement, W. C. Quantification of sleepiness: a new approach. *Psychophysiology*, 1973, 10: 431–436.
- Leger, D. The cost of sleep-related accidents: a report for the National Commission on Sleep Disorders Research. *Sleep*, 1994, 17: 84–93.
- Leger, D. The cost of sleepiness: a response to comments. *Sleep*, 1995, 18: 281–284.
- Lubin, A. Performance under sleep loss and fatigue. In: S. S. Kety, E. V. Evarts and H. L. Williams (Eds) *Sleep and Altered States of Consciousness*. Williams and Wilkins, Baltimore, 1967: 506–513.
- Manzey, D. Monitoring of mental performance during spaceflight. *Aviat. Space Environ. Med.*, 2000, 71 (Suppl.): A69–A75.
- McCarthy, M. E. and Waters, W. F. Decreased attentional responsiveness during sleep deprivation: orienting response latency, amplitude, and habituation. *Sleep*, 1997, 20: 114–123.
- Mitler, M. M., Gujavarty, K. S. and Browman, C. P. Maintenance of wakefulness test: a polysomnographic technique for evaluation of treatment efficacy in patients with excessive somnolence. *Electroencephalogr. Clin. Neurophysiol.*, 1982, 53: 658–661.
- Patrick, G. T. W. and Gilbert, J. A. On the effects of loss of sleep. *Psychol. Rev.*, 1896, 3: 469–483.
- Rajaratnam, S. M. and Arendt, J. Health in a 24-h society. *Lancet*, 2001, 358: 999–1005.
- Richardson, G. S., Carskadon, M. A., Flagg, W., van den Hoed, J., Dement, W. C. and Mitler, M. M. Excessive daytime sleepiness in man: multiple sleep latency measurement in narcoleptic and control subjects. *Electroencephalogr. Clin. Neurophysiol.*, 1978, 45: 621–627.
- Rothstein, H. R., McDaniel, M. A. and Borenstein, M. Meta-analysis: a review of quantitative cumulation methods. In: F. Drasgow and N. cSchmitt (Eds) *Measuring and Analyzing Behavior in Organizations: Advances in Measurement and Data Analysis*. Josse-Bass, San Francisco, 2002: 534–570.
- Russo, M., Thomas, M., Thorne, D., Sing, H., Redmond, D., Rowland, L., Johnson, D., Hall, S., Krichmar, J. and Balkin, T. Oculomotor impairment during chronic partial sleep deprivation. *Clin. Neurophysiol.*, 2003, 114: 723–736.
- Thorne, D. R., Genser, S. G., Sing, H. C. and Hegge, F. W. The Walter Reed performance assessment battery. *Neurobehav. Toxicol. Teratol.*, 1985, 7: 415–418.
- Tibshirani, R. Variance stabilization and the bootstrap. *Biometrika*, 1988, 75: 433–444.
- Webb, W. B. The cost of sleep-related accidents: a reanalysis. *Sleep*, 1995, 18: 276–280.
- Wilkinson, R. T. Sleep deprivation. In: O. G. Edholm and A. L. Bacharach (Eds) *The Physiology of Human Survival*. Academic Press, London, 1965: 399–430.