

Observer Variation in Drug-Induced Sleep Endoscopy: Experienced Versus Nonexperienced Ear, Nose, and Throat Surgeons

Anneclaire V. M. T. Vroegop, MD^{1,2}; Olivier M. Vanderveken, MD, PhD^{1,2}; Kristien Wouters, PhD³; Evert Hamans, MD, PhD^{1,2}; Marijke Dieltjens, MBS^{1,2}; Nele R. Michels, MD, PhD²; Winfried Hohenhorst, MD⁴; Eric J. Kezirian, MD, MPH⁵; Bhik T. Kotecha, MPhil., FRCS⁶; Nico de Vries, MD, PhD⁷; Marc J. Braem, DDS, PhD^{1,2}; Paul H. Van de Heyning, MD, PhD^{1,2}

¹ENT, Head and Neck Surgery, Antwerp University Hospital, Edegem, Antwerp, Belgium; ²Faculty of Medicine and Health Sciences, University of Antwerp, Antwerp, Belgium; ³Department of Scientific Coordination and Biostatistics, Antwerp University Hospital, Edegem, Antwerp, Belgium; ⁴ENT, Head and Neck Surgery, Alfried Krupp Hospital, Essen, Germany; ⁵Otolaryngology—Head and Neck Surgery, University of California, San Francisco, San Francisco, CA; ⁶ENT, Head and Neck Surgery, The Royal National Throat, Nose and Ear Hospital, London, United Kingdom; ⁷ENT, Head and Neck Surgery, St. Lucas Andreas Hospital, Amsterdam, Netherlands

Study Objective: To determine variations in interobserver and intraobserver agreement of drug-induced sleep endoscopy (DISE) in a cohort of experienced versus nonexperienced ear, nose, and throat (ENT) surgeons.

Design: Prospective, blinded agreement study.

Setting: Ninety-seven ENT surgeons (90 nonexperienced with DISE; seven experienced) observed six different DISE videos and were asked to score the upper airway (UA) level (palate, oropharynx, tongue base, hypopharynx, epiglottis), direction (anteroposterior, concentric, lateral), and degree of collapse (none; partial or complete collapse). Findings were collected and analyzed, determining interobserver and intraobserver agreement [overall agreement (OA), specific agreement (SA)] and kappa values per UA level.

Measurement and Results: In the nonexperienced group, overall interobserver agreement on presence of tongue base collapse (OA = 0.63; kappa = 0.33) was followed by the agreement on epiglottis (OA = 0.57; kappa = 0.23) and oropharynx collapse (OA = 0.45; kappa = 0.09). Low overall interobserver agreement in this group was found for hypopharyngeal collapse (OA = 0.33; kappa = 0.08). A similar ranking was found for degree of collapse. For direction of collapse, high interobserver agreement was found for the palate (OA = 0.57; kappa = 0.16). Among the experienced observers, overall interobserver agreement was highest for presence of tongue base collapse (OA = 0.93; kappa = 0.71), followed by collapse of the palate (OA = 0.80; kappa = 0.51). In this group, lowest agreement was also found for hypopharyngeal collapse (OA = 0.47; kappa = 0.03). Interobserver agreement on direction of collapse was highest for epiglottis collapse (OA = 0.97; kappa = 0.97). Concerning the degree of collapse, highest agreement was found for degree of oropharyngeal collapse (OA = 0.82; kappa = 0.66). Among the experienced observers a statistically significant higher interobserver agreement was obtained for presence, direction, and degree of oropharyngeal collapse, as well as for presence of tongue base collapse and degree of epiglottis collapse. Among the nonexperienced observers, high intraobserver agreement was found in particular for tongue base and epiglottis collapse. Among the experienced observers, high agreement was found for all levels but to a lesser extent for hypopharyngeal collapse. Intraobserver agreement was statistically significantly higher in the experienced group, for all UA levels expect for the hypopharynx.

Conclusion: This study indicates that both interobserver and intraobserver agreement was higher in experienced versus nonexperienced ENT surgeons. Agreement ranged from poor to excellent in both groups. The current results suggest that experience in performing DISE is necessary to obtain reliable observations.

Keywords: Drug-induced sleep endoscopy, interobserver agreement, intraobserver agreement, obstructive sleep apnea, sleep disordered breathing, sleep related breathing disorders

Citation: Vroegop AVMT; Vanderveken OM; Wouters K; Hamans E; Dieltjens M; Michels NR; Hohenhorst W; Kezirian EJ; Kotecha BT; de Vries N; Braem MJ; Van de Heyning PH. Observer variation in drug-induced sleep endoscopy: experienced versus nonexperienced ear, nose, and throat surgeons. *SLEEP* 2013;36(6):947-953.

INTRODUCTION

Obstructive sleep apnea (OSA) is a syndrome characterized by recurrent episodes of apnea and hypopnea during sleep that are caused by repetitive upper airway (UA) collapse and often result in decreased oxygen blood levels and arousal from sleep.¹ Continuous positive airway pressure (CPAP) therapy is considered the gold standard treatment for moderately severe OSA, but in patients with mild to moderate OSA or in cases

of CPAP failure, oral appliance therapy and/or UA surgery can be considered.²⁻⁴ Identification of predictors of treatment outcome is important in selecting patients who might benefit from non-CPAP therapies.⁵⁻⁷ However, the ability to preselect suitable patients for non-CPAP treatment in daily clinical practice is limited by criteria based on individual elements associated with UA behavior.

The aim of UA evaluation is not only to gain better insight into the complex pathophysiology of UA collapse but also to improve treatment success rates by selecting the most appropriate therapeutic option for the individual patient.^{8,9} Sleep nasendoscopy or drug-induced sleep endoscopy (DISE), first described by Croft and Pringle in 1991,¹⁰ uses a flexible nasopharyngoscope to visualize the UA under sedation. DISE enables evaluation of the localization of flutter and collapse in patients with sleep disordered breathing. During the procedure, UA collapse patterns can be assessed when alternatives to CPAP are con-

Submitted for publication July, 2012

Submitted in final revised form December, 2012

Accepted for publication December, 2012

Address correspondence to: Anneclaire V.M.T. Vroegop, MD, Antwerp University Hospital (UZA), Department of ENT, Head and Neck Surgery, Wilrijkstraat 10, 2650 Edegem, Antwerp, Belgium; Tel: +32 (0)3 821 33 85; Fax: +32 (0)3 821 42 71; E-mail: anneclaire.vroegop@uza.be

Table 1—Baseline characteristics of patients

Variables	All patients (n = 6)
Males (n, %)	5 (83.3%)
Age in years (mean ± SD)	51 ± 15
BMI in kg/m ² (mean ± SD)	25 ± 3
AHI, per hour sleep (mean ± SD)	21 ± 10

AHI, apnea-hypopnea index; BMI, body mass index; SD, standard deviation.

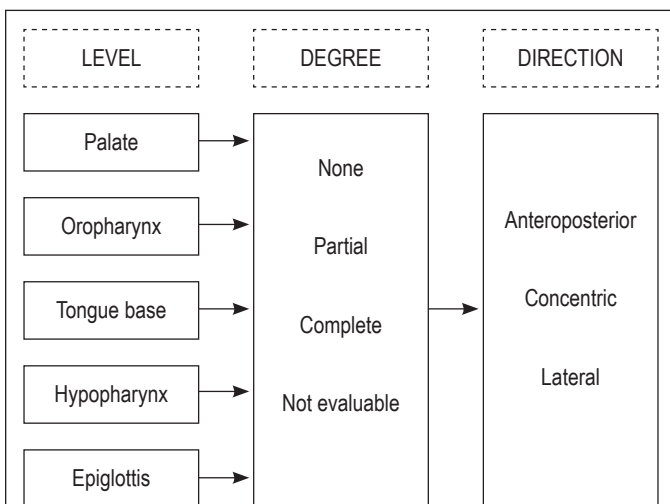


Figure 1—Reporting of the assessment of the level, the corresponding degree, and direction of upper airway collapse patterns.

sidered. It has shown its value in optimizing patient selection for surgical UA interventions and can also be helpful in selecting patients for oral appliance therapy.^{11,12} Its validity and reliability in experienced observers have been demonstrated in the literature^{13,14}; however, assessment is subjective and may vary based on clinical experience. Although several scoring systems have been introduced over the years,¹⁵⁻²³ no standard approach toward assessment and classification of DISE findings has been universally adopted.

DISE is a relatively new procedure to most general ear, nose, and throat (ENT) surgeons. Because further development and implementation of this procedure in the interdisciplinary field of sleep medicine is expected, assessment of the evaluation technique could be of interest to improve its application. The purpose of the current study was to assess differences in DISE evaluation by experienced versus nonexperienced observers. Therefore, observer agreement values were calculated in a cohort of 97 ENT surgeons, of whom seven were experienced with DISE.

METHODS

Study Design

This prospective study comprised a blinded multiobserver assessment of six DISE videos. These specific videos were selected because of the distinctive UA collapse patterns recorded. The DISE videos were shown to the observers twice in a randomized order, at two different time points.

Patients

The DISE videos originated from six different patients with OSA, who were reluctant about or non-responding to CPAP, or who had failed to use CPAP and underwent DISE (baseline characteristics shown in Table 1). Each DISE video represented one unique OSA patient. DISE was performed by an experienced ENT surgeon in a semidark and silent operating room with the patient lying supine in a hospital bed. Artificial sleep was induced by intravenous administration of midazolam (bolus injection of 1.5 mg) and propofol through a target-controlled infusion system (2.0-2.5 µg/mL). During the procedure, electrocardiography and oxygen saturation were continuously monitored. A fiberoptic laryngoscope (Olympus ENF-GP, diameter 3.7 mm, Olympus Europe GmbH, Hamburg, Germany) was used for videoscopic imaging, and simultaneous sound recording was performed.

Observers

The cohort consisted of 97 observers, who were invited to participate in this project during a 2-day conference. Ninety observers were general ENT surgeons with diverse clinical interests, who were nonexperienced with DISE. The other seven observers were experienced with DISE, as they perform the procedure on a regular basis in specialized sleep clinics and also conduct clinical research on this topic. A subset of observers was able to assess the DISE videos twice, as these observers were available on both days of the conference. All observers had no prior knowledge of patient history, including previous CPAP and non-CPAP treatment, polysomnography results, and findings on clinical examination or planned treatment. Each observer was asked to note his or her conclusions as to the level(s) of collapse and the degree and direction thereof, using a uniform scoring system (Figure 1). The scoring system consisted of anatomic UA landmarks and was based on key elements of scoring systems recently proposed in the literature.²¹⁻²⁴

Statistical Analysis

Data analysis was performed using R version 2.15.0 (R Foundation for Statistical Computing, Vienna, Austria). Descriptive statistics for clinical characteristics of patients were presented as means ± standard deviation (SD). Raw agreement among the observers was calculated by dividing the number of actual agreements by the total number of possible agreements. This was done for global agreement per UA level [overall agreement (OA)] and per specific response option [specific agreement (SA)].²⁵ Additionally, Fleiss kappa values were determined to correct for chance agreement. For the intraobserver agreement, raw agreement and kappa values were calculated per observer and averaged over all observers. For both interobserver and intraobserver agreement, 95% confidence intervals were obtained using bootstrap with 1,000 samples on the level of the observer.^{26,27} For the nonexperienced observers this means that 1,000 samples, containing 90 × 6 videos, were taken with replacement from the 90 observers. Similarly, for the experienced observers, 1,000 samples were taken from seven observers. Bootstrap P values for the differences between experienced and non-experienced raters were computed by pooling the experienced and nonexperienced raters together. Observers were then repeatedly artificially classified into experienced and nonexperienced in the same ratio as

the original sample and the number of times the differences exceeded the observed value were counted.

RESULTS

Interobserver Variation

Specific and overall interobserver agreement values per UA level are shown in Tables 2-6. In the nonexperienced group (n = 90), overall interobserver agreement was highest for presence of palatal collapse (OA = 0.88), although the kappa value for the presence of palatal collapse was lowest (-0.03). The agreement on the presence of tongue base collapse (OA = 0.63; kappa = 0.33) was followed by the agreement on epiglottis (OA = 0.57; kappa = 0.23) and oropharynx collapse (OA = 0.45; kappa = 0.09). Low overall interobserver agreement in this group was found for hypopharyngeal collapse (OA = 0.33; kappa = 0.08). A similar ranking was found for degree of collapse. For direction of collapse, high interobserver agreement was found for the palate (OA = 0.57; kappa = 0.16).

In the experienced group (n = 7), no observers scored “not evaluable” for presence of collapse at the palate, oropharynx, or tongue base (marked with a number sign (#) in Tables 2 through 6). Overall interobserver agreement was highest for presence of tongue base collapse (OA = 0.93; kappa = 0.71), followed by collapse of the palate (OA = 0.80; kappa = 0.51). In this group, lowest agreement was also found for hypopharyngeal collapse (OA = 0.47; kappa = 0.03). Interobserver agreement on direction of collapse was highest for epiglottis collapse (OA = 0.97; kappa = 0.97), followed by oropharynx (OA = 0.82; kappa = 0.35). Among the experienced observers, several specific structure-configurations for direction of collapse were not observed (marked with double number signs (##) in Tables 2 through 6). Concerning the degree of collapse, highest agreement was found for degree of oropharyngeal collapse (OA = 0.82; kappa = 0.66).

Among the experienced observers a statistically significant higher interobserver agreement was obtained for presence, direction, and degree of oropharyngeal collapse, as well as for presence of tongue base collapse and degree of epiglottis collapse.

Intraobserver Variation

The intraobserver agreement per UA level is presented in Table 7. Among the nonexperienced observers, high intraobserver agreement was found in particular for tongue base and epiglottis collapse. Among the experienced observers, high agreement was found for all levels but to a lesser extent for hypopharyngeal collapse. Intraobserver agreement was statistically significantly higher in the experienced group, for all UA levels except for the hypopharynx. In total, 52 nonexperienced observers and five experienced observers were available to score the videos twice.

DISCUSSION

This study describes variations in the interobserver and intraobserver agreement of DISE in a large cohort of

Table 2—Interobserver agreement on the palatal level

	Palate			
	Nonexp (N = 90)		Exp (N = 7)	
	N	Agreement	N	Agreement
Collapse				
Specific agreement				
Collapse	369	0.94 (0.91-0.96)	26	0.85 (0.68-0.98)
No collapse	19	0.20 (0.06-0.47)	12	0.75 (0.38-0.97)
Not evaluable	3	0.01 (0.00-0.05)	0	#
Overall agreement		0.88 (0.83-0.92)		0.80 (0.63-0.95)
Kappa		-0.03 (-0.37-0.31)		0.51 (0.45-0.84)
Direction				
Specific agreement				
Anteroposterior	174	0.64 (0.56-0.71)	13	0.60 (0.30-0.86)
Concentric	170	0.62 (0.55-0.68)	9	0.41 (0.11-0.74)
Lateral	10	0.05 (0.01-0.11)	0	##
Overall agreement		0.57 (0.51-0.63)		0.44 (0.23-0.74)
Kappa		0.16 (0.10-0.39)		0.27 (0.09-0.50)
Degree				
Specific agreement				
Complete	262	0.73 (0.68-0.78)	17	0.72 (0.47-0.92)
Partial	91	0.39 (0.30-0.48)	8	0.44 (0.11-0.79)
Overall agreement		0.59 (0.55-0.65)		0.59 (0.36-0.85)
Kappa		0.29 (0.16-0.36)		0.30 (0.06-0.71)

Interobserver agreement on the palatal level (Nonexp, nonexperienced observers; Exp, experienced observers), including 95% confidence intervals (in parenthesis). #, no experienced observers scored “not evaluable” for presence of palatal collapse; ##, none observed lateral palatal collapse.

Table 3—Interobserver agreement on the oropharyngeal level

	Oropharynx			
	Nonexp (N = 90)		Exp (N = 7)	
	N	Agreement	N	Agreement
Collapse				
Specific agreement				
Collapse	249	0.68 (0.61-0.73)	17	0.75 (0.50-0.92)
No collapse	67	0.23 (0.17-0.31)	22	0.82 (0.59-0.96)
Not evaluable	25	0.08 (0.05-0.12)	0	#
Overall agreement		0.45 (0.39-0.52)		0.80 (0.63-0.94)*
Kappa		0.09 (0.05-0.17)		0.48 (0.40-0.82)*
Direction				
Specific agreement				
Anteroposterior	19	0.12 (0.05-0.22)	0	##
Concentric	112	0.44 (0.36-0.51)	2	0.92 (0.67-1.00)
Lateral	106	0.46 (0.38-0.54)	14	0.25 (0.00-0.84)
Overall agreement		0.33 (0.28-0.39)		0.82 (0.42-1.00)**
Kappa		0.04 (0.01-0.14)		0.35 (-0.47-1.00)*
Degree				
Specific agreement				
Complete	65	0.36 (0.25-0.43)	4	0.93 (0.56-1.00)
Partial	165	0.57 (0.50-0.64)	12	0.60 (0.00-1.00)
Overall agreement		0.38 (0.31-0.44)		0.82 (0.42-1.00)*
Kappa		0.10 (0.02-0.23)		0.66 (0.22-1.00)**

Interobserver agreement on the oropharyngeal level (Nonexp, nonexperienced observers; Exp, experienced observers), including 95% confidence intervals (in parenthesis). #, no experienced observers scored “not evaluable” for presence of oropharyngeal collapse; ##, none observed anteroposterior oropharyngeal collapse. Asterisks indicate significant differences in agreement between nonexperienced and experienced observers: * < 0.05, ** < 0.01.

Table 4—Interobserver agreement on the tongue base level

	Tongue base			
	Nonexp (N = 90)		Exp (N = 7)	
	N	Agreement	N	Agreement
Collapse				
Specific agreement				
Collapse	276	0.79 (0.74-0.84)	29	0.96 (0.83-1.00)
No collapse	86	0.43 (0.36-0.52)	10	0.83 (0.57-1.00)
Not evaluable	4	0.01 (0.00-0.05)	0	#
Overall agreement		0.63 (0.57-0.70)		0.93 (0.80-1.00)*
Kappa		0.33 (0.26-0.41)		0.71 (0.40-1.00)*
Direction				
Specific agreement				
Anteroposterior	199	0.69 (0.63-0.75)	25	0.88 (0.72-0.98)
Concentric	60	0.33 (0.25-0.43)	0	##
Lateral	4	0.02 (0.00-0.06)	0	##
Overall agreement		0.51 (0.45-0.58)		0.76 (0.51-0.97)
Kappa		0.04 (-0.02-0.14)		-0.14 (-0.36-0.26)
Degree				
Specific agreement				
Complete	76	0.33 (0.27-0.39)	14	0.58 (0.30-0.84)
Partial	184	0.62 (0.56-0.68)	12	0.49 (0.24-0.78)
Overall agreement		0.45 (0.40-0.52)		0.50 (0.32-0.75)
Kappa		0.12 (0.05-0.21)		0.03 (-0.01-0.58)

Interobserver agreement on the tongue base level (Nonexp, nonexperienced observers; Exp, experienced observers), including 95% confidence intervals (in parenthesis). #, no experienced observers scored "not evaluable" for presence of tongue base collapse; ##, none of them observed concentric or lateral tongue base collapse. *Significant differences in agreement between nonexperienced and experienced observers.

Table 5—Interobserver agreement on the hypopharyngeal level

	Hypopharynx			
	Nonexp (N = 90)		Exp (N = 7)	
	N	Agreement	N	Agreement
Collapse				
Specific agreement				
Collapse	166	0.51 (0.44-0.58)	10	0.45 (0.16-0.72)
No collapse	114	0.37 (0.31-0.44)	21	0.59 (0.38-0.79)
Not evaluable	43	0.11 (0.08-0.16)	5	0.016 (0.00-0.42)
Overall agreement		0.33 (0.28-0.38)		0.47 (0.31-0.66)
Kappa		0.08 (0.05-0.14)		0.03 (0.00-0.47)
Direction				
Specific agreement				
Anteroposterior	37	0.19 (0.12-0.27)	0	##
Concentric	71	0.53 (0.40-0.65)	1	0.00 (0.00-0.71)
Lateral	45	0.21 (0.14-0.27)	8	0.71 (0.25-1.00)
Overall agreement		0.24 (0.17-0.31)		0.44 (0.14-0.95)
Kappa		0.02 (0.00-0.12)		0.05 (-0.18-0.64)
Degree				
Specific agreement				
Complete	32	0.17 (0.11-0.25)	2	0.71 (0.25-1.00)
Partial	118	0.51 (0.44-0.59)	8	0.33 (0.00-1.00)
Overall agreement		0.28 (0.22-0.35)		0.54 (0.20-0.98)
Kappa		-0.06 (-0.09-0.03)		0.09 (-0.19-0.68)

Interobserver agreement on the hypopharyngeal level (Nonexp, nonexperienced observers; Exp, experienced observers), including 95% confidence intervals (in parenthesis). ##, among the experienced observers, anteroposterior hypopharyngeal collapse was not observed.

ENT surgeons, consisting of both experienced and non-experienced DISE observers. The current results indicate that as a rule, agreement is higher among experienced observers, and therefore experience in performing DISE is necessary to obtain reliable results. Training could be beneficial to ENT surgeons new to this emerging technique. Higher interobserver agreement was found in particular for oropharyngeal, tongue base, and epiglottis collapse. Intraobserver agreement was higher among the experienced raters, also in particular for the oropharynx, tongue base, and epiglottis (Table 7).

The difference in interobserver agreement on palatal collapse between the nonexperienced and experienced ENT surgeons might be due to a higher tendency of non-experienced observers to see a collapse at this specific level, as it is the most common site of snoring¹⁹ but does not necessarily implicate collapse. The differences in overall interobserver agreement for this specific UA level might thus be explained by an artificially high agreement among the nonexperienced observers, rather than a substantially lower agreement among the experienced raters. The low kappa value for presence of collapse of this specific UA level in the nonexperienced group might endorse this explanation. Least interobserver agreement was found for hypopharyngeal collapse in both the non-experienced and experienced groups. General ENT surgeons regularly perform endoscopic UA procedures as part of their routine clinical practice but might be less accustomed to evaluation of the hypopharynx as an isolated entity of the UA tract. The low agreement for hypopharyngeal collapse in the experienced group might be attributed to the use of different scoring systems in specific specialized centers,^{13,15,21} in which qualification of the hypopharynx is particularly subject to different approaches. Also, evaluation of the hypopharyngeal level in general can be hindered by collapse at a higher UA level where the tip of the nasopharyngoscope is positioned to observe the hypopharynx. This phenomenon is not present when assessing the palate, as the nasopharyngoscope is then positioned in the nasopharynx, not being a part of the collapsible segment of the UA. Furthermore, the hypopharynx has less defined structural boundaries, compared with other UA levels, e.g., the palate. The experienced observers appeared to find identification of the presence of UA collapse at the level of the palate, oropharynx, and tongue base less cumbersome, having not scored "not evaluable" for these levels, unlike the nonexperienced observers. Also, the absence of certain structure-configurations in the ratings of the experienced observers is prominent. This is particularly true for the absence of lateral palatal collapse, anteroposterior oropharyngeal collapse, concentric or lateral tongue base collapse, anteroposterior hypopharyngeal collapse, and lateral epiglottis collapse.

This study is the first to specifically describe differences in interobserver and intraobserver agreement between nonexperienced and experienced ENT surgeons. The current results on interobserver agreement in our subgroup of experienced ENT surgeons are in accor-

dance with previous findings in the field, although study designs are roughly comparable.^{13,14} Kezirian et al.¹³ found that the reliability of global assessment of obstruction was somewhat higher than the degree of obstruction, especially for the hypopharynx. Rodriguez-Bruno et al.¹⁴ found high interobserver and intraobserver agreement for obstruction at the level of the tonsils, followed by the epiglottis. There are major differences with the current study: the number of experienced observers (Kezirian et al. study: two; Rodriguez-Bruno et al. study: two; current study: seven), number of video segments shown (Kezirian et al. study: 108; Rodriguez-Bruno et al. study: 32; current study: six), and the type of scoring system used. Furthermore, Kezirian et al.¹³ found that among experienced surgeons, assessment of the palate, tongue, and epiglottis showed greater reliability than for other structures, whereas in the current study, higher agreement was found for palatal, oropharyngeal, and tongue base collapse. As the palate and tongue base are better-defined UA structures, this might partially explain the findings of consistently better reliability on these specific levels.

The current study provides important additional insights into what is known of interobserver and intraobserver variability in DISE, as it was conducted in a large cohort of general ENT surgeons, most of whom were nonexperienced with DISE and not yet specifically trained to assess DISE findings. All observers were blinded to baseline characteristics of the study participants, so no confounding effects of variables such as the apnea-hypopnea index or body mass index were to be expected. In recent literature on the value of DISE, it was suggested that DISE is most relevant when considering tongue base surgery or oral appliance therapy⁹; therefore, the high interobserver and intraobserver agreement found for tongue base collapse among the experienced ENT surgeons in the current study might be of particular

importance. Furthermore, results on interobserver and intraobserver agreement in a cohort of both experienced and nonexperienced observers may help to detect differences in assessment for varying levels of training. A possible solution to the unreliability of DISE findings on particular levels could be training ENT surgeons with a set of standardized DISE videos to foster consistency for scoring the UA level variables. To do so, the

Table 6—Interobserver agreement on the epiglottis level

	Epiglottis			
	Nonexp (N = 90)		Exp (N = 7)	
	N	Agreement	N	Agreement
Collapse				
Specific agreement				
Collapse	189	0.71 (0.63-0.78)	14	0.82 (0.53-0.97)
No collapse	156	0.56 (0.49-0.63)	19	0.71 (0.45-0.92)
Not evaluable	13	0.19 (0.06-0.47)	5	0.67 (0.00-1.00)
Overall agreement		0.57 (0.51-0.63)		0.74 (0.57-0.92)
Kappa		0.23 (0.16-0.34)		0.51 (0.41-0.77)
Direction				
Specific agreement				
Anteroposterior	153	0.82 (0.75-0.88)	12	1.00 (1.00-1.00)
Concentric	19	0.22 (0.09-0.42)	1	0.33 (0.00-1.00)
Lateral	8	0.13 (0.02-0.73)	0	##
Overall agreement		0.63 (0.53-0.73)		0.97 (0.72-1.00)
Kappa		-0.17 (-0.25-0.15)		0.97 (0.71-1.00)
Degree				
Specific agreement				
Complete	77	0.51 (0.42-0.61)	9	0.75 (0.35-1.00)
Partial	98	0.48 (0.40-0.57)	4	0.40 (0.00-1.00)
Overall agreement		0.41 (0.35-0.48)		0.61 (0.30-0.99)
Kappa		-0.05 (-0.08-0.19)		0.61 (-0.02-0.79)*

Interobserver agreement on the epiglottis level (Nonexp, nonexperienced observers; Exp, experienced observers), including 95% confidence intervals (in parenthesis). ##, among the experienced observers, lateral epiglottis collapse was not observed. *Significant differences in agreement between nonexperienced and experienced observers.

Table 7—Intraobserver agreement

	Nonexperienced (N = 52)				Experienced (N = 5)			
	Raw agreement		Kappa		Raw agreement		Kappa	
	Median Q1-Q3	Bootstrap Mean	Median Q1-Q3	Bootstrap Mean	Median Q1-Q3	Bootstrap Mean	Median Q1-Q3	Bootstrap Mean
Palate	1.0 (0.8-1.0)	0.92 (0.88-0.95)	0.00 (-0.05-0.00)	-0.01 (-0.04-0.02)	1.0 (1.0-1.0)	0.96 (0.94-0.98)	0.43 (0.00-1.00)	0.49*** (0.39-0.59)
Oropharynx	0.8 (0.4-0.8)	0.62 (0.56-0.68)	0.00 (-0.11-0.17)	0.07 (-0.01-0.15)	0.8 (0.6-0.8)	0.76 (0.72-0.80)	0.52 (0.27-0.52)	0.49** (0.42-0.56)
Tongue base	0.8 (0.6-1.0)	0.74 (0.69-0.80)	0.40 (0.00-0.64)	0.38 (0.28-0.48)	1.0 (0.8-1.0)	0.92 (0.90-0.94)	1.00 (0.52-1.00)	0.81* (0.75-0.87)
Hypopharynx	0.6 (0.2-0.8)	0.56 (0.48-0.64)	0.09 (-0.01-0.53)	0.19 (0.11-0.28)	0.6 (0.4-0.8)	0.60 (0.53-0.66)	0.38 (-0.10-0.52)	0.30 (0.20-0.41)
Epiglottis	0.8 (0.6-1.0)	0.76 (0.71-0.81)	0.52 (0.30-0.67)	0.52 (0.45-0.60)	1.0 (0.8-1.0)	0.92 (0.90-0.99)	1.00 (0.66-1.00)	0.84* (0.79-0.89)

Intraobserver agreement on levels of upper airway collapse (Q, quartile; CI, confidence interval). 95% confidence intervals in parenthesis. Asterisks indicate significant differences in agreement between nonexperienced and experienced observers: * < 0.05, ** < 0.01, *** < 0.001.

use of a standardized and universally accepted DISE scoring system will also be essential.

There are several limitations to this study. Ideally, intraobserver agreement is to be determined based on more ratings, which was practically not feasible in the study setting. A fuller set of DISE samples, potentially including more different patterns, would have allowed more robust reliability testing of the scoring system. Furthermore, there were only seven observers in the experienced group. However, the current setting was considered sufficient to demonstrate and verify differences in agreement in experienced versus nonexperienced observers—the main goal of this study. As for the statistical analysis, it is to be mentioned that kappa coefficients have an important drawback in the current study design, because the kappa value is heavily dependent on the observed marginal frequencies. Some specific UA collapse configurations (i.e., lateral palatal collapse, anteroposterior oropharyngeal collapse) are generally rare or unlikely to be observed. This leads to lower kappa values and can therefore be severely misleading in interpreting the agreement among the observers. The observers participating in the current study were possibly acquainted with different DISE scoring systems at their home clinics. Using an unknown uniform study scoring system while assessing the DISE video segment may have been cumbersome. Furthermore, only segments of the DISE videos were shown, although this was considered a minor limitation, as the main goal was to assess interobserver and intraobserver variation and not to provide a treatment plan. The heterogeneity of the observer group is to be considered relative in this context. Indeed, there are differences in levels of experience within the subcategories of nonexperienced and experienced. However, this aspect reflects clinical reality and may help to emphasize the need for specific training in DISE assessment in daily clinical practice.

In conclusion, the reported findings on interobserver and intraobserver agreement in a cohort of ENT surgeons, both nonexperienced and experienced with DISE, demonstrate that overall observer agreement was higher in experienced versus nonexperienced ENT surgeons, suggesting that experience in performing DISE is necessary to obtain reliable observations. Training under guidance of an experienced surgeon might be helpful for those unexperienced with DISE. These findings are new to the field and add additional insights into the validity of the technique. If DISE is to be extended into general ENT practice, the need for a standardized scoring system increases.

ABBREVIATIONS

CPAP, continuous positive airway pressure
DISE, drug-induced sleep endoscopy
ENT, ear, nose and throat
OSA, obstructive sleep apnea
OA, overall agreement
SA, specific agreement
SD, standard deviation
UA, upper airway

ACKNOWLEDGMENTS

The authors express their gratitude to all ENT surgeons who collaborated in this project. Parts of this study were presented

as a poster at the SLEEP 2012 26th Annual Meeting of the Associated Professional Sleep Societies (APSS). The work was supported by the Agency for Innovation by Science and Technology (IWT) of the Flemish Government (Belgium), project number 090864.

DISCLOSURE STATEMENT

This was not an industry supported study. Dr. Hamans has consulted for Philips Healthcare. Dr. Hohenhorst is investigator for Inspire Medical Systems, Inc. Dr. Kezirian is on the Medical Advisory Board of Apnex Medical and ReVENT Medical and has consulted for Apnex Medical, ArthroCare, Medtronic, Pavad Medical, and Magnap (Intellectual Property Rights). Dr. de Vries is a medical advisor for MSD, ReVENT, and Night-Balance; investigator for Inspire Medical Systems, Inc.; consultant for Philips; and has stock options in ReVENT. The other authors indicated no financial conflicts of interest. No product was used for an indication not in the approved labeling, and no investigational products were used.

REFERENCES

1. Sleep-related breathing disorders in adults: recommendations for syndrome definition and measurement techniques in clinical research. The Report of an American Academy of Sleep Medicine Task Force. *Sleep* 1999; Aug 1:667-89.
2. Sullivan CE, Issa FG, Berthon-Jones M, Eves L. Reversal of obstructive sleep apnoea by continuous positive airway pressure applied through the nares. *Lancet* 1981;1:862-5.
3. Epstein LJ, Kristo D, Strollo PJ Jr, et al. Clinical guideline for the evaluation, management and long-term care of obstructive sleep apnea in adults. *J Clin Sleep Med* 2009;5:263-76.
4. Kushida CA, Morgenthaler TI, Littner MR, et al. Practice parameters for the treatment of snoring and Obstructive Sleep Apnea with oral appliances: an update for 2005. *Sleep* 2006;29:240-3.
5. Kezirian EJ, Goldberg AN. Hypopharyngeal surgery in obstructive sleep apnea: an evidence-based medicine review. *Arch Otolaryngol Head Neck Surg* 2006;132:206-13.
6. Vanderveken OM, Vroegop AV, Van de Heyning P, Braem MJ. Drug-induced sleep endoscopy completed with a simulation bite approach for the prediction of the outcome of treatment of obstructive sleep apnea with mandibular repositioning appliances. *Operative Techniques in Otolaryngology - head and neck surgery* 2011;22:175-82.
7. Johal A, Battagel JM, Kotecha BT. Sleep nasendoscopy: a diagnostic tool for predicting treatment success with mandibular advancement splints in obstructive sleep apnoea. *Eur J Orthod* 2005;27:607-14.
8. Vanderveken OM. Drug-induced sleep endoscopy (DISE) for non-CPAP treatment selection in patients with sleep-disordered breathing. *Sleep Breath* 2012 (Epub ahead of print).
9. Eichler C, Sommer JU, Stuck BA, Hormann K, Maurer JT. Does drug-induced sleep endoscopy change the treatment concept of patients with snoring and obstructive sleep apnea? *Sleep Breath* 2012 (Epub ahead of print).
10. Croft CB, Pringle M. Sleep nasendoscopy: a technique of assessment in snoring and obstructive sleep apnoea. *Clin Otolaryngol Allied Sci* 1991;16:504-9.
11. Johal A, Hector MP, Battagel JM, Kotecha BT. Impact of sleep nasendoscopy on the outcome of mandibular advancement splint therapy in subjects with sleep-related breathing disorders. *J Laryngol Otol* 2007;121:668-75.
12. Hessel NS, de Vries N. Results of uvulopalatopharyngoplasty after diagnostic workup with polysomnography and sleep endoscopy: a report of 136 snoring patients. *Eur Arch Otorhinolaryngol* 2003;260:91-5.
13. Kezirian EJ, White DP, Malhotra A, Ma W, McCulloch CE, Goldberg AN. Interrater reliability of drug-induced sleep endoscopy. *Arch Otolaryngol Head Neck Surg* 2010;136:393-7.
14. Rodriguez-Bruno K, Goldberg AN, McCulloch CE, Kezirian EJ. Test-retest reliability of drug-induced sleep endoscopy. *Otolaryngol Head Neck Surg* 2009;140:646-51.

15. Pringle MB, Croft CB. A grading system for patients with obstructive sleep apnoea-based on sleep nasendoscopy. *Clin Otolaryngol Allied Sci* 1993;18:480-4.
16. Camilleri AE, Ramamurthy L, Jones PH. Sleep nasendoscopy: what benefit to the management of snorers? *J Laryngol Otol* 1995;109:1163-5.
17. Sadaoka T, Kakitsuba N, Fujiwara Y, Kanai R, Takahashi H. The value of sleep nasendoscopy in the evaluation of patients with suspected sleep-related breathing disorders. *Clin Otolaryngol Allied Sci* 1996;21:485-9.
18. Higami S, Inoue Y, Higami Y, Takeuchi H, Ikoma H. Endoscopic classification of pharyngeal stenosis pattern in obstructive sleep apnea hypopnea syndrome. *Psychiatry Clin Neurosci* 2002;56:317-8.
19. Quinn SJ, Daly N, Ellis PD. Observation of the mechanism of snoring using sleep nasendoscopy. *Clin Otolaryngol Allied Sci* 1995;20:360-4.
20. Iwanaga K, Hasegawa K, Shibata N, et al. Endoscopic examination of obstructive sleep apnea syndrome patients during drug-induced sleep. *Acta Otolaryngol Suppl* 2003;550:36-40.
21. Kezirian EJ, Hohenhorst W, de Vries N. Drug-induced sleep endoscopy: the VOTE classification. *Eur Arch Otorhinolaryngol* 2011;268:1233-6.
22. Bachar G, Nageris B, Feinmesser R, et al. Novel grading system for quantifying upper-airway obstruction on sleep endoscopy. *Lung* 2012;190:313-8.
23. Vicini C, De Vito A, Benazzo M, et al. The nose oropharynx hypopharynx and larynx (NOHL) classification: a new system of diagnostic standardized examination for OSAHS patients. *Eur Arch Otorhinolaryngol* 2012;269:1297-300.
24. Hamans E, Meeus O, Boudewyns A, Saldien V, Verbraecken J, Van de Heyning P. Outcome of sleep endoscopy in obstructive sleep apnoea: the Antwerp experience. *B-ENT* 2010;6:97-103.
25. Uebersax JS. A design-independent method for measuring the reliability of psychiatric diagnosis. *J Psychiatr Res* 1982;17:335-42.
26. Efron B. Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. *Biometrika* 1981;68:589-99.
27. Efron B, Tibshirani R. An introduction to the bootstrap. Boca Raton, FL: Chapman & Hall / CRC, 1993.