Received: 25 March 2023 | Revised: 6 May 2023 | Accepted: 20 May 2023

DOI: 10.1111/jsr.13961

RESEARCH ARTICLE

Revisiting the maintenance of wakefulness test: from intra-/inter-scorer agreement to normative values in patients treated for obstructive sleep apnea

Pierre Tankéré[1,2] | Jacques Taillard[3,4] | Marc-Antoine Armeni[2] | Thierry Petitjean[2] | Christian Berthomier[5] | Mélanie Strauss[6,7] | Laure Peter-Derex[2,8,9]

[1]Reference Center for Rare Pulmonary Diseases, Pulmonary Medicine and Intensive Care Unit, Dijon University Hospital, Dijon, France

[2]Center for Sleep Medicine and Respiratory Disease, Croix-Rousse Hospital, Hospices Civils de Lyon, Lyon, France

[3]Sommeil, Addiction et Neuropsychiatrie, Université de Bordeaux, SANPSY, USR 3413, Bordeaux, France

[4]CNRS, SANPSY, USR 3413, Bordeaux, France

[5]Physip, Paris, France

[6]Hôpital Universitaire de Bruxelles, Site Erasme, Services de Neurologie, Psychiatrie et Laboratoire du Sommeil, Université Libre de Bruxelles, Brussels, Belgium

[7]Neuropsychology and Functional Imaging Research Group (UR2NF), Center for Research in Cognition and Neurosciences and ULB Neuroscience Institute, Université Libre de Bruxelles, Brussels, Belgium

[8]Lyon Neuroscience Research Center, PAM Team, INSERM U1028, CNRS UMR 5292, Lyon, France

[9]Claude Bernard Lyon 1 University, Lyon, France

Correspondence
Laure Peter-Derex, Centre for Sleep Medicine and Respiratory Diseases, Croix-Rousse Hospital, University Hospital of Lyon, 103 Grande rue de la Croix-Rousse, 69004 Lyon, France.
Email: laure.peter-derex@chu-lyon.fr; laure.peter-derex@univ-lyon1.fr

Summary

The Maintenance of Wakefulness Test is widely used to objectively assess sleepiness and make safety-related decisions, but its interpretation is subjective and normative values remain debated. Our work aimed to determine normative thresholds in non-subjectively sleepy patients with well-treated obstructive sleep apnea, and to assess intra- and inter-scorer variability. We included maintenance of wakefulness tests of 141 consecutive patients with treated obstructive sleep apnea (90% men, mean (SD) age 47.5 (9.2) years, mean (SD) pre-treatment apnea–hypopnea index of 43.8 (20.3) events/h). Sleep onset latencies were independently scored by two experts. Discordant scorings were reviewed to reach a consensus and half of the cohort was double-scored by each scorer. Intra- and inter-scorer variability was assessed using Cohen's kappa for 40, 33, and 19 min mean sleep latency thresholds. Consensual mean sleep latencies were compared between four groups according to subjective sleepiness (Epworth Sleepiness Scale score < versus ≥11) and residual apnea–hypopnea index (< versus ≥15 events/h). In well-treated non-sleepy patients ($n = 76$), the consensual mean (SD) sleep latency was 38.4 (4.2) min (lower normal limit [mean − 2SD] = 30 min), and 80% of them did not fall asleep. Intra-scorer agreement on mean sleep latency was high but inter-scorer was only fair (Cohen's kappa 0.54 for 33-min threshold, 0.27 for 19-min threshold), resulting in changes in latency category in 4%–12% of patients. A higher sleepiness score but not the residual apnea–hypopnea index was significantly associated with a lower mean sleep latency. Our findings suggest a higher than usually accepted normative threshold (30 min) in this context and emphasise the need for more reproducible scoring approaches.

KEYWORDS
driving, security, sleep scoring, sleepiness, threshold, vigilance

1 | INTRODUCTION

Excessive daytime sleepiness affects >5% of the general population, and may contribute to up to 30% of fatal car accidents (Bioulac et al., 2017; Dinges, 1995; Ohayon, 2008). Among causes of abnormal sleepiness, obstructive sleep apnea (OSA) is frequently involved in traffic and occupational accidents, as well as impaired productivity at the workplace (Mulgrew et al., 2007; Ward et al., 2013). Given the growing prevalence of OSA associated with the obesity epidemic, and the persistent under diagnosis of OSA, it is even thought to be the leading cause of accidents at work and on the road (Garbarino et al., 2016; Peppard et al., 2013). Remarkably, OSA treatment, particularly continuous positive airway pressure (CPAP), significantly and possibly rapidly reduces the risk of a car accident among drivers with OSA (Tregear et al., 2010). This implies a major need not only for diagnosis and treatment of OSA, but also for reliable tools to assess sleepiness objectively in patients with OSA, especially following CPAP or oral appliance (OA) therapy. This assessment is particularly relevant considering the recent entry into the marketplace of wake-promoting medications targeting residual sleepiness in treated OSA (Pépin et al., 2021; Strollo et al., 2019).

Since it was first described 40 years ago, the Maintenance of Wakefulness Test (MWT) has become a 'gold standard' measure for objective sleepiness (Mitler et al., 1982). It is mainly used for the evaluation of the efficacy of treatment for sleepiness and/or the assessment of vigilance in professionals whose sleepiness may impact public security (Littner et al., 2005; Wise, 2006). The MWT is included in both French and European Legislation as part of sleepiness evaluation for decisions related to occupational driving. However, no mention is made of the critical thresholds associated with 'normal' or 'expected' vigilance, and the lack of data is clearly stated in the American Academy of Sleep Medicine 2021 recommendations for the MWT (Krahn et al., 2021). Thus, the 8-min threshold is accepted as a lower normal limit, but no sleep onset in any of the four tests is recommended for maximum safety level jobs (Krahn et al., 2021; Sullivan & Kushida, 2008). This paradox reflects both the lack of robust normative values for the MWT and the imperfect correlation between the MWT and driving performance both on driving simulators and in real life (Banks et al., 2005; Bijlenga et al., 2022; Littner et al., 2005; Pizza et al., 2009; Schreier et al., 2018). This also explains a huge variability in MWT interpretation. In France, the 19-min threshold is accepted in most centres based on the normative values obtained in 64 healthy individuals (Doghramji et al., 1997). However, several centres use the 33-min threshold, which has been associated with normal driving performance in patients with OSA (Philip et al., 2008).

In addition, MWT interpretation relies on human scoring, which raises legitimate questions about intra- and inter-scorer reliability, as well as inter-centre agreement in MWT results. The determination of sleep onset latency in the MWT relies on the detection of any-stage first epoch of sleep (Krahn et al., 2021). This first epoch is usually N1 stage, which is the very stage associated with the lowest inter-scorer agreement (Lee et al., 2022). Even if, in the context of the MWT, the shorter duration of scoring and an increased attention paid to the first sleep epoch may influence this agreement, the reproducibility of the MWT scoring remains questionable. These limitations are a major issue in clinical practice, with potential crucial consequences for patients in terms of work ability, but also safety for themselves and for others. Finally, data are lacking about both reproducibility in MWT scoring and normative values in wide clinical populations of interest, i.e., treated patients with OSA.

In this study, leveraging a large database of MWTs performed in treated patients with OSA and the contribution of two major sleep centres, we aimed to: (i) assess intra- and inter-scorer scoring variability in the MWT, and (ii) provide normative values based on a double scoring in well-treated patients with OSA without subjective residual sleepiness. We hypothesised that: (i) several patients might change 'driving authorisation' category according to the scorer, and (ii) normative values would be higher than the values usually used in French sleep centres.

2 | METHODS

2.1 | Participants

In this retrospective study, we reviewed the files of all patients with OSA hospitalised to undergo both polysomnography (PSG) and the 40-min MWT in the Center for Sleep Medicine and Respiratory Disease, Lyon Academic Hospital from September 2017 to March 2020. No patient had undergone a MWT in the past. Exclusion criteria were: age <18 years, refusal to participate in the study, diagnosis of central disorder of hypersomnolence, and missing data on the MWT, Epworth Sleepiness Scale (ESS) or apnea–hypopnea index (AHI).

2.2 | Recordings

Full-night PSG recordings were conducted in the Center for Sleep Medicine and Respiratory Disease, Lyon University Hospital. Due to French legislation, the vast majority of patients were on sick leave due to excessive daytime sleepiness; MWTs were mandatory before resuming work, as required by the occupational health physician. All patients had therefore a normal and regular sleep–wake rhythm. Patients arrived in the late afternoon and underwent instrumentation for the electrodes and sensors required for PSG. The following signals were recorded: electroencephalogram (Fp2, C4, O2, T4, Cz, Pz, A1, A2), electro-oculogram, chin and tibialis electromyogram, electrocardiography, nasal airflow (nasal pressure and thermistor), pulse oximetry, and respiratory efforts (thoracic and abdominal belts). The PSGs were performed under current treatment for sleep apnea (CPAP or OA). Patients remained on their usual chronic medication, abstained from sedating substances such as alcohol and marijuana on the day of the test, and typical caffeine use was allowed according to recommendations (Krahn et al., 2021). Stimulating activities such as consuming nicotine and the use of electronic devices and cell phones should end at least 30 min before each wake trial. Recordings were performed using

a Deltamed® Natus Amplifier device. Bedtime was decided by patients, but they were woken up at 7:00 a.m. to ensure sufficient time before the first MWT. Four MWT trials were conducted according to current international guidelines (Krahn et al., 2021; Littner et al., 2005) the day following the PSG recording at 9:00 a.m., 11:00 a.m., 1:00 p.m., and 3:00 p.m., with the first test beginning 1.5–3 h after the end of the PSG recording. Each MWT trial ended 'once the patient has three consecutive epochs of stage N1 sleep or one epoch of any other sleep stage or after 40 min' (Krahn et al., 2021).

2.3 | Sleep recording analysis

- Night PSG sleep scoring was performed according to the American Academy of Sleep Medicine (AASM) scoring rules (Berry et al., 2017). The main sleep parameters were extracted: total time of sleep, sleep onset latency, rapid eye movement (REM) sleep latency, wake after sleep onset, sleep efficiency, percentage and duration of sleep stages (N1, N2, N3, R), arousal index, AHI, index of >3% desaturation, and time spent with an arterial oxygen saturation <90%.
- The MWT latencies were scored independently by two board-certified experts in sleep medicine (LP-D and JT), both with >15 years of daily scoring experience in two sleep regional reference centres and involved as teachers in the National Post-graduate Diploma in Sleep Medicine. The sleep onset latency, defined as 'the time from lights out until the start of the first epoch of any stage of sleep (an epoch of N1, N2, N3, or R)' for each of the four trials and the mean sleep latency (MSL) were assessed (Krahn et al., 2021). In case of discordance, the MWT latencies were reviewed by the two scorers to reach a consensual scoring (consensual MSL [cMSL]) >6 months after the initial assessment and without knowledge of the initial scores. Moreover, half of the MWT latencies were scored twice (with a 6-month interval) by each of the two scorers to assess intra-scorer agreement.

2.4 | Data collection

Data collected included demographic and anthropomorphic data such as age, sex, body mass index (BMI), smoking history, cardiovascular comorbidities, depression history, medications, estimated time in bed, working time, MWT context, and history of motor vehicle accident or near miss due to sleepiness.

Several scales were also available including the ESS at OSA diagnosis and at the time of the MWT recording, Horne and Ostberg scale, Pichot scale, Beck Depression Inventory (BDI), and the Observation and Interview-based Diurnal Sleepiness Inventory (ODSI) (Beck et al., 1961; Horne & Ostberg, 1976; Johns, 1991; Onen et al., 2016; Pichot & Brun, 1984). The OSA treatment characteristics were collected: OA or CPAP and its parameters (mode, pressure level, mask, mean use, and leaks).

2.5 | Statistical analysis

Four groups of patients were defined according to the presence of residual subjective sleepiness (as assessed by the ESS at the time of the MWT) and the efficacy of CPAP treatment (as measured by the residual AHI on pre-MWT PSG). For residual AHI, we used the dichotomy <15 versus ≥15 events/h, with a residual AHI of ≥15 events/h defining a residual moderate OSA according to current AASM rules, keeping in mind that the acceptable residual AHI under CPAP treatment remains debated (Kapur et al., 2017; Li et al., 2022). Thus, the groups of patients were defined as follows: ENAN (ESS score <11, AHI <15 events/h); ENAX (ESS score <11, AHI ≥15 events/h); EXAN (ESS score ≥11, AHI <15 events/h), and EXAX (ESS score ≥11, AHI ≥15 events/h). With these four groups, our aim was to define normative values in non-sleepy well-treated patients with OSA, and then to compare them to the other groups.

Qualitative variables were described as percentages, and quantitative variables as mean and standard deviation (SD), median and interquartile range (IQR, i.e., the 25% and 75% quartiles). Patient groups were compared for clinical characteristics as well as questionnaires and PSG results using chi-square, Fisher's, Student's tests, Mann–Whitney, and Kruskal–Wallis rank sum, as appropriate. Normal distribution was tested with a Shapiro–Wilk test. A $p < 0.05$ was considered significant. All tests were two-sided. Statistical analyses were performed using R version 4.0.5 software, as well as R studio.

Inter- and intra-scorer agreements were measured with: (i) the percentage agreement, defined as the percentage of tests for which the first sleep epoch (sleep onset latency) was assigned at the same recorded epoch, and (ii) the Cohen's kappa coefficient (Cohen's κ) with distinct MSL thresholds. These thresholds were: 40 min (i.e., no sleep), 33 min, or 19 min, which are commonly used thresholds in French sleep reference centres based on the current available literature (Doghramji et al., 1997; Philip et al., 2008).

2.6 | Ethics approval

This study was conducted in accordance with the Declaration of Helsinki (as revised in 2013) and Guidelines of the International Conference on Harmonization. It was approved by the institutional ethics committee of our university Hospital – Hospices Civils de Lyon (HCL) under approval N°20-04 issued on January 13, 2020, and registered with the French data protection authority (Commission Nationale de l'Informatique et des Libertés [CNIL]), HCL record 19-327. All patients gave informed consent for the use of their data for research purposes.

3 | RESULTS

3.1 | Population characteristics and sleep parameters

A total of 141 patients (90.1% male, mean [SD] age 47.5 [9.2] years) were included. The mean (SD) BMI was 31.6 (5.9) kg/m^2 and 29.8%

had a smoking history. The mean (SD) initial AHI was 43.8 (20.3) events/h with 76.8% of patients having severe OSA (AHI >30 events/h), 20.4% moderate OSA (AHI = 15–29 events/h) and 1.8% mild OSA (<15 events/h). In 91.4% of the population, the MWT was performed in a context of work safety or driving authorisation (Table 1). When we categorised the population according to the ESS and AHI, 76 patients (53.9%) were in the ENAN group, 34 (24.1%) in the ENAX group, 21 (14.8%) in the EXAN group, and 10 (7.1%) in the EXAX group (Table 1). There were no differences between the groups for age, sex, BMI, smoking history, cardiovascular comorbidities, depression, reason for the MWT, working time, time in bed, and CPAP parameters except for CPAP mean duration of use, which was greater in the ENAN group than in the other subgroups (mean [SD] 6.25 [1.9] versus 4.2 [1.9] h, $p = 0.002$). Few (8.5%) of the patients were taking sedating medications, none in the groups with an AHI of >15 events/h but 19% in the EXAN group (post-hoc: ENAX versus EXAN, $p = 0.0181$). A non-significant trend was observed for history of car accident/near miss, which was higher in the EXAN (28.6%) and in the EXAX (30.0%) subgroups than in the other subgroups (ENAN, 15.8%; and ENAX, 15.7%). The pre-treatment ESS score was significantly lower in the ENAN and ENAX subgroups i.e., with a residual ESS score of <11. Adherence to CPAP treatment was higher in the ENAN and EXAN subgroups, i.e., with residual AHI of <15 events/h (respectively 98.7% and 90.5%) than in EXAN and EXAX subgroups (76.5% and 54.5%, respectively) (Table 1).

Regarding the questionnaires, there was no difference between groups for the Horne and Ostberg score (mean [SD] 16.2 [5]), but more intermediate profiles were observed in patients with an AHI of <15 events/h (Table 2). For depression evaluation, the mean (SD) BDI score was 5.8 (6.9); 6.4% of patients had abnormal values, and there were no differences among subgroups. The mean (SD) ESS score was 7.4 (4.5), with differences among subgroups due to their very definition. The mean (SD) Pichot scale score was 7.6 (7.5), with significantly higher values in individuals reporting greater sleepiness ($p < 0.01$) (Table 2).

Sleep parameters are presented in Table 3. The mean (SD) total time of sleep on the pre-MWT PSG night was 399.0 (75.4) min, the mean (SD) sleep latency was 19.9 (15.6) min, and the mean (SD) sleep efficiency was 81.0% (12.0%). There were no significant differences between subgroups. The mean (SD) N3 duration was 91.6 (34.1) min, and the mean (SD) REM sleep duration was 82.0 (32.6) min. The mean N3 duration was significantly higher and mean arousal index lower in the subgroups with an AHI of <15 events/h.

There was no difference between men ($N = 127$) and women ($N = 14$) for clinical and sleep characteristics, except for depression history (64% in women versus 21% in men, $p = 0.001$) (Table S1).

3.2 | Inter- and intra-scorer agreement

The MSL differed between the two scorers in 14.8% (84/564) of individual MWTs. This led to 32.6% of patients (46/141) having discordant MSL between the two scorers. The mean (SD) difference between MSL values was 5.4 (5.9) min (Table 4). Figure 1 shows the difference between the two scores versus the corresponding mean. For a MSL of 40 min, i.e., no sleep onset, the inter-scorer comparison showed a difference in 15 patients (11%) with a Cohen's κ of 0.74 (95% confidence interval [CI] 0.62–0.87). For a MSL of <33 min, the inter-scorer comparison showed a difference in 17 patients (12%) with a Cohen κ of 0.54 (95% CI 0.35–0.73). For a MSL of <19 min, the interscorer comparison showed a difference in five patients (4%) with a Cohen's κ of 0.27 (95% CI 0.18–0.71) (Table 4). Intra-scorer agreement, which was evaluated on half of the population ($n = 72$), was higher than inter-scorer agreement. Nonetheless, there were differences in patient categorisation ranging from one in 72 (1.4%) to six in 72 (8.3%) with Cohen's κ ranging from 0.49 to 0.94 depending on the scorer and the thresholds considered (Table 5 and Table S2).

3.3 | The MWT latencies as assessed by consensual rating

The mean (SD) cMSL in the whole population was 37.4 (5.4) min with a median (IQR) of 40 (36.8–40) min. Overall, 69.5% of the population did not have sleep onset, 16.3% had a cMSL between 33 and 40 min, 14.2% had a cMSL <33 min, and 2.1% <19 min. The minimum value was 14.5 min, which was thus above the 8-min threshold (Littner et al., 2005). The distribution of cMSL in the whole population is presented in the Figure S1.

The MWT results in the four groups are presented Table 6. In the ENAN subgroup, the mean (SD) cMSL was 38.4 (4.2) min, leading to a lower limit of mean – 2SD = 30 min, with a median (IQR) of 40 (40–40) min, and 80.3% of the population had no sleep onset. Categories of cMSL according to old (19 and 33 min) and new MWT thresholds (30 and 38.4 min) in this group of well-treated OSA without subjective sleepiness are presented in Figure 2. In the ENAX subgroup, the mean (SD) cMSL was 37.7 (4.8) min, with a median (IQR) of 40 (39.1–40) min and 70.6% of the population had no sleep onset. In the EXAN subgroup the mean (SD) cMSL was 34.6 (7.1) min, with median (IQR) of 40 (32.8–40) min. Only 38.1% of the population had no sleep onset. In the EXAX subgroup, the mean (SD) cMSL was 34.5 (8.4) min, with median (IQR) of 39.1 (31.5–40) min. Only 50% of the population had no sleep onset.

The cMSL differed significantly between the four groups ($p = 0.0015$): it was lower ($p < 0.001$) in patients with abnormal ESS scores (EXAN + EXAX) than in non-sleepy patients (ENAN + ENAX), but no significant difference was observed ($p = 0.5$) between groups with and without residual sleep apnea, i.e., ENAX + EXAX versus ENAN + EXAN. Interestingly, in the whole population, the prevalence of a history of car accident/near miss was significantly lower in the group with a MSL of ≥30 versus <30 min (16% versus 50%, $p < 0.01$).

No effect of gender was found for MSL and between group differences (for neither of the two scorers nor the cMSL, Tables S3 and S4).

4 | DISCUSSION

In this study, we demonstrated that: (i) despite high intra-scorer agreement, inter-scorer agreement was weak, leading to changes in

TABLE 1 Population characteristics.

	All	ENAN	ENAX	EXAN	EXAX	*p* (Kruskal–Wallis rank-sum test; Fisher's exact test)	Missing values
Number of patients (%)	141 (100)	76 (53.9)	34 (24.1)	21 (14.9)	10 (7.1)		
Age, years							
Mean (SD)	47.5 (9.1)	47.4 (9.8)	48.2 (9.3)	47.7 (7.8)	45.6 (6.8)	0.6	0
Median (IQR)	48.0 (42.0–54.0)	48.0 (42.5–54.0)	48.5 (42.0–54.0)	50.0 (42.0–52.0)	45.5 (39.8–51.5)		
Sex							
% female	9.9	11.8	5.9	14.3	0.0	0.5	0
BMI, kg/m^2							
Mean (SD)	31.6 (5.9)	31.9 (5.5)	32.6 (6.8)	29.3 (5.7)	30.9 (6.6)	0.3	0
Median (IQR)	31 (27.7–34.7)	31.1 (28.0–35.0)	31.4 (28.2–34.7)	30.4 (25.2–33)	29.2 (26–34.6)		
Smoking history, %	29.8	30.3	38.2	23.8	10.0	0.6	0
CV history, %							
HBP	24.8	23.7	32.4	23.8	10.0	0.5	0
Stroke	2.8	2.6	2.9	4.8	0.0	0.9	0
Myocardial infraction	0.7	0.0	2.9	0.0	0.0	0.5	0
Depression history, %	25.5	28.9	17.6	23.8	30.0	0.7	0
MWT context, %							1
Professional driver	75.7	81.3	70.6	81.0	40.0	0.07	
Security	15.7	10.7	23.5	4.8	50.0		
Prior to driving licence	1.4	1.3	2.9	0.0	0.0		
Other	7.1	6.7	2.9	14.3	10.0		
Working time, %						0.3	8
Regular hours	32.3	32.4	36.4	35.0	11.1		
Night shift	24.1	23.9	27.3	10.0	44.4		
Day shifts (<6:00 a.m. or >9:00 p.m.)	43.6	43.7	36.4	55.0	44.4		
TIB, h							
Workdays, mean (SD)	6.5 (1.2)	6.6 (0.7)	6.5 (0.04)	6.5 (0.1)	6.5 (1.3)	0.8	26
Median (IQR)	6.5 (6–7.0)	6.5 (6–7.2)	6.5 (6.0–7.0)	7.01 (6.0–7.0)	7.0. (6.0–7.2)		
Day off, mean (SD)	7.9 (1.3)	8.1 (0.7)	7.9 (0.1)	8.2 (0.1)	7.7 (1.5)	0.8	26
Median (IQR)	7.9 (7.0–9.1)	7.9 (7.0–9.1)	7.9 (7.4–8.2)	7.9 (7.0–9.6)	7.9 (7.0–8.4)		
Before treatment ESS, h							
Mean (SD)	10.4 (4.4)	9.8 (4.5)*	8.5 (4.0)###,Δ	13.4 (3.0)*,###	12.4 (3.3)Δ	**<0.001**	26
Median (IQR)	11 (8–13)	11 (7–13)	8 (6–10)	14 (11.5–15)	11.5 (11–13)		

(Continues)

TABLE 1 (Continued)

	All	ENAN	ENAX	EXAN	EXAX	*p* (Kruskal–Wallis rank-sum test; Fisher's exact test)	Missing values
Initial AHI, events/h							
Mean (SD)	43.8 (20.3)	43.5 (20.9)	45.0 (21.4)	41.9 (19.0)	43.7 (17.0)	0.9	2
Median (IQR)	39.0 (30.0–56.2)	38.0 (30.0–55.0)	38.1 (30.7–63.2)	41.0 (28.0–50.0)	41.0 (35.6–47.2)		
AHI >30 events/h, *n* (%)	109 (76.8)	59 (77.6)	26 (76.5)	15 (71.4)	9 (90.0)		
AHI ≥15 and <30 events/h, *n* (%)	29 (20.4)	15 (19.7)	8 (23.5)	5 (23.8)	1 (10.0)		
Car accident/near miss, %	18.4	15.8	14.7	28.6	30.0	0.2	0
Under-treatment PSG, %	88.7	98.7	76.5	90.5	50.0	**<0.001**	0
With CPAP	97.6	100.0	88.5	100.0	100.0		
With OA	2.4	0.0	11.5	0.0	0.0		
Sedating medication	8.5	10.6	0[#]	19.0[#]	0.0	0.04	0
CPAP treatment ($N = 122$ [86.5%])							
CPAP treatment use, h, mean (SD)	5.3 (2.3)	6.2 (2.0)[ΨΨΨ]	3.5 (2.0)[ΨΨΨ]	5.4 (1.5)	4.1 (2.1)	**0.002**	28
Median (IQR)	5.6 (4.3–6.7)	6 (5.1–7.0)	4.1 (2.8–6.2)	5.3 (4.0–6.0)	4.7 (3.3–6.13)		
CPAP leaks, L/min, mean (SD)	10.2 (9.6)	8.2 (6.5)	12.6 (14.3)	14.0 (9.4)	12.7 (13.1)	0.2	52
Median (IQR)	7.8 (2.8–14.7)	7.0 (2.7–12.0)	9.5 (2.3–16.5)	14.0 (5.0–22.0)	9.5 (2.7–19.5)		
CPAP autoset mode, %	38.0	48.7	20.6	28.6	40.0	0.1	2
CPAP facial mask, %	14.1	17.1	8.8	14.3	10.0	0.2	26
CPAP pressure (mean if constant, p95 if autoset)	10.3 (2.2)	10.8 (2.2)	10.3 (2.3)	10.7 (2.0)	6 (2.3)	0.06	25
Median (IQR)	10.0 (9.0–12.0)	10.0 (8.9–11.8)	10.0 (8.7–11.8)	10.6 (10.0–12.2)	8.0 (6.5–10.0)		

Note: post hoc comparisons (with Dunn's correction) for (1) pre-treatment ESS: ENAN versus EXAN, $p < 0.05^{*}$; ENAX versus EXAN, $p < 0.001^{\#\#\#}$; ENAX versus EXAX, $p < 0.05^{\Delta}$; (2) CPAP treatment use: ENAN versus ENAX, $p < 0.001^{\Psi\Psi\Psi}$; (3) sedating treatment use: ENAN versus ENAX, $p < 0.05^{\Psi}$. Bold values statistically significant at $p < 0.05$.

Abbreviations: AHI, Apnea–Hypopnea Index; BMI, body mass index; CPAP, continuous positive airway pressure; CV, cardiovascular; ESS, Epworth Sleepiness Scale; HBP, high blood pressure; IQR, interquartile range (25% and 75% quartiles); MWT, Maintenance of Wakefulness Test; OA, oral appliance therapy; PSG, polysomnography; SD, standard deviation; TIB, time in bed.

TABLE 2 Results of sleep and depression questionnaires.

Questionnaire	Overall $N = 141$	ENAN $n = 76$ (53.9%)	ENAX $n = 34$ (24.1%)	EXAN $n = 21$ (14.9%)	EXAX $n = 10$ (7.1%)	p	Missing values
Horne and Ostberg score							
Mean (SD)	16.2 (5)	16 (4.9)	16.4 (5.9)	16.7 (5.6)	16.6 (5)	0.8	4
Median (IQR)	17 (13–20)	17 (13–19)	18 (12–20)	17 (14–20)	18 (16–19)		
Horne and Ostberg category, *n/N* (%)						**0.04**	6
1		30/74 (41)	18/32 (56)	10/20 (50)	5/10 (50)		
2		37/74 (50)	7/32 (22)	9/20 (45)	2/10 (20)		
3		7/74 (9.5)	7/32 (22)	1/20 (5)	3/10 (30)		
Pichot scale score							
Mean (SD)	7.6 (7.5)	6.4 (7.3)***	5.6 (4.9)##	13.9 (8.8)***,##	9.5 (7.1)	**0.01**	2
Median (IQR)	5 (2–12)	3 (1–9)	5 (1–9)	15 (9–19)	8 (4.5–11)		
Abnormal score (threshold ≥22), %	7.8	7.9	0.0	19.0	10.0	0.06	2
BDI score							
Mean (SD)	5.8 (6.9)	5.9 (7.9)	4.8 (4.4)	6.6 (6.1)	6.7 (7.6)	0.7	2
Median (IQR)	4 (1–8)	3 (0–8)	4 (1–6.8)	6 (2–8.3)	4 (0.3–12.3)		
Abnormal BDI score (threshold ≥20), %	6.4	7.9	0.0	9.5	10.0	0.5	2
ESS							
Mean (SD)	7.4 (4.5)	5.4 (2.9)***,◆◆◆	5.6 (2.5)###,ΔΔΔ	14.6 (2.6)***,###	13.3 (2.1)ΔΔΔ,◆◆◆	**<0.01**	0
Median (IQR)	7 (4–10)	6 (3–8)	5.5 (3.3–7)	15 (12–16)	13 (12–14.5)		
Abnormal ESS score (threshold >10), %	77	0.0	0.0	100	100	**<0.01**	0
ODSI							
Mean (SD)	4.7 (5.3)	3.2 (4.2)**	4.3 (4.1)##	9 (6.8)**,##	8.8 (6.6)	**<0.01**	2
Median (IQR)	2 (1–8)	2 (0–4)	2 (1.3–7.5)	9 (4–14)	6 (3.5–14)		
Abnormal ODSI score (threshold >5), %	29.1	9.2	7.1	8.5	4.3	**<0.01**	2

Note: post hoc comparisons (with Dunn's correction) for (1) ESS: ENAN versus EXAN, $p < 0.001$***; ENAN versus EXAX, $p < 0.001$◆◆◆; ENAX versus EXAN, $p < 0.001$### and ENAX versus EXAX, $p < 0.001$ΔΔΔ; (2) Pichot scale: ENAN versus EXAN, $p < 0.001$*** and ENAX versus EXAN, $p < 0.01$##; (3) ODSI: ENAN versus EXAN, $p < 0.01$** and ENAX versus EXAN, $p < 0.01$##. Bold values statistically significant at $p < 0.05$.
Abbreviations: BDI, Beck Depression Inventory; ESS, Epworth Sleepiness Scale; ODSI; Observation and Interview-based Diurnal Sleepiness Inventory.ESS, Epworth Sleepiness Scale; BDI, Beck Depression Inventory, ODSI; Observation and Interview-based Diurnal Sleepiness Inventory.

MSL category (and thus potential medical decisions regarding driving authorisation) in 4%–12% of patients depending on the threshold considered; and (ii) the vast majority of subjectively non-sleepy and well-treated patients did not fall asleep and the mean (SD) cMSL in this group was 38.4 (4.2) min.

4.1 | The challenge of defining normative values for the MWT

The normative MWT latency values currently used in most French Sleep centres are 33 min (normal >33 min) and 19 min (abnormal <19 min), with unknown significance of MSL between 19 and 33 min. These values stem from a population of 64 healthy individuals, different standard of MWT and correspond to respectively MSL and lower normal limit (LNL) calculated as the MSL – 2SDs. The American literature tends to consider a LNL of 8 min but recommends no sleep onset in any of the four tests for maximum safety level jobs, with values between 8 and 40 min being of unknown significance (Krahn et al., 2021; Sullivan & Kushida, 2008). The absence of a recognised abnormal cut-off for the MSL obviously translates into heterogeneous clinical practices, underscoring the urgent need to define reliable thresholds. However, there are many challenges relative to the determination of normative values. First, the ceiling effect of the MWT results in a non-normal distribution (Doghramji et al., 1997; Littner et al., 2005). Second, both MWT protocols (40 versus 20 min) and sleep onset latency definitions ('three continuous epochs of stage 1' versus 'one epoch of whatever sleep stage' versus '10 s of sleep')

TABLE 3 Sleep variables (polysomnography).

Variable	Overall $N = 141$	ENAN $n = 76$ (53.9%)	ENAX $n = 34$ (24.1%)	EXAN $n = 21$ (14.9%)	EXAX $n = 10$ (7.1%)	*p*	Missing values
TST, min							
Mean (SD)	399.0 (75.4)	406.4 (72.2)	377.4 (88)	424 (48.4)	363.4 (80.4)	0.095	0
Median (IQR)	405 (358–452)	400.5 (360.8–455)	383 (349.3–425.3)	422 (405–450)	373 (308.3–432)		
Sleep efficiency, %							
Mean (SD)	81 (12)	83.2 (9.4)	75.7 (15.3)	83.3 (10.3)	78.4 (15.5)	0.38	0
Median (IQR)	83.2 (72.9–89.7)	86.2 (75.8–90.5)	78.7 (71.3–83.9)	85.4 (78.6–91.5)	83.7 (68.4–90.6)		
N3, min							
Mean (SD)	91.6 (34.1)	98.2 (31)$^{\Psi\Psi,\blacklozenge}$	74.3 (32.9)$^{\Psi\Psi,\#\#\#}$	105.7 (37.1)$^{\#\#\#}$	70.3 (23.8)$^{\blacklozenge}$	**<0.001**	0
Median (IQR)	91 (69–114)	101.5 (82.8–119)	70.5 (55.8–93)	109 (80–120)	71.5 (51.5–90.5)		
% of TST, mean (SD)	23.4 (8.9)	25 (9.1)	20.3 (8.7)	25 (8.1)	19.5 (6.2)		
REM sleep, min							
Mean (SD)	82 (32.6)	84.7 (31.7)	70.4 (29.9)$^{\#\#}$	100.2 (27.8)$^{\#\#,\circ}$	63.2 (37.1)$^{\circ}$	**0.004**	0
Median (IQR)	81 (58–104)	85 (61–111)	68.5 (51–88.8)	101 (80–126)	66.5 (31.5–91.8)		
% of TST, mean (SD)	20.3 (6.7)	20.7 (6.7)	18.4 (5.7)	23.5 (5.7)	16.4 (8.5)		
Arouxal index, events/h							
Mean (SD)	24 (16.6)	17.2 (7.9)$^{\Psi\Psi\Psi,\blacklozenge\blacklozenge\blacklozenge}$	35.2 (21)$^{\Psi\Psi\Psi,\#\#}$	20.2 (11.1)$^{\#\#,\circ\circ}$	45.5 (21.7)$^{\blacklozenge\blacklozenge\blacklozenge,\circ\circ}$	**<0.001**	0
Median (IQR)	19.9 (13.3–28.6)	16.5 (11.8–22.4)	28.3 (21.9–44.6)	18.5 (12.7–24.5)	38.5 (31.4–49.7)		
AHI, events/h							
Mean (SD)	14.9 (19.3)	5.2 (4)$^{\Psi\Psi\Psi,\blacklozenge\blacklozenge\blacklozenge}$	33.3 (21.7)$^{\Psi\Psi\Psi,\#\#\#}$	7.8 (8.8)$^{\#\#\#,\circ\circ}$	40.3 (29)$^{\blacklozenge\blacklozenge\blacklozenge,\circ\circ}$	**<0.001**	0
Median (IQR)	7.8 (3–17.3)	4.4 (2–7.7)	23.2 (18.2–38.7)	5.2 (3–10.1)	36.9 (19.8–40.5)		
Among which central apnea, events/h							
Mean (SD)	0.7 (1.3)	0.5 (0.8)	1.4 (1.7)	0.6 (1.3)	0.9 (1.7)		0
Median (IQR)	0.2 (0–1)	0 (0–0.6)	0.6 (0–2)	0 (0–0.5)	0.2 (0–0.9)		
AHI + RERA, events/h							
Mean (SD)	17.3 (19.9)	7.5 (5.4)$^{\Psi\Psi\Psi,\blacklozenge\blacklozenge\blacklozenge}$	36.6 (22.1)$^{\Psi\Psi\Psi,\#\#\#}$	9.4 (9.5)$^{\#\#\#,\circ\circ}$	43.2 (28.1)$^{\blacklozenge\blacklozenge\blacklozenge,\circ\circ}$	**<0.001**	0
Median (IQR)	11.1 (4.9–22)	6.7 (2.8–11.7)	25.2 (22.1–48.3)	6.2 (3.3–11.5)	37.1 (27.3–42.4)		
Sleep onset latency, min							
Mean (SD)	19.9 (15.6)	17.5 (10.1)	27.7 (23)$^{\Delta}$	19.4 (16.2)	12.1 (8.8)$^{\Delta}$	**0.014**	0

TABLE 3 (Continued)

Variable	Overall $N = 141$	ENAN $n = 76$ (53.9%)	ENAX $n = 34$ (24.1%)	EXAN $n = 21$ (14.9%)	EXAX $n = 10$ (7.1%)	p	Missing values
Median (IQR)	16 (10.5–25.5)	15 (9.2–25)	21.7 (13.9–28.6)	16.8 (8.5–19)	10.3 (6.6–12.6)		
Time SpO_2 <90%, s							
Mean (SD)	345.4 (1730.7)	358.7 (2183.8)$^{\blacklozenge\blacklozenge}$	539.9 (1308.2)	19.1 (43.8)$^{\circ}$	268.9 (420)$^{\blacklozenge\blacklozenge,\circ}$	**0.003**	0
Median (IQR)	0 (0–30)	0 (0–16)	2.5 (0–315.8)	0 (0–11)	107 (46.5–336.1)		

Note: post hoc comparisons (with Dunn's correction) for (1) N3: ENAN versus ENAX, $p < 0.01^{\Psi\Psi}$; ENAN versus EXAX, $p < 0.05^{\blacklozenge}$; ENAX versus EXAN, $p < 0.001^{\#\#\#}$; (2) REM: ENAX versus EXAN, $p < 0.01^{\#\#}$ and EXAN versus EXAX, $p < 0.05^{\circ}$; (3) Arousal index: ENAN versus ENAX, $p < 0.001^{\Psi\Psi\Psi}$; ENAN versus EXAX, $p < 0.001^{\blacklozenge\blacklozenge\blacklozenge}$; ENAX versus EXAN, $p < 0.01^{\#\#}$ and EXAN versus EXAX, $p < 0.01^{\circ\circ}$; (4) AHI: ENAN versus ENAX, $p < 0.001^{\Psi\Psi\Psi}$; ENAN versus EXAX, $p < 0.001^{\blacklozenge\blacklozenge}$; ENAX versus EXAN, $p < 0.001^{\#\#\#}$ and EXAN versus EXAX, $p < 0.01^{\circ\circ}$; (5) AHI + RERA: ENAN versus ENAX, $p < 0.001^{\Psi\Psi\Psi}$; ENAN versus EXAX, $p < 0.001^{\blacklozenge\blacklozenge}$; ENAX versus EXAN, $p < 0.001^{\#\#\#}$ and EXAN versus EXAX, $p < 0.01^{\circ\circ}$; (6) sleep onset latency: ENAX versus EXAX, $p < 0.05^{\Delta}$ (7) time spent with SpO_2 < 90%: ENAN versus EXAX, $p < 0.01^{\blacklozenge\blacklozenge}$; EXAN versus EXAX, $p < 0.05^{\circ}$. Bold values statistically significant at $p < 0.05$.

Abbreviations: AHI, apnea–hypopnea index; IQR, interquartile range (25% and 75% quartiles); REM, rapid eye movement; RERA, respiratory effort related arousal; SpO_2, oxygen saturation of the blood; TST, total sleep time.

have evolved over time, leading to poor reproducibility of results across studies (Doghramji et al., 1997; Krahn et al., 2021). Third, the population in which normative values should be determined is not clear. Theoretically, this population should include healthy subjects who do not complain of sleepiness, and in whom the main causes of sleepiness (sleep deprivation, depression, obesity, sedative medications, sleep disorders including OSA) have been carefully ruled out, which is rarely the case in studies (Berger et al., 2021; Jaussent et al., 2020; Li et al., 2022). Interestingly, another study in 31 healthy individuals found a greater mean (SD) value of 36.9 (5.4) min, and suggested that even the method of recruitment of subjects (advertisement versus random selection) could influence the MWT results (Banks et al., 2004b; Krahn et al., 2021). In practice, MWTs are usually performed to evaluate OSA treatment efficacy on sleepiness. Thus, several studies have evaluated the MWT in patients with OSA, with heterogeneous results according to OSA severity and treatment status. In 110 patients with mild-to-moderate untreated OSA, the mean (SD) MWT MSL was 30.7 (10.2) min (Banks et al., 2004a). Another study reported a mean (SD) MSL of 25.9 (11.8) min in 322 untreated patients with OSA, with an increase from 18 (12) min to 31.9 (10.4) min in the 24 patients treated with CPAP (Poceta et al., 1992). In our work, with a pragmatic approach, we evaluated the MSL in a large population of treated patients with OSA. This reflects clinical practice, as, according to the French legislation, MWTs are required in OSA associated with excessive daytime sleepiness following at least 1 month of treatment before resuming professional driving of safety occupations. We focused on a clinical population with demonstrated treatment efficacy on the AHI as assessed by PSG and without residual subjective sleepiness. In this group (ENAN), the mean (SD) cMSL was 38.4 (4.2) min, leading to a LNL of 30 min, which is higher than the threshold (based on normal individuals) used in many centres.

4.2 | Issues associated with subjective scoring

Interpretation of the MWT relies on visual human scoring, resulting in an inter-rater variability in sleep stage assignment. The reported inter-rater agreement in sleep stage is ~80% in healthy individuals, with lower values in patients with sleep disorders (Norman et al., 2000; Younes et al., 2016). It depends on sleep stage and is lower for N1, which might translate to a high variability in sleep onset latency between scorers. Our study is the first to provide valuable data through the comparison of scores from two sleep experts from two different sleep centres. Inter- and intra-scorer Cohen's κ based on the usual thresholds of 40, 33 and 19 min suggested fair to substantial inter-scorer agreement and moderate to almost perfect intra-scorer agreement. Although there were differences in scoring context, our results for inter-scorer agreement are similar to those published in a recent meta-analysis (Lee et al., 2022). However, scoring discrepancies have much more impact in the context of the MWT than in the context of assessment of sleep parameters. Using a 33-min threshold, different classifications were attributed to 12% of individuals by the two sleep centres, meaning that 17 patients would have obtained

TABLE 4 Results of Maintenance of Wakefulness Test scoring.

Variable	All $N = 141$	ENAN $n = 76$ (53.9%)	ENAX $n = 34$ (24.1%)	EXAN $n = 21$ (14.9%)	EXAX $n = 10$ (7.1%)	p (Kruskal–Wallis/ Wilcoxon rank sum)
S1, min, mean (SD)	37.6 (5.2)	38.7 (3.6)	38.2 (4)	34.5 (7.5)	34 (8.9)	**<0.01**
S1, min, median (IQR)	40 (38–40)	40 (40–40)	40 (39.1–40)	37.3 (31.3–40)	39.1 (31.9–40)	
			G2/3/4 36.4 (6.4)			**0.003**
S1-Patients with 1 sleep onset, %	29.1	18.4	29.4	57.1	50.0	
S2, min, mean (SD)	37.1 (5.7)	37.8 (5.4)	37.1 (6.2)	34.8 (6.7)	36.8 (4)	**0.03**
S2, min, median (IQR)	40 (36.5–40)	40 (40–40)	40 (39.4–40)	36.4 (32.1–40)	39.1 (34.1–40)	
			G2/3/4 36.3 (6.1)			**0.025**
S2-Patients with 1 sleep onset, %	29.8	21.1	29.4	52.4	50.0	
Scoring differences in sleep onset latency						
Number of differences considering each MWT trial, (%)	84/564 (15)					
Number of differences considering MSL (%)	46/141 (32.6)	20 (26.3)	10 (29.4)	12 (57.1)	4 (40.0)	
Absolute value of MSL difference of scoring, min, mean (SD)	5.4 (5.9)	5.6 (6.3)	5.2 (7.4)	4.5 (2.8)	7.7 (8.5)	
Number of differences, threshold 40 min, *n* patients (%)	15 (10.6)	10 (13.2)	2 (5.9)	3 (14.3)	0 (0.0)	
Cohen's κ, (95% CI)	0.74 (0.62–0.87)					
Number of differences, threshold <33 min, *n* patients (%)	17 (12.1)	9 (11.8)	3 (8.8)	4 (19.0)	1 (10.0)	
Cohen's κ (95% CI)	0.54 (0.35–0.73)					
Number of differences, threshold <19 min, *n* patients (%)	**5 (3.5)**	2 (2.6)	1 (2.9)	1 (4.8)	1 (10.0)	
Cohen's κ (95% CI)	0.27 (−0.18–0.71)					

Note: Bold values statistically significant at $p < 0.05$.

Abbreviations: CI, confidence interval; IQR, interquartile range (25% and 75% quartiles); MSL, (mean) sleep latency; MWT, Maintenance of Wakefulness Test; S1, scorer 1; S2, scorer 2; SD, standard deviation.

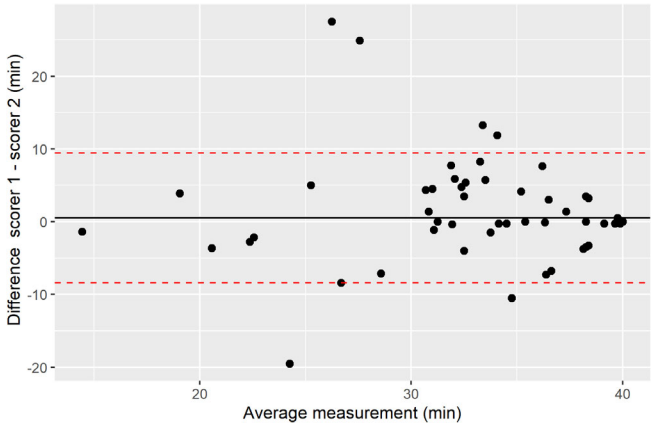

FIGURE 1 Bland–Altman plot of inter-scorer difference. The mean difference (0.52) is represented as a black line and the limits of agreement (Lower, −8.40; Upper, +9.43) are represented as red dotted lines.

differing driving authorisation results depending on the sleep centre they were referred to, even though the scorers were experienced. This discrepancy raises the question of whether an automatic sleep scoring approach would be a more appropriate way to ensure inter-centre reproducibility (Peter-Derex et al., 2021).

4.3 | Application to the prediction of accidental risk in real life?

The aim of the MWT in a clinical context is to predict the risk of accidents. However, despite its wide use around the world, the MWT protocol remains criticised for its artificial nature (Wise, 2006). Therefore, several studies have attempted to investigate the relationship between MWT findings, simulator driving performance, real driving performance, and/or accident risk. The demonstration of impaired driving performance in patients with OSA with MSL below thresholds

TABLE 5 Intra- and inter-scorer agreement.

	Intra-scorer (Scorer 1)	Intra-scorer (Scorer 2)	Inter-scorer on same subset	Inter-scorer on overall population
0 sleep onset				
% of difference	2.8	5.6	9.9	10.6
Cohen's κ (95% CI)	0.94 (0.86 to 1)	0.88 (0.76 to 0.99)	0.79 (0.65 to 0.94)	0.74 (0.62 to 0.87)
Threshold <33 min				
% of difference	8.5	7.0	12.7	12.1
Cohen's κ (95% CI)	0.75 (0.56 to 0.94)	0.77 (0.58 to 0.96)	0.61 (0.38 to 0.84)	0.54 (0.35 to 0.73)
Threshold <30 min				
% of difference	2.8	7.0	12.7	9.9
Cohen's κ (95% CI)	0.87 (0.7 to 1)	0.67 (0.39 to 0.94)	0.4 (0.08 to 0.72)	0.41 (0.16 to 0.66)
Threshold <19 min				
% of difference	1.4	2.8	4.2	3.5
Cohen's κ (95% CI)	0.79 (0.4–1)	0.49 (−0.13 to 1)	0.38 (−0.17 to 0.93)	0.27 (−0.18 to 0.71)

TABLE 6 MWT results on consensual rating.

	All $N = 141$	ENAN $n = 76$ (53.9%)	ENAX $n = 34$ (24.1%)	EXAN $n = 21$ (14.9%)	EXAX $n = 10$ (7.1%)	p (Kruskal–Wallis/Wilcoxon rank sum)
cMSL, min Mean (SD)	37.4 (5.4)	38.4 (4.2)	37.7 (4.8)	34.6 (7.1)	34.5 (8.4)	**<0.01**
			G2/3/4 36.2 (6.3)			**0.003**
Median (IQR)	40 (36.8–40)	40 (40–40)	40 (39.1–40)	36.4 (32.8–40)	39.1 (31.5–40)	
Patients with 1 sleep onset, %	30.5	19.7	29.4	61.9	50.0	
cMSL <38.4 min, %	29.1	19.7	23.5	61.9	50.0	
cMSL <33 min, %	14.2	6.6	17.6	28.6	30.0	
cMSL <30 min, %	8.5	3.9	11.8	14.3	20.0	
cMSL <19 min, %	2.1	1.3	0.0	4.8	10.0	
cMSL <8 min, %	0.0	0.0	0.0	0.0	0.0	

Note: Bold values statistically significant at $p < 0.05$.

Abbreviations: cMSL, consensual mean sleep latency; IQR, interquartile range (25% and 75% quartiles); SD, standard deviation.

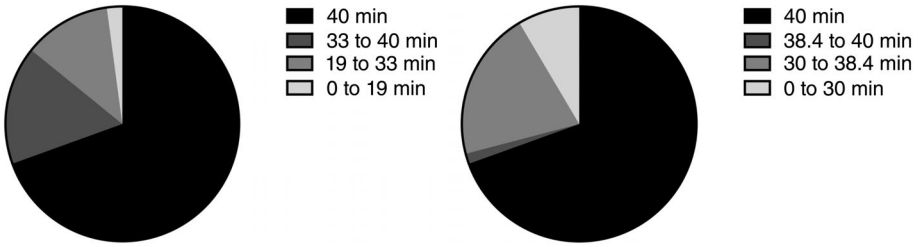

FIGURE 2 Distribution of consensual mean sleep latencies in the whole population with usual thresholds (left) and new suggested thresholds based on our study for treated severe obstructive sleep apnea (right).

of 19 and 33 min was provided on 38 and 30 untreated patients with OSA on a driving simulator (Philip et al., 2008; Sagaspe et al., 2007), and then confirmed by the same Bordeaux team in other sleep disorders treated or not (Philip et al., 2021). It was also shown that the MWT was better correlated with performance on driving simulator than the Multiple Sleep Latency Test (MSLT) (Pizza et al., 2009) and that driving simulator performance had a good correlation with real driving performance although not perfect (Philip et al., 2005). However, several questions remain unsolved, such as the prevalence, normative latencies, and meaning of microsleep episodes often recorded during the MWT and the dynamics of sleepiness over time (and thus the recommended periodicity of sleepiness assessment) (Boyle et al., 2008; Des Champs de Boishebert et al., 2021; Hertig-Godeschalk et al., 2020; Morrone et al., 2020). Interestingly, we observed that an ESS score of ≥11 but not an AHI of ≥15 events/h was associated with lower MSL, which is in line with previous finding about association subjective sleepiness, impaired driving performance, and history of car crash (Budhiraja et al., 2017; Pizza et al., 2009).

4.4 | Strengths and limitations

The strengths of our study are the large size of our population as compared to most studies in the field (Banks et al., 2004a, 2004b; Doghramji et al., 1997; Mitler et al., 1982; Poceta et al., 1992), which was representative of patients with OSA usually explored with the MWT, the availability of subjective scales performed at the time of the MWT, and a full-night PSG the night before the MWT, which allowed us to assess residual AHI and sleep quality. The use of multiple evaluations of each MWT trial for each patient, first independent and then consensual, by two sleep experts is a novel approach. We also acknowledge several limitations. First, there was no longitudinal follow-up of MWT results before and after CPAP or OA initiation, so we were unable to observe the influence of OSA treatment on objective sleepiness. Second, only two experienced scorers participated in the study. However we believe that the assessment of both intra- and inter-scorer agreement provides a good estimate of scoring variability. Future work including more scorers with various level of expertise would be interesting. Third, our work focused on the population of patients with moderate-to-severe OSA in the context of professional ability evaluation; the age, high BMI and gender imbalance reflects the reality of clinical practice. To note, sub-analyses accounting for gender, although of limited power, did not point toward a gender effect. A high residual AHI on PSG in a patient supposed to be well treated at the time of the MWT may reflect the discrepancy between device automatic detection and PSG visual scoring of residual respiratory events (Fanfulla et al., 2021; Georges et al., 2015). In addition, no control group (without OSA) was available, although this would have provided valuable information about 'normality' of the MWT results in well-treated non-sleepy patients with OSA. However, our findings suggest that one should expect in well-treated non-sleepy (and probably motivated given professional issues) patients with OSA higher MWT latencies than those commonly used in sleep centres, based on healthy individuals.

In conclusion, our study, which provides the first normative values and scorer variability for the MWT obtained with double multicentric scoring in treated mostly severe OSA, suggests the need for higher values (LNL = 30 min) than those currently used. Further longitudinal studies are warranted to evaluate the real-life risk of accidents associated with such a threshold.

AUTHOR CONTRIBUTIONS

Laure Peter-Derex, Jacques Taillard, Mélanie Strauss, and Christian Berthomier were involved in the conception and design of the study. Laure Peter-Derex was the coordinator of the study. Pierre Tankéré, Thierry Petitjean, Marc-Antoine Armeni, and Laure Peter-Derex were responsible for the data collection. Pierre Tankéré and Laure Peter-Derex were in charge of the analysis and wrote the first draft. All authors were involved in the interpretation, critically reviewed the first draft, and approved the final version.

ACKNOWLEDGMENTS

We thank Suzanne Rankin for reviewing the English.

FUNDING INFORMATION

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CONFLICT OF INTEREST STATEMENT

All authors declare no competing interests related to this work.

DATA AVAILABILITY STATEMENT

Data and material are available upon request at University Hospital, CHU de Lyon – Hôpital Lyon Croix Rousse, France.

ORCID

Christian Berthomier https://orcid.org/0000-0002-2300-9476

Mélanie Strauss https://orcid.org/0000-0002-7283-7374

Laure Peter-Derex https://orcid.org/0000-0002-9938-9639

REFERENCES

Banks, S., Barnes, M., Tarquinio, N., Pierce, R. J., Lack, L. C., & McEvoy, R. D. (2004a). Factors associated with maintenance of wakefulness test mean sleep latency in patients with mild to moderate obstructive sleep apnoea and normal subjects. *Journal of Sleep Research*, *13*(1), 71–78. https://doi.org/10.1111/j.1365-2869.2003.00383.x

Banks, S., Barnes, M., Tarquinio, N., Pierce, R. J., Lack, L. C., & McEvoy, R. D. (2004b). The maintenance of wakefulness test in normal healthy subjects. *Sleep*, *27*(4), 799–802.

Banks, S., Catcheside, P., Lack, L. C., Grunstein, R. R., & McEvoy, R. D. (2005). The maintenance of wakefulness test and driving simulator performance. *Sleep*, *28*(11), 1381–1385.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry*, *4*, 561–571.

Berger, M., Hirotsu, C., Haba-Rubio, J., Betta, M., Bernardi, G., Siclari, F., Waeber, G., Vollenweider, P., Marques-Vidal, P., & Heinzer, R. (2021). Risk factors of excessive daytime sleepiness in a prospective population-based cohort. *Journal of Sleep Research*, *30*(2), e13069. https://doi.org/10.1111/jsr.13069

Berry, R. B., Brooks, R., Gamaldo, C., Harding, S. M., Lloyd, R. M., Quan, S. F., Troester, M. T., & Vaughn, B. V. (2017). AASM scoring manual updates for 2017 (version 2.4). *Journal of Clinical Sleep Medicine: JCSM: Official Publication of the American Academy of Sleep Medicine*, *13*(5), 665–666. https://doi.org/10.5664/jcsm.6576

Bijlenga, D., Urbanus, B., van der Sluiszen, N. N. J. J. M., Overeem, S., Ramaekers, J. G., Vermeeren, A., & Lammers, G. J. (2022). Comparing objective wakefulness and vigilance tests to on-the-road driving performance in narcolepsy and idiopathic hypersomnia. *Journal of Sleep Research*, *31*(3), e13518. https://doi.org/10.1111/jsr.13518

Bioulac, S., Micoulaud-Franchi, J.-A., Arnaud, M., Sagaspe, P., Moore, N., Salvo, F., & Philip, P. (2017). Risk of motor vehicle accidents related to sleepiness at the wheel: A systematic review and meta-analysis. *Sleep*, *40*(10), zsx134. https://doi.org/10.1093/sleep/zsx134

Boyle, L. N., Tippin, J., Paul, A., & Rizzo, M. (2008). Driver performance in the moments surrounding a microsleep. *Transportation Research. Part F, Traffic Psychology and Behaviour*, *11*(2), 126–136. https://doi.org/10.1016/j.trf.2007.08.001

Budhiraja, R., Kushida, C. A., Nichols, D. A., Walsh, J. K., Simon, R. D., Gottlieb, D. J., & Quan, S. F. (2017). Predictors of sleepiness in obstructive sleep apnea at baseline and after 6 months of continuous positive airway pressure therapy. *The European Respiratory Journal*, *50*(5), 1700348. https://doi.org/10.1183/13993003.00348-2017

Des Champs de Boishebert, L., Pradat, P., Bastuji, H., Ricordeau, F., Gormand, F., Le Cam, P., Stauffer, E., Petitjean, T., & Peter-Derex, L. (2021). Microsleep versus sleep onset latency during maintenance wakefulness tests: Which one is the best marker of sleepiness? *Clocks Sleep*, *3*(2), 259–273. https://doi.org/10.3390/clockssleep3020016

Dinges, D. F. (1995). An overview of sleepiness and accidents. *Journal of Sleep Research*, *4*(S2), 4–14.

Doghramji, K., Mitler, M. M., Sangal, R. B., Shapiro, C., Taylor, S., Walsleben, J., Belisle, C., Erman, M. K., Hayduk, R., Hosn, R., O'Malley, E. B., Sangal, J. M., Schutte, S. L., & Youakim, J. M. (1997). A normative study of the maintenance of wakefulness test (MWT). *Electroencephalography and Clinical Neurophysiology*, *103*(5), 554–562.

Fanfulla, F., D'Artavilla Lupo, N., Malovini, A., Arcovio, S., Prpa, A., Mogavero, M. P., Pronzato, C., & Bonsignore, M. R. (2021). Reliability of automatic detection of AHI during positive airway pressure treatment in obstructive sleep apnea patients: A "real-life study". *Respiratory Medicine*, *177*, 106303. https://doi.org/10.1016/j.rmed.2021.106303

Garbarino, S., Guglielmi, O., Sanna, A., Mancardi, G. L., & Magnavita, N. (2016). Risk of occupational accidents in workers with obstructive sleep apnea: Systematic review and meta-analysis. *Sleep*, *39*(6), 1211–1218. https://doi.org/10.5665/sleep.5834

Georges, M., Adler, D., Contal, O., Espa, F., Perrig, S., Pépin, J. L., & Janssens, J. P. (2015). Reliability of apnea-hypopnea index measured by a home Bi-level pressure support ventilator versus a polysomnographic assessment. *Respiratory Care*, *60*(7), 1051–1056. https://doi.org/10.4187/respcare.03633

Hertig-Godeschalk, A., Skorucak, J., Malafeev, A., Achermann, P., Mathis, J., & Schreier, D. R. (2020). Microsleep episodes in the borderland between wakefulness and sleep. *Sleep*, *43*(1), zsz163. https://doi.org/10.1093/sleep/zsz163

Horne, J. A., & Ostberg, O. (1976). A self-assessment questionnaire to determine morningness-eveningness in human circadian rhythms. *International Journal of Chronobiology*, *4*(2), 97–110.

Jaussent, I., Morin, C. M., Ivers, H., & Dauvilliers, Y. (2020). Natural history of excessive daytime sleepiness: A population-based 5-year longitudinal study. *Sleep*, *43*(3), zsz249. https://doi.org/10.1093/sleep/zsz249

Johns, M. W. (1991). A new method for measuring daytime sleepiness: The Epworth sleepiness scale. *Sleep*, *14*(6), 540–545.

Kapur, V. K., Auckley, D. H., Chowdhuri, S., Kuhlmann, D. C., Mehra, R., Ramar, K., & Harrod, C. G. (2017). Clinical practice guideline for diagnostic testing for adult obstructive sleep apnea: An American Academy of sleep medicine clinical practice guideline. *Journal of Clinical Sleep Medicine: JCSM: Official Publication of the American Academy of Sleep Medicine*, *13*(3), 479–504. https://doi.org/10.5664/jcsm.6506

Krahn, L. E., Arand, D. L., Avidan, A. Y., Davila, D. G., DeBassio, W. A., Ruoff, C. M., & Harrod, C. G. (2021). Recommended protocols for the multiple sleep latency test and maintenance of wakefulness test in adults: Guidance from the American Academy of sleep medicine. *Journal of Clinical Sleep Medicine: JCSM: Official Publication of the American Academy of Sleep Medicine*, *17*(12), 2489–2498. https://doi.org/10.5664/jcsm.9620

Lee, Y. J., Lee, J. Y., Cho, J. H., & Choi, J. H. (2022). Interrater reliability of sleep stage scoring: A meta-analysis. *Journal of Clinical Sleep Medicine: JCSM: Official Publication of the American Academy of Sleep Medicine*, *18*(1), 193–202. https://doi.org/10.5664/jcsm.9538

Li, Z., Cai, S., Wang, J., & Chen, R. (2022). Predictors of the efficacy for daytime sleepiness in patients with obstructive sleep apnea with continual positive airway pressure therapy: A meta-analysis of randomized controlled trials. *Frontiers in Neurology*, *13*, 911996. https://doi.org/10.3389/fneur.2022.911996

Littner, M. R., Kushida, C., Wise, M., Davila, D. G., Morgenthaler, T., Lee-Chiong, T., Hirshkowitz, M., Daniel, L. L., Bailey, D., Berry, R. B., Kapen, S., & Kramer, M. (2005). Practice parameters for clinical use of the multiple sleep latency test and the maintenance of wakefulness test. *Sleep*, *28*(1), 113–121.

Mitler, M. M., Gujavarty, K. S., & Browman, C. P. (1982). Maintenance of wakefulness test: A polysomnographic technique for evaluation treatment efficacy in patients with excessive somnolence. *Electroencephalography and Clinical Neurophysiology*, *53*(6), 658–661.

Morrone, E., D'Artavilla Lupo, N., Trentin, R., Pizza, F., Risi, I., Arcovio, S., & Fanfulla, F. (2020). Microsleep as a marker of sleepiness in obstructive sleep apnea patients. *Journal of Sleep Research*, *29*(2), e12882. https://doi.org/10.1111/jsr.12882

Mulgrew, A. T., Ryan, C. F., Fleetham, J. A., Cheema, R., Fox, N., Koehoorn, M., Fitzgerald, J. M., Marra, C., & Ayas, N. T. (2007). The

impact of obstructive sleep apnea and daytime sleepiness on work limitation. *Sleep Medicine*, *9*(1), 42–53. https://doi.org/10.1016/j.sleep.2007.01.009

Norman, R. G., Pal, I., Stewart, C., Walsleben, J. A., & Rapoport, D. M. (2000). Interobserver agreement among sleep scorers from different centers in a large dataset. *Sleep*, *23*(7), 901–908.

Ohayon, M. M. (2008). From wakefulness to excessive sleepiness: What we know and still need to know. *Sleep Medicine Reviews*, *12*(2), 129–141. https://doi.org/10.1016/j.smrv.2008.01.001

Onen, F., Lalanne, C., Pak, V. M., Gooneratne, N., Falissard, B., & Onen, S.-H. (2016). A three-item instrument for measuring daytime sleepiness: The observation and interview based diurnal sleepiness inventory (ODSI). *Journal of Clinical Sleep Medicine: JCSM: Official Publication of the American Academy of Sleep Medicine*, *12*(4), 505–512. https://doi.org/10.5664/jcsm.5676

Pépin, J.-L., Georgiev, O., Tiholov, R., Attali, V., Verbraecken, J., Buyse, B., Partinen, M., Fietze, I., Belev, G., Dokic, D., Tamisier, R., Lévy, P., Lecomte, I., Lecomte, J.-M., Schwartz, J.-C., Dauvilliers, Y., & HAROSA I Study Group. (2021). Pitolisant for residual excessive daytime sleepiness in OSA patients adhering to CPAP: A randomized trial. *Chest*, *159*(4), 1598–1609. https://doi.org/10.1016/j.chest.2020.09.281

Peppard, P. E., Young, T., Barnet, J. H., Palta, M., Hagen, E. W., & Hla, K. M. (2013). Increased prevalence of sleep-disordered breathing in adults. *American Journal of Epidemiology*, *177*(9), 1006–1014. https://doi.org/10.1093/aje/kws342

Peter-Derex, L., Berthomier, C., Taillard, J., Berthomier, P., Bouet, R., Mattout, J., Brandewinder, M., & Bastuji, H. (2021). Automatic analysis of single-channel sleep EEG in a large spectrum of sleep disorders. *Journal of Clinical Sleep Medicine: JCSM: Official Publication of the American Academy of Sleep Medicine*, *17*(3), 393–402. https://doi.org/10.5664/jcsm.8864

Philip, P., Guichard, K., Strauss, M., Léger, D., Pepin, E., Arnulf, I., Sagaspe, P., Barateau, L., Lopez, R., Taillard, J., Micoulaud-Franchi, J.-A., & Dauvilliers, Y. (2021). Maintenance of wakefulness test: How does it predict accident risk in patients with sleep disorders? *Sleep Medicine*, *77*, 249–255. https://doi.org/10.1016/j.sleep.2020.04.007

Philip, P., Sagaspe, P., Taillard, J., Chaumet, G., Bayon, V., Coste, O., Bioulac, B., & Guilleminault, C. (2008). Maintenance of wakefulness test, obstructive sleep apnea syndrome, and driving risk. *Annals of Neurology*, *64*(4), 410–416. https://doi.org/10.1002/ana.21448

Philip, P., Sagaspe, P., Taillard, J., Valtat, C., Moore, N., Åkerstedt, T., Charles, A., & Bioulac, B. (2005). Fatigue, sleepiness, and performance in simulated versus real driving conditions. *Sleep*, *28*(12), 1511–1516. https://doi.org/10.1093/sleep/28.12.1511

Pichot, P., & Brun, J. P. (1984). Brief self-evaluation questionnaire for depressive, asthenic and anxious dimensions. *Annales Medico-Psychologiques*, *142*(6), 862–865.

Pizza, F., Contardi, S., Mondini, S., Trentin, L., & Cirignotta, F. (2009). Daytime sleepiness and driving performance in patients with obstructive sleep apnea: Comparison of the MSLT, the MWT, and a simulated driving task. *Sleep*, *32*(3), 382–391.

Poceta, J. S., Timms, R. M., Jeong, D.-U., Ho, S.-L., Erman, M. K., & Mitler, M. M. (1992). Maintenance of wakefulness test in obstructive sleep apnea syndrome. *Chest*, *101*(4), 893–897. https://doi.org/10.1378/chest.101.4.893

Sagaspe, P., Taillard, J., Chaumet, G., Guilleminault, C., Coste, O., Moore, N., Bioulac, B., & Philip, P. (2007). Maintenance of wakefulness test as a predictor of driving performance in patients with untreated obstructive sleep apnea. *Sleep*, *30*(3), 327–330.

Schreier, D. R., Banks, C., & Mathis, J. (2018). Driving simulators in the clinical assessment of fitness to drive in sleepy individuals: A systematic review. *Sleep Medicine Reviews*, *38*, 86–100. https://doi.org/10.1016/j.smrv.2017.04.004

Strollo, P. J., Hedner, J., Collop, N., Lorch, D. G., Chen, D., Carter, L. P., Lu, Y., Lee, L., Black, J., Pépin, J.-L., Redline, S., & Tones 4 Study Investigators. (2019). Solriamfetol for the treatment of excessive sleepiness in OSA: A placebo-controlled randomized withdrawal study. *Chest*, *155*(2), 364–374. https://doi.org/10.1016/j.chest.2018.11.005

Sullivan, S. S., & Kushida, C. A. (2008). Multiple sleep latency test and maintenance of wakefulness test. *Chest*, *134*(4), 854–861. https://doi.org/10.1378/chest.08-0822

Tregear, S., Reston, J., Schoelles, K., & Phillips, B. (2010). Continuous positive airway pressure reduces risk of motor vehicle crash among drivers with obstructive sleep apnea: Systematic review and meta-analysis. *Sleep*, *33*(10), 1373–1380. https://doi.org/10.1093/sleep/33.10.1373

Ward, K., Hillman, D. R., James, A., Bremner, A., Simpson, L., Cooper, M. N., Palmer, L. J., Fedson, A. C., & Mukherjee, S. (2013). Excessive daytime sleepiness increases the risk of motor vehicle crash in obstructive sleep apnea. *Journal of Clincal Sleep Medicine*, *9*(10), 1013–1021. https://doi.org/10.5664/jcsm.3072

Wise, M. S. (2006). Objective measures of sleepiness and wakefulness: Application to the real world? *Journal of Clinical Neurophysiology: Official Publication of the American Electroencephalographic Society*, *23*(1), 39–49. https://doi.org/10.1097/01.wnp.0000190416.62482.42

Younes, M., Raneri, J., & Hanly, P. (2016). Staging sleep in Polysomnograms: Analysis of inter-scorer variability. *Journal of Clinical Sleep Medicine: JCSM: Official Publication of the American Academy of Sleep Medicine*, *12*(6), 885–894. https://doi.org/10.5664/jcsm.5894

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Tankéré, P., Taillard, J., Armeni, M.-A., Petitjean, T., Berthomier, C., Strauss, M., & Peter-Derex, L. (2023). Revisiting the maintenance of wakefulness test: from intra-/inter-scorer agreement to normative values in patients treated for obstructive sleep apnea. *Journal of Sleep Research*, e13961. https://doi.org/10.1111/jsr.13961