http://dx.doi.org/10.5665/sleep.1606

Reliability of a Single Objective Measure in Assessing Sleepiness

Bernie Y. Sunwoo, BSc MBBS[1,3]; Nicholas Jackson, MPH[2,3]; Greg Maislin, MS, MA[2,3]; Indira Gurubhagavatula, MD, MPH[2,3]; Charles F. George, MD[4]; Allan I. Pack, MBChB, PhD[2,3]

[1]Department of Medicine, Division of Pulmonary and Critical Care, University of Pennsylvania, Philadelphia, PA; [2]Department of Medicine, Division of Sleep Medicine, University of Pennsylvania, Philadelphia, PA; [3]Center for Sleep and Circadian Neurobiology, University of Pennsylvania, Philadelphia, PA; [4]Division of Respirology, University of Western Ontario, Ontario, Canada

Study Objectives: To evaluate reliability of single objective tests in assessing sleepiness.

Design: Subjects who completed polysomnography underwent a 4-nap multiple sleep latency test (MSLT) the following day. Prior to each nap opportunity on MSLT, subjects performed the psychomotor vigilance test (PVT) and divided attention driving task (DADT). Results of single versus multiple test administrations were compared using the intraclass correlation coefficient (ICC) and adjusted for test administration order effects to explore time of day effects. Measures were explored as continuous and binary (i.e., impaired or not impaired).

Setting: Community-based sample evaluated at a tertiary, university-based sleep center.

Participants: 372 adult commercial vehicle operators oversampled for increased obstructive sleep apnea risk.

Interventions: N/A

Measurements and Results: As continuous measures, ICC were as follows: MSLT 0.45, PVT median response time 0.69, PVT number of lapses 0.51, 10-min DADT tracking error 0.87, 20-min DADT tracking error 0.90. Based on binary outcomes, ICC were: MSLT 0.63, PVT number of lapses 0.85, 10-min DADT 0.95, 20-min DADT 0.96. Statistically significant time of day effects were seen in both the MSLT and PVT but not the DADT. Correlation between ESS and different objective tests was strongest for MSLT, range [−0.270 to −0.195] and persisted across all time points.

Conclusions: Single DADT and PVT administrations are reliable measures of sleepiness. A single MSLT administration can reasonably discriminate individuals with MSL < 8 minutes. These results support the use of a single administration of some objective tests of sleepiness when performed under controlled conditions in routine clinical care.

Keywords: Excessive sleepiness, multiple sleep latency test, psychomotor vigilance test, divided attention driving task, objective tests

Citation: Sunwoo BY; Jackson N; Maislin G; Gurubhagavatula I; George CF; Pack AI. Reliability of a single objective measure in assessing sleepiness. *SLEEP* 2012;35(1):149-158.

INTRODUCTION

Excessive sleepiness is common, affecting 5% to 20% of the general population, and is likely to increase in prevalence with growing pressure towards a 24-hour society.[1,2] Identification of excessive sleepiness may help identify treatable sleep disorders (such as obstructive sleep apnea [OSA]), and medical and psychological disorders. Moreover, excessive sleepiness has been associated with adverse consequences to both the individual and society, including cardiovascular disease, insulin resistance, neurobehavioral and performance deficits, job errors, and motor vehicle crashes.[3-9] Accordingly, identification and assessment of sleepiness is important and particularly relevant in high-risk occupations, such as driving commercial vehicles.

Subjective and objective tools are available for assessing sleepiness. The Epworth Sleepiness Scale (ESS) asks patients to rate their likelihood of falling asleep under 8 sedentary conditions. It is currently among the most commonly used measures of sleepiness in clinical practice, but subjective measures have limitations. Individuals are not always aware of their degree of sleepiness or their susceptibility to impairment, with significant inter-individual differences in performance described following sleep deprivation.[10-14] Moreover, in the occupational setting, underreporting of sleepiness may exist.[15] There is a need for complementary simple, objective measures.

The multiple sleep latency test (MSLT) has been endorsed as the *de facto* gold standard objective measure of excessive sleepiness by the American Academy of Sleep Medicine (AASM).[16] Subjects are given multiple nap opportunities at 2-h intervals following nocturnal polysomnography (PSG) under standardized test conditions, and the average time to fall asleep or the mean sleep latency (MSL) is measured. The MSL has been shown to be sensitive to conditions expected to increase sleepiness, including sleep deprivation, and shows expected changes with sedative and stimulant medications as well as treatment of sleep disorders.[2] The psychomotor vigilance test (PVT) and divided attention driving task (DADT) focus more on neurocognitive function. This is especially relevant among commercial drivers who undertake the complex task of driving.[17,18] While not direct tests of physiological sleepiness, the PVT and DADT have been shown to be sensitive to sleepiness, and for simplicity are referred to as tests of sleepiness in our study.

Subjective and objective measures of sleepiness like the ESS and MSLT correlate only weakly at best in populations with excessive daytime sleepiness, many with OSA.[2,19-24] The poor agreement between the various measures of sleepiness may be a reflection of the multidimensional nature of sleepiness.[22] Thus, one metric of sleepiness may be inadequate in a comprehensive evaluation of sleepiness. Despite this, the ESS remains the preferred commonly used subjective measure of sleepiness in clinical practice. A barrier to the routine implementation of the MSLT and other objective measures of sleepiness is that they

Submitted for publication March, 2011
Submitted in final revised form May, 2011
Accepted for publication May, 2011
Address correspondence to: Bernie Y. Sunwoo, BSc, MBBS, Center for Sleep and Circadian Neurobiology, 3624 Market Street, Suite 205, Philadelphia, PA 19104; Tel: (215) 615-4868; Fax: (215) 615-4874; E-mail: bernie.sunwoo@uphs.upenn.edu

are cumbersome to administer, usually requiring a full day for testing. A single objective test, administered quickly and easily, could be used more readily (much like the ESS), and perhaps become part of the initial assessment of patients and part of an outcomes assessment battery to follow responses to therapy.

We assessed the value of a single objective test sensitive to sleepiness. We argued that, if used routinely in clinical practice, such a test should have a high intraclass correlation when applied to the same subject on multiple occasions. We evaluated the following primary measures: (1) sleep latency on a single nap on MSLT; (2) median response time and number of lapses on the PVT[6,25]; and (3) mean tracking error on the DADT at 10 minutes and 20 minutes.[26] All tests were done on the same individual so that we could compare results for different assessment approaches using a within-subject design. We assessed whether a single administration of the DADT and PVT, or determination of sleep latency based on a single nap on the MSLT might be suitable for routine use in clinical practice to assess sleepiness for diagnostic purposes and for follow-up of outcomes of therapy.

METHODS

Data from this study have been used to look at the role of short sleep duration and sleep apnea on performance and occupational screening in commercial drivers, and have been previously published.[27,28] We present here results of a new analysis of data to assess the goals of our study.

Participants

A list of n = 4,826 randomly selected holders of commercial drivers licenses living within 50 miles of the University of Pennsylvania, Philadelphia, was obtained from the state. The Multivariable Apnea Prediction questionnaire, including information on age, sex, body mass index (BMI), and 3 questions to determine an apnea symptom-frequency index was mailed to determine the relative likelihood of apnea.[29] Stratified sampling was employed among the n = 1,329 of 4,410 valid contacts who completed the questionnaire to oversample for drivers at higher risk of sleep apnea. By design, polysomnography was performed in n = 247 (44.8%) of 551 highest risk drivers for apnea and then, in randomized order, n = 159 (20.4%) of 778 lower risk drivers. In total, 406 subjects with polysomnography were included. Overnight PSG was followed by the MSLT, the PVT, and the DADT. These were carried out 4 times a day after the PSG, each test being separated by 2 h. The study was approved by the Institutional Review Board of the University of Pennsylvania, and all subjects gave written informed consent.

Experimental Procedure

Prior to in-laboratory polysomnography, mean sleep duration at home was assessed for one week by actigraphy (Ambulatory Monitoring, Ardsley, NY) and daily sleep diary. Mean sleep duration was estimated as the mean cumulative duration of relative inactivity during the main sleep bout. For ≥ 2 weeks before testing, subjects were instructed to stop medications that were considered to be stimulants, stimulant-like medications, sedative hypnotics, or REM-suppressing agents. Urine screening for illicit drugs and alcohol but not caffeine was performed prior to the in-laboratory study. Subjects who tested positive were excluded.

Full overnight in-laboratory polysomnography was performed on all subjects, using electroencephalogram (EEG); electrooculogram (EOG); submental and pretibial electromyogram (EMG); electrocardiography; body position; respiratory effort by belts around the chest and abdomen; oximetry (N-200 oximeter, Nellcor Inc., Pleasanton, CA, with sampling frequency 3 Hz, paper speed 15 cm/h); and airflow measured with nasal pressure and oral thermistors. The apnea-hypopnea index ([AHI], the number of apneas plus hypopneas per hour of sleep), was defined on the basis of current American Academy of Sleep Medicine (AASM) criteria.[30] An apnea was defined as ≥ 10 sec of airflow cessation. A hypopnea was defined as ≥ 50% airflow reduction for ≥ 10 sec associated with ≥ 3% fall in oxyhemoglobin saturation and/or an arousal.

The PSG was followed by the 4-nap MSLT. The MSLT was performed using the standard research protocol published by the AASM.[16] Subjects were given 20-min nap opportunities at 09:30, 11:30, 13:30, 15:30, with a standard lunch provided between 11:50 and 12:10. Only decaffeinated beverages were provided on the day of testing. The 10-min long PVT and 20-min long DADT were administered within 1 h prior to each nap opportunity. In half the subjects, the PVT was administered first, 1 h prior to each nap opportunity, followed by the DADT 30 min prior to each nap opportunity. In the other half of subjects, the DADT was administered first, 1 h prior to each nap opportunity, followed by the PVT, 30 min prior to each nap opportunity. The ESS was administered to assess subjective sleepiness.

Mean Sleep Latency Test

A 4-nap MSLT was administered using the standard AASM research protocol.[16] Subjects were asked to stop smoking 30 min prior to each nap opportunity, although smoking breaks were deliberately built into the protocol. Standard instructions for biocalibrations were performed before each nap per AASM protocol. Sleep latency was defined by the time from lights out to the first epoch > 15 sec of any stage of cumulative sleep in a 30-sec epoch. A nap session was terminated after 20 min if sleep did not occur. Individuals not reaching any stage of sleep during this 20-min period were considered to have a sleep latency of 20 min. If sleep occurred, subjects were awakened after 5 min of stage 1 sleep or the first sign of stage 2 sleep. REM latency was not calculated, as no subjects reached REM within the 20-min test period.

Psychomotor Vigilance Test (PVT)

Subjects were familiarized with the equipment and underwent a practice run of testing on the previous evening, prior to polysomnography. Subjects were instructed to respond as quickly as possible by pressing a button in response to a bright, millisecond signal occurring at random inter-stimulus intervals over a 10-min period on a small, dark, rectangular screen. A lapse was defined as a response time ≥ 500 ms. Measured outcomes included the median response time (in ms) and the number of response lapses. As the number of lapses exhibited a strong right skew, square root transformation of the number of lapses was also examined.

Divided Attention Driving Task (DADT)

The 20-min DADT was used to assess driving performance, where tracking was the primary task and visual search

the secondary task. Subjects were instructed to steer a cursor within 2 lines on the computer screen that were continuously moved laterally in random amounts (tracking). While tracking activity, the numbers 0-9 appeared at random intervals in one of the corners of the screen, and subjects were instructed to press a button in response to the number 2 (visual search). The corner in which the number appeared was selected randomly by the computer. The main outcome was mean tracking error over the time of the test. The tracking error is the distance of the cursor from the center of the lines in centimeters. To allow comparison with the PVT and to assess whether test duration would improve discriminatory power, we assessed tracking error over the first 10 min of the test and then over the full 20 min.

Statistical Analysis

The intraclass correlation coefficient (ICC) was used to determine within-subject agreement of repeated test administration. Van Dongen et al. discuss the validity of using the ICC to assess test-retest reliability of objective measures of sleepiness.[13] It is a common method of determining conformity among repeated measures. Landis and Koch have previously defined ICC values to reflect increasing stability of observed inter-individual differences as 0.0-0.2 (slight); 0.2-0.4 (fair); 0.4-0.6 (moderate); 0.6-0.8 (substantial); and 0.8-1.0 (near perfect).[13,31] The ICC was calculated as the ratio of between subject variance to total variance using a random effects models for continuous (mixed effects linear regression) and binary (mixed effects logistic regression) outcomes. Binary models were based on definitions of what constitutes impaired or non-impaired for each test. Sleep latency, psychomotor vigilance lapses, and DADT mean tracking error were transformed according to the functions $x^{0.5}$, $(x^{0.5} + (x+1)^{0.5})$, and $\log_e(x + 1)$, respectively, to better fit parametric assumptions before analysis. An adjusted ICC model incorporating test administration order was used to determine variance adjusted for test order, which was largely interpreted in terms of time of day effects. However, there was no attempt to distinguish true time of day effects from other effects such as learning or changes in motivation, the net effects of which are assumed to be reflected in the generic order effect included in the statistical models. Pearson correlations were used to examine the bivariate relations between different objective and subjective tests within test administration times. Analyses were conducted using Stata Version 11 (StataCorp, College Station, TX).

RESULTS

Population Sample

A total of 372 out of 406 subjects with full polysomnography completed all 3 objective tests of sleepiness—the MSLT, PVT, and DADT—on 4 occasions. The study population was predominantly male (94.1%), Caucasian (85.1%), with a mean age of 45.7 y, a mean BMI of 30.2 kg/m^2, and a mean AHI of 5.3 ± 9.8 (65.6% AHI < 5, 20.7% AHI 5-15, 7.80% AHI 15-30, 5.91% AHI > 30) (Table 1). These numbers are slightly different from those previously reported for this sample.[27,28] This is the result of a different number of subjects being included in this report. The subjects who did not complete all 3 objective tests of sleepiness were similar in demographics but more obese (P = 0.001; see Supplemental Table S1).

Table 1—Overall demographic information

Age (years) Mean ± SD	45.7 ± 11.3
BMI (kg/m^2) Mean ± SD	30.2 ± 5.7
Sex % Male	94.1
Race % White	85.1
Actigraphy total sleep time (min) Mean ± SD	376.9 ± 76.8
PSG total sleep time (min) Mean ± SD	382.5 ± 58.1
PSG AHI (events/h) Mean ± SD	5.3 ± 9.8

Characteristics of study population, n = 372.

Single Versus Multiple Test Administrations

The mean sleep latency (MSL) was explored as both a continuous and binary measure, using a cutpoint of < 8 min to define excessive sleepiness. This cutpoint was selected based on the 2005 International Classification of Sleep Disorders' definition of sleepiness.[32] The PVT and DADT were also examined as both continuous and binary measures. We defined impairment on the PVT as > 3 lapses per 10-min trial, based on a prior dose-response study looking at the effects of alcohol.[33] For the DADT, we used a mean absolute tracking error of > 250 cm based on studies looking at the effects of alcohol ingestion in patients with untreated OSA and healthy controls.[26] The first 10 min of the 20-min DADT was also analyzed to explore the characteristics of a shorter test.

Continuous measures

Sleep latency on MSLT increased from 412.9 ± 344.2 sec at first test administration to 601.5 ± 397 seconds at fourth test administration (Table 2). ICC showed moderate agreement of 0.45 as a continuous measure (Table 3 and Figure 1). The PVT showed substantial between test agreement as a continuous measure, with ICC values of 0.69 for median response time, 0.51 for number of lapses, and 0.60 for square-root transformed number of lapses. The DADT had the highest between-test agreement, with an ICC value of 0.90 for mean tracking error at 20 minutes. The 10-min DADT mean tracking error ICC was only marginally lower at 0.87 (Table 3 and Figure 1).

Binary measures

As a binary measure, using a cutpoint of 8 min to define impairment, the MSLT had an ICC value of 0.63 (Table 4 and Figure 1). The PVT, using a cutpoint of > 3 lapses, and 10-min DADT, with a cutpoint of > 250 cm for tracking error, had substantial between-test agreement, with ICC values of 0.85 and 0.95, respectively (Table 4). The ICC values for binary measures are not directly comparable to those derived for continuous measures due to differences in the approach to variance component estimation.

Time of Day Effects

Given known circadian influences on sleepiness and tests of sleepiness, we explored the magnitude of time of day fixed effects.

Table 2—Objective test results across test administrations: continuous variables

		Time 1 Mean ± SD	Time 2 Mean ± SD	Time 3 Mean ± SD	Time 4 Mean ± SD
MSLT	Sleep latency (seconds)	412.9 ± 344.2	456.1 ± 346.9	503.0 ± 348.1	601.5 ± 397.0
PVT	Median response Time (ms)	257.4 ± 30.5	266.7 ± 44.0	268.4 ± 45.4	266.6 ± 40.7
	Number of lapses	1.8 ± 3.0	3.1 ± 5.9	3.1 ± 5.1	2.9 ± 4.5
	Number of lapses transformed	2.5 ± 1.6	3.1 ± 2.1	3.2 ± 2.0	3.1 ± 1.9
DADT	Mean tracking error 10 min (cm)	245.0 ± 169.7	241.8 ± 156.6	248.9 ± 152.4	246.4 ± 145.5
	Mean tracking error 20 min (cm)	247.9 ± 172.3	245.2 ± 165.4	252.4 ± 164.4	248.9 ± 155.7

Mean results of each individual objective test in the overall study population at each test administration across time.

Table 3—Intraclass correlation coefficients (ICC): continuous variables

		ICC - Completely Random Model *ICC*	ICC - Order Effects Adjusted *ICC*
MSLT	Sleep latency	0.45	0.49
PVT	Median response time	0.69	0.71
	Number of lapses	0.51	0.53
	Number of lapses transformed	0.60	0.63
DADT	Mean tracking error 10 min	0.87	0.87
	Mean tracking error 20 min	0.90	0.90

ICC of each objective test as a continuous measure, both unadjusted and adjusted for order effects.

Table 4—Intraclass correlation coefficients (ICC): binary (impaired or not impaired)

		ICC - Completely Random Model *ICC*	ICC - Order Effects Adjusted *ICC*
MSLT	% Impaired (Latency < 8 Min)	0.63	0.67
PVT	% Impaired (Lapses > 3)	0.85	0.86
DADT	% Impaired 10 min (Mean tracking error > 250 cm)	0.95	0.96
	% Impaired 20 min (Mean tracking error > 250 cm)	0.96	0.96

ICC of each objective test as a binary measure, both unadjusted and adjusted for order effects

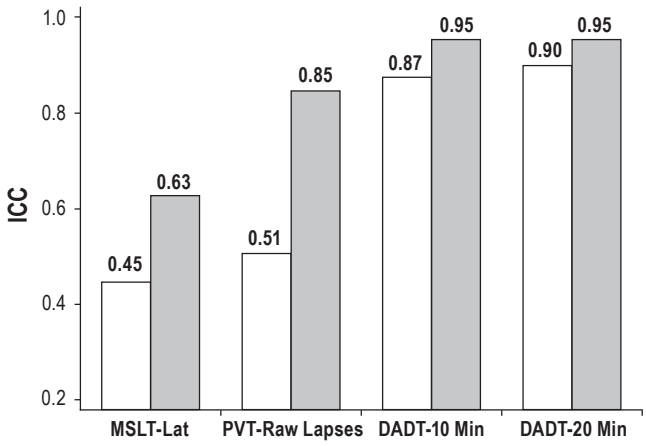

Figure 1—ICC values of MSLT (latency), PVT (raw number of lapses), DADT (tracking error) at 10 min and 20 min, explored as both a continuous (white) and binary measure (gray), showing highest reliability for the DADT.

Continuous measures

There were statistically significant order effects on the MSLT ($P < 0.001$). Comparison of each objective measure between tests administered at different times is shown in Table 5. While the differences between test administrations were statistically significant, the average difference in sleep latency between test administrations was small, i.e., 102.5 seconds. The greatest differences were seen between tests administered the furthest apart from each other in time, i.e., test 1 compared to test 4. A statistically significant order effect was also seen in the PVT ($P < 0.001$) as a continuous measure (Table 5). The average differences between test administrations were again small (4.89 ms for median response time, 0.56 and 0.32 for raw and transformed lapses, respectively; Table 5 and Figure 2). The DADT showed no statistically significant order effects (Table 5 and Figure 2).

After introducing time of day effects, the ICC values thus adjusted changed minimally from unadjusted calculations (Table 3). ICC increased to 0.49 (0.45) (adjusted ICC with unadjusted in parentheses) for MSLT, 0.71 (0.69) for PVT median response time, 0.53 (0.51) and 0.63 (0.60) for raw and transformed PVT lapses, respectively. ICC values for the 10- and 20-min DADT did not change at the level of 2 significant digits.

Binary measures

Statistically significant order effects were seen on MSLT when explored as a binary measure to identify impairment ($P < 0.001$) with an average change from the earlier to later tests in impairment of 12.8% (Table 6 and Figure 3). Table 6 shows comparison of each objective test measure as a binary measure between tests administered at different times. Significant order effects were also in the PVT ($P < 0.001$), with an average change from

Table 5—Order effects. Comparison between different time administrations: continuous variables

			Overall	T1 vs. T2	T1 vs. T3	T1 vs. T4	T2 vs. T3	T2 vs. T4	T3 vs. T4
MSLT	Sleep latency (sec)	Difference (seconds)	**−102.5**	**−49.79**	**−95.39**	**−186.83**	**−45.02**	**−142.47**	**−95.46**
		P value	**< 0.001**	**0.005**	**< 0.001**	**< 0.001**	**0.004**	**< 0.001**	**< 0.001**
PVT	Median response time (ms)	Difference	**−4.89**	**−9.47**	**−10.84**	**−9.27**	−1.54	0.08	1.67
		P value	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	0.355	0.962	0.332
	Number of lapses	Difference	**−0.56**	**−1.30**	**−1.32**	**−1.10**	−0.02	0.18	0.21
		P value	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	0.096	0.844	0.144
	Number of lapses transformed	Difference	**−0.32**	**−0.61**	**−0.72**	**−0.59**	−0.11	0.01	0.12
		P value	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	0.052	0.622	0.148
DADT	Mean tracking error 10 min (cm)	Difference	−1.43	3.79	−3.21	−0.45	**−6.96**	−4.48	2.70
		P value	0.135	0.643	0.112	0.252	**0.039**	0.106	0.656
	Mean tracking error 20 min (cm)	Difference	−1.37	3.12	−4.06	−0.30	**−7.00**	−3.53	3.57
		P value	0.215	0.438	0.204	0.618	**0.040**	0.201	0.438

Difference in mean values between test administrations at different times for each objective test, with P values (bold = statistically significant, P < 0.05).

the earlier to later tests in impairment of 4.2%, as defined by > 3 lapses (Table 6). Significant DADT order effects (P < 0.001) showed the same type of average difference in impairment of 4.8% for 10 min between test administrations. The DADT had the least change in classification (Table 6 and Figure 3).

ICC values for impairment, after adjusting for order effects, increased for all of the objective tests (Table 4). ICC increased to 0.67 (0.63) (adjusted ICC with unadjusted in parentheses) for the MSLT, 0.86 (0.85) for the PVT, and 0.96 (0.95) for the 10-min DADT.

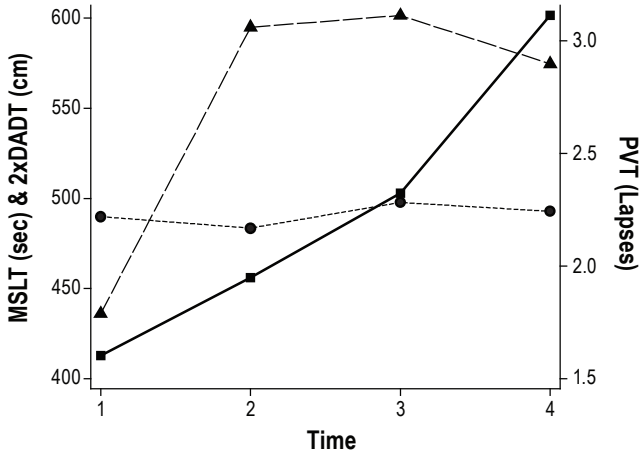

Figure 2—Time of day effects of continuous measures. Change in MSLT mean sleep latency (squares), PVT number of lapses (triangles), and DADT tracking error (circles), across test administrations from time 1 to time 4. The DADT shows the least change across the day.

Correlation between Subjective and Different Objective Measures of Sleepiness

We next assessed correlation between the ESS as a subjective measure of sleepiness and the different objective measures of sleepiness, as well as between the different individual objective measures. This was performed across all 4 time points (Table 7). Supplemental Table S4a-S4d shows correlation at each test administration at all time points from time 1 (Supplemental Table S4a) to time 4 (Supplemental Table S4d). The correlation between the ESS and objective tests of sleepiness was strongest for the MSLT with correlation ranging from −0.270 to −0.195 across different test administrations (Table 7). While there was moderate correlation between the DADT and PVT, there was weak correlation between the MSLT and other objective measures (Table 7). Time of day effects on the magnitude of relations between subjective and the different objective measures and between the different objective tests of sleepiness was explored by comparing these relations between the 4 test administrations (Supplemental Table S5). Correlation between the ESS and MSLT did not significantly change when assessed across different times of test administration (P = 0.35) (Supplemental Table S5).

DISCUSSION

Currently, no simple, single objective measure of sleepiness is in routine clinical use. Our study is the first to our knowledge to compare reliability of a single MSLT, PVT, and DADT test administration within the same subject. Our results indicate that a single administration of either the PVT or 10-minute DADT provide reliable objective measures of sleepiness, with ICC values of 0.69 and 0.87 respectively. The divided attention task is slightly superior but there are substantially more data on the PVT.[6,25] Both the PVT and DADT have a high signal load, making them sensitive to changes in vigilant attention.[5,6,25,34,35] Combined with their relative ease in administration, our results support routine use of either of these simple 10-min objective tests (PVT or DADT) in assessment of excessive sleepiness.

In contrast, the MSLT showed significant variations with repeated testing, with an ICC value of 0.45 for sleep latency when measured as a continuous measure consistent with prior studies, showing only fair agreement.[36] Zwyghuizen-Doorenbos et al. did not find a single measure of latency reliable.[36]

Table 6—Order effects. Comparison between different time administrations: binary (impaired compared to not impaired)

			Overall	T1 vs. T2	T1 vs. T3	T1 vs. T4	T2 vs. T3	T2 vs. T4	T3 vs. T4
MSLT	% Impaired (Latency < 8 min)	Difference	**12.8**	**6.6**	**15.3**	**22.6**	**8.7**	**16.0**	**7.4**
		P value	**< 0.001**	**0.034**	**< 0.001**	**< 0.001**	**0.011**	**< 0.001**	**0.034**
PVT	% Impaired (number of lapses > 3)	Difference	**−4.2**	**−11.5**	**−11.8**	**−8.2**	−0.3	3.3	3.6
		P value	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	0.942	0.179	0.158
DADT	% Impaired 10 min (mean tracking error > 250 cm)	Difference	**−4.8**	−1.8	**−7.8**	**−7.6**	**−6.0**	**−5.8**	0.2
		P value	**< 0.001**	0.385	**0.001**	**0.001**	**0.008**	**0.012**	0.891
	% Impaired 20 min (mean tracking error > 250 cm)	Difference	−2.9	−1.8	**−4.5**	**−4.9**	−2.8	−3.1	−0.4
		P value	0.086	0.387	**0.038**	**0.027**	0.216	0.170	0.892

Difference in mean values between test administrations at different times for each objective test, with P values, (bold = statistically significant, P < 0.05), explored binary.

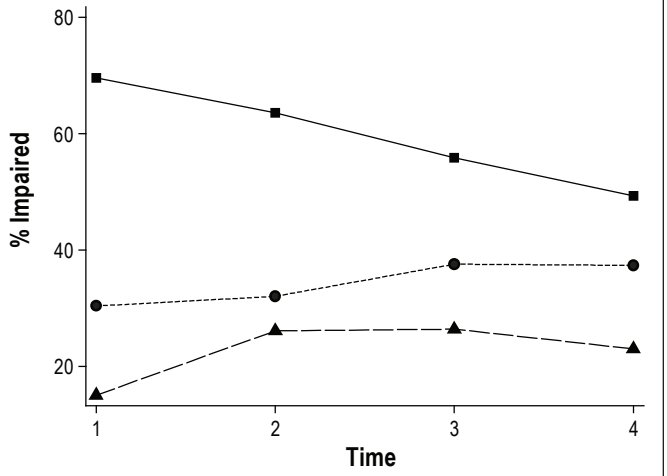

Figure 3—Time of day effects of binary measures. Change in impairment on MSLT defined by latency < 8 minutes (squares), PVT defined by number of lapses > 3 (triangles), and DADT defined by tracking error > 250 cm (circles) across test administrations from time 1 to time 4. The DADT continues to show little change. See Supplemental S3.

When comparing the mean sleep latency from the 4-nap MSLT in 14 healthy normal subjects over a period of 4 to 14 months, strong test-retest reliability was only seen with 3 or more tests of latency, where r = 0.97 (P < 0.001) for 4 naps decreased to r = 0.65 for 2 naps.[36] The retest interval or degree of sleepiness had no significant effect. Similarly, a review by Arand found a statistically significant difference in the mean latency between the 4-nap and 5-nap MSLT (P < 0.01), concluding that a single MSLT test administration was not reliable.[2]

However, an absolute mean sleep latency value may not be the most important aspect clinically. While excessive sleepiness is accepted as inappropriate sleepiness occurring in a situation when an individual would be expected to be awake and alert, no standardized definition of excessive sleepiness exists.[2] Normative data for the MSLT is lacking. In 2005, Arand et al. tried to assemble a surrogate database of MSL values in a comprehensive review. Pooling data from normal subjects over all ages, they found a mean sleep latency using a 4-nap MSLT of 10.4 ± 4.3 minutes. They demonstrated significant overlap in the MSL between normal and excessively sleepy subjects that did not allow for discrimination between normal and sleepy populations, but their analysis was limited by variations in definitions, methodology, and the possibility of a non-normal distribution of scores.[2] MSL was square root transformed for normality, but we found a similarly large standard deviation in the MSL on MSLT in our study population. A floor effect has also been described, such that while a single MSLT administration may inadequately discriminate an absolute MSL value of 1 from 4 minutes, the clinical relevance of this distinction is uncertain.[2] Moreover, MSL values vary with different disease populations. Significantly shorter MSL values have been shown in narcolepsy allowing for strong discriminatory power in the right clinical context, but studies on MSL ranges in different disease populations are needed.[2,16]

Given these deficiencies, we chose as an approach looking at the ability of a single MSLT administration to reliably differentiate subjects who are sleepy from those who are not, using a binary definition. We used the cutpoint offered by the 2005 International Classification of Sleep Disorders that defines excessive sleepiness as a mean sleep latency < 8 min.[32] Explored as a binary measure, ICC for the MSL showed substantial agreement, with an ICC value of 0.63 for sleep latency, demonstrating substantial reliability in a single administration of the MSLT in classifying patients into the more clinically useful criterion, MSL < 8 minutes or not. However, with this cutpoint, a very high proportion of subjects (69.5% at first test administration) were considered impaired, many more than with cutpoints for other tests. Thus, one needs to question whether a MSL of 8 minutes is an optimal cutpoint, especially in the general population where normative data are lacking. While exploring binary cutpoints facilitates clinical interpretability of these findings, magnitudes of ICCs are not directly comparable between the two types of outcome definitions. This is due to differences in the statistical models used to characterize between- and within-subject variances and the dependency of binary ICC on somewhat arbitrary cutpoints.

Even explored as a binary measure, i.e., separating individuals into impaired or not impaired, the PVT continued to show stronger between-test agreement than the MSLT, with an ICC value of 0.85. The PVT has become an accepted measure of sustained attention, with multiple studies now demonstrating sensitivity to the effects of sleep deprivation, stimulants, as well

as impaired alertness associated with OSA.[6,25] Our results were similar to a prior test-retest ICC of 0.888 ($P < 0.0001$) for number of lapses and 0.826 for median response time.[25,34] Moreover, the standard 10-minute test is easy to use, is only minimally influenced by aptitude, has little learning effect, and is sensitive.[25,35] These data, as well as our study results, support more widespread use of the PVT in assessing impairment due to sleepiness.[25] Basner and Dinges recently explored the ability of various PVT metrics to differentiate sleep deprived from alert subjects and observed higher effect sizes for number of lapses and the reciprocal metrics, mean 1/response time, and mean slowest 10% 1/response time, compared to mean and median response times, supporting the use of these specific PVT metrics for future studies.[35] While a PVT of shorter duration offers even greater practicality, studies comparing PVT of shorter duration to the conventional 10-minute test have shown reduced sensitivity to the effects of sleepiness for certain PVT outcome variables.[35,37-39]

Table 7—Range [min, max] of correlation of measures on 4 different test assessments

TIME 1		ESS	MSLT	PVT	DADT	
		ESS	Sleep Latency	Number of Lapses	Mean Tracking Error 10 Min	Mean Tracking Error 20 Min
ESS	ESS	1.000				
MSLT	Sleep latency	**[−0.270, −0.195]**	1.000			
PVT	Number of lapses	[0.039, 0.142]	[−0.111, −0.008]	1.000		
DADT	Mean tracking error 10 min	[0.046, 0.140]	[−0.114, −0.046]	**[0.264, 0.408]**	1.000	
	Mean tracking error 20 min	[0.053, 0.145]	[−0.123, −0.069]	**[0.307, 0.457]**	**[0.977, 0.985]**	1.000

Correlation coefficient range, [minimum, maximum] between the different tests of sleepiness over the 4 test administrations. Bold = statistically significant, $P < 0.05$, at all 4 test administrations.

There are substantially fewer studies published on the test characteristics of the DADT when compared to the PVT and MSLT. Nevertheless, our study found the DADT to be the most reliable of the objective measures studied, with "near-perfect" ICC values for 10 minutes of 0.87 and 0.95, explored as continuous and binary variables, respectively. Juniper et al. assessed reproducibility of test results on a driving simulator, much like the DADT, between consecutive days and a median of 61 days apart in a subset of healthy young subjects. They found no significant differences between test performances at both time intervals, suggesting high test-retest reliability.[40] DADT performance has been shown to be influenced by alcohol, OSA, and CPAP treatment in OSA.[18,26,41] Studies have compared the effects of sleep deprivation to the effects of alcohol to define impairment. A mean tracking error cutpoint > 250 cm was chosen in our study to determine impairment based on studies looking at the effects of alcohol on performance; while this has some limitations, our study also demonstrated substantial between-test agreement when tracking was assessed as a continuous measure.

A major limitation to tests of sleepiness is the known sensitivity to internal and external alerting factors, including circadian and homeostatic effects.[2] A standardized protocol established by the AASM was used in our study to control for many of the potential internal and external influences. Excessive sleepiness requires exclusion of factors expected to cause sleepiness, and the AASM protocol attempts to assess for some of these expected causes that could result in misclassification of subjects, including sleep deprivation and medications. For example, urine drug and alcohol screens were performed in all subjects; subjects positive for either were excluded, given known impaired performance on PVT and DADT with excessive alcohol.[26,33] Our testing was done rigorously, and all tests were conducted in quiet, environmentally controlled rooms. Despite this, it is difficult to control for all possible factors. In trying to control for internal and external factors, it is recognized the AASM protocol is exhaustive, making it difficult to implement in routine clinical care; but without studies comparing simpler protocols to existing ones, it is currently recommended that clinical application of tests of sleepiness employ similar control of the testing environment.

Given the known circadian variations in task performance,[42,43] we particularly explored time of day effects in our study. Significant time of day effect was seen in the MSLT and PVT. Zwyghuizen-Doorenbos et al. found a correlation coefficient $r = 0.65$ ($P < 0.01$) when combining morning tests (10:00 and 12:00) between days and a correlation coefficient $r = 0.79$ ($P < 0.008$) combining afternoon tests (14:00 and 16:00) between days, but found single test-to-test correlations were only significant for the 10:00 test ($r = 0.78$, $P < 0.01$).[36] The greatest differences between latencies in our study were seen in test administrations that were the furthest apart in time, but the average absolute difference in MSL latency was small (approximately 102 sec). Similarly, the average differences in PVT response time and number of lapses between test administrations, while statistically significant, were again small. The clinical significance of this is uncertain, but support administration of single tests around the same time at each clinical visit. Without a test-retest experiment across similar time points, our study could not address the optimal timing of single test administration. No time of day effect was seen in the DADT. While significant circadian time of day effects have been demonstrated in healthy subjects with repeated assessments across the day, including during sleep deprivation,[5,43-46] sleepiness and performance are known to be influenced by both homeostatic sleep drive and endogenous circadian variation. In individuals impaired by different sleep duration and/or OSA, the variation in sleepiness between in-

dividuals is expected to be large and potentially greater than the circadian oscillations within a subject.

Differences were also seen in the percentage impaired in our study population using the chosen binary cutpoints between the different objective tests across test administrations, with 69.6% impaired using a MSL cutpoint < 8 minutes on MSLT compared to 15.1% using > 3 lapses on PVT and 30.4% using > 250 cm tracking error on 10-minute DADT at first test administration. These differences may have been due to the specific cutpoints used in our study to define impairment. Sleep latency on MSLT is likely overly sensitive using a cutpoint of 8 minutes. Alternatively, these differences in impairment may be capturing different domains of sleepiness, reinforcing the multidimensional nature of sleepiness.[22] This is further supported by the weak correlation between the different tests of sleepiness.

When comparing the different measures of sleepiness, the strongest correlation was between the ESS and MSLT, and the DADT and PVT. This is not surprising, given that the ESS and MSLT measure tendency to fall asleep, while the DADT and PVT measure neurocognitive functional deficits related to sleepiness. Prior studies comparing different measures of sleepiness have been conflicting, but overall show poor agreement between objective and subjective measures of sleepiness, largely in populations with excessive daytime sleepiness, usually from OSA and narcolepsy.[2,19-24] Fong et al. found significantly shorter mean sleep latency in patients with severe OSA compared to moderate and mild OSA, but only weak correlation between the MSL and ESS score ($\rho = -0.149$), with no significant correlation between the ESS and OSA severity.[21] The poor correlation between the ESS and different objective measures of sleepiness does not support the use of the ESS alone as a means to assess sleepiness. While we found some correlation between the DADT and PVT, we found little correlation with the MSLT. George et al. found only a weak relation between the MSL and tracking error on the DADT, $r = -0.42$ ($P = 0.01$) in OSA patients.[47] Our study was not designed to address validity of each of these individual objective measures, an area in need of future research.

How performance on neurocognitive and simulation tests translate into real life performance is uncertain.[18,48-52] Patients with OSA are more likely to have crashes,[53] but OSA patients who have had crashes have not consistently been shown to self-report more sleepiness on ESS or have shorter MSL than OSA patients without accidents, again highlighting the need for better objective predictors.[7,54-57] While Drake et al. demonstrated a significant linear relation between MSL and the prevalence of verified motor vehicle crashes ($P < 0.05$),[58] the predictive validity of the MSL for safety is not established.[2,42,48,56] Mazza et al. compared driving performance using a more real-life road safety platform test with several in-laboratory objective measures, including the DADT, in OSA patients before and after treatment.[59] Driving performance was significantly impaired in OSA patients and improved following continuous positive airway pressure treatment. Association between mean reaction time in real-life driving condition performance and reaction time in the DADT was seen, but no other associations between real-life driving condition performance and in-laboratory objective measures or between any of the objective tests and the ESS score were observed.[59] Thus, currently these laboratory tests cannot be used to determine who is particularly at risk for motor vehicle crashes.

Our study was restricted to holders of commercial drivers' licenses and consisted largely of Caucasian, middle-aged, overweight males, limiting generalizability of our findings to the general population. However, this is also the population at risk for OSA, and one in which a single reliable objective measure of sleepiness is of public health interest. The small difference in distribution of AHI severity in our study population compared to the study by Gurubhagavatula et al.[27] and Pack et al.[28] was due to our sample size being limited only to those subjects completing all three objective tests of sleepiness. While our study oversampled for patients at higher risk for OSA, the average AHI was only 5.3 events/hour. Further studies in a clinical population are needed. Our study was also conducted among holders of commercial drivers' licenses, where anxiety over employment or legal consequences may have been present. However, all participants in this study were informed that all information in this research study was confidential and would not be shared with their employer or licensing authority. Given sleep tendency can be affected by anxiety or tension,[2] the role of any measure of sleepiness in the occupational and driving setting remains uncertain. At present, we do not advocate use of these tests for pre-employment screening.

In conclusion, excessive sleepiness is common and a growing problem with multiple causes and adverse outcomes, making identification and characterization of sleepiness clinically relevant. Traditionally, objective measures of sleepiness have relied on multiple test administrations, making them impractical for routine clinical use. We compared single to multiple test administrations of various objective measures of sleepiness, and found a single administration of the 10-minute DADT and PVT reliable. A single administration of the MSLT was reasonably reliable in discriminating individuals with MSL < 8 minutes. These findings support the use of single administration of some objective tests of sleepiness in clinical care when performed under standardized conditions to try and control for internal and external influences. Future studies should elucidate their effectiveness for identifying individuals who are impaired, and for following response to treatment.

ACKNOWLEDGMENTS

The authors thank David F. Dinges, PhD, for his contribution to the original study design and the current manuscript and Mr. Daniel C. Barrett in the preparation of this manuscript. This research was supported by a contract from the Trucking Research Institute, American Trucking Association (DTFH61-93-C-00088) funded by the Federal Highway Administration (now the Federal Motor Carriers Safety Administration), NIH grants HL094307, T32 HL07713 and R01 OH009149.

DISCLOSURE STATEMENT

This was not an industry supported study. Dr. George is a member of the medical advisory board for SleepTech LLC. Dr. Gurubhagavatula received equipment on loan from Embra Corporation for research. Dr. Maislin is the principal biostatistician of Biomedical Statistical Consulting. The other authors have indicated no financial conflicts of interest.

REFERENCES

1. Ohayon MM. From wakefulness to excessive sleepiness: what we know and still need to know. Sleep Med Rev 2008;12:129-41.
2. Arand D, Bonnet M, Hurwitz T, Mitler M, Rosa R, Sangal RB. The clinical use of the MSLT and MWT. Sleep 2005;28:123-44.
3. Balkin TJ, Rupp T, Picchioni D, Wesensten NJ. Sleep loss and sleepiness: current issues. Chest 2008;134:653-60.
4. Knutson KL, Spiegel K, Penev P, Van Cauter E. The metabolic consequences of sleep deprivation. Sleep Med Rev 2007;11:163-78.
5. Dinges DF, Pack F, Williams K, 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;20:267-7.
6. Lim J, Dinges DF. Sleep deprivation and vigilant attention. Ann N Y Acad Sci 2008;1129:305-22.
7. Ellen RL, Marshall SC, Palayew M, Molnar FJ, Wilson KG, Man-Son-Hing M. Systematic review of motor vehicle crash risk in persons with sleep apnea. J Clin Sleep Med 2006;2:193-200.
8. Rajaratnam SM, Arendt J. Health in a 24-h society. Lancet 2001;358:999-1005.
9. Kapur VK, Resnick HE, Gottlieb DJ. Sleep disordered breathing and hypertension: does self-reported sleepiness modify the association? Sleep 2008;31:1127-32.
10. Guilleminault C, Brooks SN. Excessive daytime sleepiness: a challenge for the practising neurologist. Brain 2001;124:1482-91.
11. Wise MS. Objective measures of sleepiness and wakefulness: application to the real world? J Clin Neurophysiol 2006;23:39-49.
12. Dinges DF Kribbs NB. Performing while sleepy: Effects of experimentally-induced sleepiness. Monk TH, ed. Sleep, sleepiness and performance. 1991:97-128.
13. Van Dongen HP, Maislin G, Dinges DF. Dealing with inter-individual differences in the temporal dynamics of fatigue and performance: importance and techniques. Aviat Space Environ Med 2004;75:A147-54.
14. Van Dongen HP, Vitellaro KM, Dinges DF. Individual differences in adult human sleep and wakefulness: Leitmotif for a research agenda. Sleep 2005;28:479-96.
15. Carper DL, Lewandowski DJ. Assessment of obstructive sleep apnea risk and severity in truck drivers: commentary on the legal implications for ignoring a national safety concern. Sleep Diag Ther 2007;3:27-30.
16. Littner MR, Kushida C, Wise M, et al. Practice parameters for clinical use of the multiple sleep latency test and the maintenance of wakefulness test. Sleep 2005;28:113-21.
17. George CF. Sleep apnea, alertness, and motor vehicle crashes. Am J Respir Crit Care Med 2007;176:954-6.
18. George CF. Sleep. 5: Driving and automobile crashes in patients with obstructive sleep apnoea/hypopnoea syndrome. Thorax 2004;59:804-7.
19. Olson LG, Cole MF, Ambrogetti A. Correlations among Epworth Sleepiness Scale scores, multiple sleep latency tests and psychological symptoms. J Sleep Res 1998;7:248-53.
20. Johns MW. Sensitivity and specificity of the multiple sleep latency test (MSLT), the maintenance of wakefulness test and the epworth sleepiness scale: failure of the MSLT as a gold standard. J Sleep Res 2000;9:5-11.
21. Fong SY, Ho CK, Wing YK. Comparing MSLT and ESS in the measurement of excessive daytime sleepiness in obstructive sleep apnoea syndrome. J Psychosom Res 2005;58:55-60.
22. Kim H, Young T. Subjective daytime sleepiness: dimensions and correlates in the general population. Sleep 2005;28:625-34.
23. Benbadis SR, Mascha E, Perry MC, Wolgamuth BR, Smolley LA, Dinner DS. Association between the Epworth sleepiness scale and the multiple sleep latency test in a clinical population. Ann Intern Med 1999;130:289-92.
24. Pizza F, Contardi S, Mondini S, Trentin L, Cirignotta F. Daytime sleepiness and driving performance in patients with obstructive sleep apnea: comparison of the MSLT, the MWT, and a simulated driving task. Sleep 2009;32:382-91.
25. Dorrian J, Rogers NL, Dinges DF. Psychomotor vigilance perfomance: neurocognitive assay sensitive to sleep loss In: Kushida C, ed. Sleep deprivation: clinical issues, pharmacology and sleep loss effects. New York, Marcel Dekker, 2005:39-70.
26. George CF, Boudreau AC, Smiley A. Simulated driving performance in patients with obstructive sleep apnea. Am J Respir Crit Care Med 1996;154:175-81.
27. Gurubhagavatula I, Maislin G, Nkwuo JE, Pack AI. Occupational screening for obstructive sleep apnea in commercial drivers. Am J Respir Crit Care Med 2004;170:371-6.
28. Pack AI, Maislin G, Staley B, et al. Impaired performance in commercial drivers: role of sleep apnea and short sleep duration. Am J Respir Crit Care Med 2006;174:446-54.
29. Maislin G, Pack AI, Kribbs NB, et al. A survey screen for prediction of apnea. Sleep 1995;18:158-66.
30. Iber C, Ancoli-Israel S, Chesson Jr AL, Quan SF. The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications, 1st ed. Westchester, IL: American Academy of Sleep Medicine, 2007.
31. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-74.
32. American Academy of Sleep Medicine. International classification of sleep disorders. diagnostic and coding manual. 2nd ed. Westchester, IL: American Academy of Sleep Medicine, 2005.
33. Powell NB, Riley RW, Schechtman KB, Blumen MB, Dinges DF, Guilleminault C. A comparative model: reaction time performance in sleep-disordered breathing versus alcohol-impaired controls. Laryngoscope 1999;109:1648-54.
34. Van Dongen HPA MG, Mullington JM, Dinges DF. The cumulative cost of additional wakefulness: Dose-response effects on neurobehavioral functions and sleep physiology from chronic sleep restriction and total sleep deprivation. Sleep 2003;26:117-26.
35. Basner M, Dinges DF. Maximizing sensitivity of the Psychomotor Vigilance Test (PVT) to sleep loss. Sleep 2011;34:581-91.
36. Zwyghuizen-Doorenbos A, Roehrs T, Schaefer M, Roth T. Test-retest reliability of the MSLT. Sleep 1988;11:562-5.
37. Loh S, Lamond N, Dorrian J, Roach G, Dawson D. The validity of psychomotor vigilance tasks of less than 10-minute duration. Behav Res Methods Instrum Comput 2004;36:339-46.
38. Lamond N, Jay SM, Dorrian J, Ferguson SA, Roach GD, Dawson D. The sensitivity of a palm-based psychomotor vigilance task to severe sleep loss. Behav Res Methods 2008;40:347-52.
39. Roach GD, Dawson D, Lamond N. Can a shorter psychomotor vigilance task be used as a reasonable substitute for the ten-minute psychomotor vigilance task? Chronobiol Int 2006;23:1379-87.
40. Juniper M, Hack MA, George CF, Davies RJ, Stradling JR. Steering simulation performance in patients with obstructive sleep apnoea and matched control subjects. Eur Respir J 2000;15:590-5.
41. George CF, Boudreau AC, Smiley A. Effects of nasal CPAP on simulated driving performance in patients with obstructive sleep apnoea. Thorax 1997;52:648-53.
42. Richardson GS, Carskadon MA, Orav EJ, Dement WC. Circadian variation of sleep tendency in elderly and young adult subjects. Sleep 1982;5 Suppl 2:S82-94.
43. Dijk DJ, Czeisler CA. Contribution of the circadian pacemaker and the sleep homeostat to sleep propensity, sleep structure, electroencephalographic slow waves, and sleep spindle activity in humans. J Neurosci 1995;15:3526-38.
44. Mollicone DJ, Van Dongen HP, Rogers NL, Banks S, Dinges DF. Time of day effects on neurobehavioral performance during chronic sleep restriction. Aviat Space Environ Med 2010;81:735-44.
45. Akerstedt T, Hume K, Minors D, Waterhouse J. Experimental separation of time of day and homeostatic influences on sleep. Am J Physiol 1998;274:R1162-8.
46. Silva EJ, Wang W, Ronda JM, Wyatt JK, Duffy JF. Circadian and wake-dependent influences on subjective sleepiness, cognitive throughput, and reaction time performance in older and young adults. Sleep;33:481-90.
47. George CF, Boudreau AC, Smiley A. Comparison of simulated driving performance in narcolepsy and sleep apnea patients. Sleep 1996;19:711-7.
48. Philip P, Sagaspe P, Taillard J, et al. Fatigue, sleepiness, and performance in simulated versus real driving conditions. Sleep 2005;28:1511-6.
49. Findley L, Unverzagt M, Guchu R, Fabrizio M, Buckner J, Suratt P. Vigilance and automobile accidents in patients with sleep apnea or narcolepsy. Chest 1995;108:619-24.
50. Pichel F, Zamarron C, Magan F, Rodriguez JR. Sustained attention measurements in obstructive sleep apnea and risk of traffic accidents. Respir Med 2006;100:1020-7.
51. Barbe, Pericas J, Munoz A, Findley L, Anto JM, Agusti AG. Automobile accidents in patients with sleep apnea syndrome. An epidemiological and mechanistic study. Am J Respir Crit Care Med 1998;158:18-22.

52. Turkington PM, Sircar M, Allgar V, Elliott MW. Relationship between obstructive sleep apnoea, driving simulator performance, and risk of road traffic accidents. Thorax 2001;56:800-5.
53. Sassani A, Findley LJ, Kryger M, Goldlust E, George C, Davidson TM. Reducing motor-vehicle collisions, costs, and fatalities by treating obstructive sleep apnea syndrome. Sleep 2004;27:453-8.
54. Teran-Santos J, Jimenez-Gomez A, Cordero-Guevara J. The association between sleep apnea and the risk of traffic accidents. Cooperative Group Burgos-Santander. N Engl J Med 1999;340:847-51.
55. George CF, Smiley A. Sleep apnea & automobile crashes. Sleep 1999;22:790-5.
56. Young T, Blustein J, Finn L, Palta M. Sleep-disordered breathing and motor vehicle accidents in a population-based sample of employed adults. Sleep 1997;20:608-13.
57. Birketvedt GS, Florholmen J, Sundsfjord J, et al. Behavioral and neuroendocrine characteristics of the night-eating syndrome. JAMA 1999;282:657-63.
58. Drake C, Roehrs T, Breslau N, et al. The 10-year risk of verified motor vehicle crashes in relation to physiologic sleepiness. Sleep 2010;33:745-52.
59. Mazza S, Pepin JL, Naegele B, et al. Driving ability in sleep apnoea patients before and after CPAP treatment: evaluation on a road safety platform. Eur Respir J 2006;28:1020-8.

SUPPLEMENT

Population Sample

In total, 372 of 406 subjects with polysomnography (n = 247 [44.8%] of 551 highest risk drivers for apnea and n = 159 [20.4%] of 778 lower risk drivers) performed all 3 objective tests of sleepiness—MSLT, PVT and DADT—with complete data available for analysis. The study population was predominantly obese, middle-aged, Caucasian males, with a mean (SD) AHI of 5.3 (9.8) events/h. The subjects missing data were more obese than subjects with complete data, mean (SD) BMI 34.1(6.7) compared to 30.2 (5.6) (P = 0.001), but otherwise similar in demographics (Table S1).

Single versus Multiple Test Administrations

Using the intraclass correlation coefficient (ICC) to determine agreement within-subject between-test administrations, the MSLT, PVT, and DADT were explored as both continuous and binary measures. Cutpoints of <8 min for MSL, >3 lapses for PVT, and >250 cm for tracking error on DADT were used to define impairment. The MSLT had a number of subjects who did not fall asleep during the testing period (Table S2). The frequency of subjects who did not fall asleep during nap opportunities increased across test administrations from 9.9% at time 1 to 21.7% at time 4. Additionally, 19.4% of those who did not fall asleep on their first test administration did not fall asleep on follow-up test administrations (Table S2).

As continuous measures, ICCs were 0.45 for MSLT, 0.69 for median response time, 0.51 for number of lapses and 0.60 for transformed number of lapses on PVT, and 0.89 for mean tracking error at 10 min for DADT. As binary measures, there were differences in the percentage impaired in the study population, using the chosen cutpoints described above, between the different objective tests across test administrations (Table S3): 69.6% were impaired using MSL <8 min on MSLT at 1st test administration compared to 15.1% using >3 lapses on PVT and 30.4% using >250 cm tracking error on 10-min DADT (Table S3). These differences were smaller by the 4th test administration. Sleep latency on MSLT may be overly sensitive using a cutpoint of 8 min, or these differences may be capturing different domains of sleepiness, further supported by weak correlation between the different tests of sleepiness described below.

Correlation between Subjective and Different Objective Measures of Sleepiness

Correlation between the Epworth Sleepiness Scale and the different objective measures of sleepiness, as well as between the different individual objective measures, was performed across all 4 time points. Supplemental Table S4a-S4d shows correlation between the different measures of sleepiness at each test administration from time 1 (Table S4a) to time 4 (Table S4d). The correlation between the ESS and objective tests of sleepiness was strongest for the MSLT with correlation ranging from −0.270 to −0.195 across test administrations. While there was moderate correlation between the DADT and PVT, there was weak correlation between the MSLT and other objective measures across all time points.

Time of day effects on the magnitude of correlation between subjective and the different objective measures and between the different objective tests of sleepiness was explored by comparing correlation between the 4 test administrations (Table S5). Correlation between the ESS and MSLT did not significantly change when assessed across different times of test administration (P = 0.35) (Table S5).

Table S1—Demographic information for subjects with and without missing data

	Non-Missing	Missing	P value
Age, Mean ± SD	45.7 ± 11.3	47.2 ± 9.7	0.406
BMI, Mean ± SD	30.2 ± 5.6	34.1 ± 6.7	0.001
Sex, %Male	94.0	100	
Race, %White	85.1	81.0	0.498

Characteristics of study population, with missing and no missing data.

Table S2—MSLT exploration of normal group latency = 1200 sec

		Time 1	Time 2	Time 3	Time 4
MSLT	% with Normal Latency N1 (MSL ≥1200 sec)	9.9	10.0	11.4	21.7

		Time 1	Normal on All MSLT Times	
			%Of Total	%of Time 1 Sample
MSLT	Normal Latency N1 (MSL ≥1200 sec)	9.9	1.9	19.4

Percentage of study population with normal MSLT defined by N1 sleep latency >1200 sec across test administrations from time 1 to time 4 (top), and percentage of study population with normal MSLT at all test administrations who were normal at first test administration (bottom).

Table S3—Objective test results across test administrations: binary

		Time 1	Time 2	Time 3	Time 4
MSLT	% Impaired (Latency <8 Min)	69.6	63.6	55.9	49.3
PVT	% Impaired (Lapses >3)	15.1	26.1	26.4	23.0
DADT	% Impaired 10 min (mean tracking error >250 cm)	30.4	32.1	37.6	37.4
	% Impaired 20 min (mean tracking error >250 cm)	31.8	33.4	36.0	36.3

Percentage impaired for each individual objective test in the overall study population at each test administration across time

Table S4A—Correlation of measures: Time 1

TIME 1		**ESS**	**MSLT**	**PVT**	**DADT**	
		ESS	**Sleep Latency**	**Number of Lapses**	**Mean Tracking Error 10 Min**	**Mean Tracking Error 20 Min**
ESS	**ESS**	1.000				
MSLT	**Sleep latency**	**−0.195***	1.000			
PVT	**Number of lapses**	0.039	−0.008	1.000		
DADT	**Mean tracking error 10 min**	0.084	**−0.114***	**0.264***	1.000	
	Mean tracking error 20 min	0.100	**−0.123***	**0.307***	**0.979***	1.000

Correlation between subjective and different objective tests of sleepiness and between different objective tests of sleepiness at 1st test administration. (bold = statistically significant, *P < 0.05).

Table S4B—Correlation of measures: Time 2

TIME 2		ESS	MSLT	PVT	DADT	
		ESS	Sleep Latency	Number of Lapses	Mean Tracking Error 10 Min	Mean Tracking Error 20 Min
ESS	ESS	1.000				
MSLT	Sleep latency	**−0.248***	1.000			
PVT	Number of lapses	**0.142***	−0.050	1.000		
DADT	Mean tracking error 10 min	**0.120***	−0.046	**0.331***	1.000	
	Mean tracking error 20 min	**0.126***	−0.069	**0.350***	**0.985***	1.000

Correlation between subjective and different objective tests of sleepiness and between different objective tests of sleepiness at 2nd test administration. (bold = statistically significant, *P < 0.05).

Table S4C—Correlation of measures: Time 3

TIME 3		ESS	MSLT	PVT	DADT	
		ESS	Sleep Latency	Number of Lapses	Mean Tracking Error 10 Min	Mean Tracking Error 20 Min
ESS	ESS	1.000				
MSLT	Sleep latency	**−0.203***	1.000			
PVT	Number of lapses	0.046	**−0.111***	1.000		
DADT	Mean tracking error 10 min	**0.140***	**−0.093***	**0.397***	1.000	
	Mean tracking error 20 min	**0.145***	**−0.117***	**0.407***	**0.977***	1.000

Correlation between subjective and different objective tests of sleepiness and between different objective tests of sleepiness at 3rd test administration. (bold = statistically significant, *P < 0.05).

Table S4D—Correlation of measures: Time 4

TIME 4		ESS	MSLT	PVT	DADT	
		ESS	Sleep Latency	Number of Lapses	Mean Tracking Error 10 Min	Mean Tracking Error 20 Min
ESS	ESS	1.000				
MSLT	Sleep latency	**−0.270***	1.000			
PVT	Number of lapses	0.091	−0.049	1.000		
DADT	Mean tracking error 10 min	0.046	**−0.081***	**0.408***	1.000	
	Mean tracking error 20 min	0.053	**−0.093***	**0.457***	**0.979***	1.000

Correlation between subjective and different objective tests of sleepiness and between different objective tests of sleepiness at 4th test administration. (bold = statistically significant, *P < 0.05).

Table S5—Correlation between different tests of sleepiness across time

P Value for order effects of correlation		ESS	Sleep Latency	Number of Lapses	Mean Tracking Error
ESS	ESS				
MSLT	Sleep latency	0.350			
PVT	Number of lapses	0.092	0.567		
DADT	Mean tracking error 10 min	0.109	0.555	0.424	
	Mean tracking error 20 min	**0.031**	0.699	0.162	**< 0.001**

P value for change in correlation between different tests of sleepiness across time (bold = statistically significant, i.e. P < 0.05).