Test-retest reliability of drug-induced sleep endoscopy

Krista Rodriguez-Bruno, MD,
Department of Otolaryngology—Head and Neck Surgery, University of California, San Francisco

Andrew N. Goldberg, MD, MSCE,
Department of Otolaryngology—Head and Neck Surgery, University of California, San Francisco

Charles E. McCulloch, PhD, and
Department of Epidemiology and Biostatistics, University of California, San Francisco

Eric J. Kezirian, MD, MPH
Department of Otolaryngology—Head and Neck Surgery, University of California, San Francisco

Abstract

Objective—To determine the test-retest reliability of drug-induced sleep endoscopy (DISE).

Methods—Prospective cohort study. Patients with OSA underwent two separate DISE examinations. The following outcomes were measured: a global assessment of obstruction at the palate and/or hypopharynx; the degree of obstruction at the palate and hypopharynx; and contribution of individual structures (palate, tonsils, tongue, epiglottis, and lateral pharyngeal walls) to obstruction.

Results—32 patients underwent two separate DISE examinations. The apnea-hypopnea index was 40.7 ± 21.1, and the lowest oxygen saturation was $79.8\pm17.4\%$. Point estimates for the intraclass correlation coefficient analogs related to palatal obstruction ranged from 0.41-0.89; related to the hypopharyngeal airway, the point estimates ranged from 0.57-0.84.

Conclusion—The test-retest reliability of DISE appears good, especially in the evaluation of the hypopharyngeal airway. Larger studies can generate more-precise confidence interval estimates and test the generalizability of these findings.

Introduction

Obstructive sleep apnea (OSA) is characterized by intermittent, repeated upper airway narrowing or obstruction occurring during sleep. Treatment options include behavioral measures, positive airway pressure, surgery, and oral appliances. Positive airway pressure therapy is recognized as the first-line treatment for moderate-severe OSA patients because it

eliminates disordered breathing events, but patients who do not tolerate this treatment modality may consider other options, including surgery.

Airway obstruction in OSA can occur at many levels, and the principal regions of dynamic obstruction are the palate and the so-called hypopharynx (actually corresponding to the hypopharynx and the retrolingual portion of the oropharynx). Surgical procedures are inherently directed at specific regions of the upper airway, and by addressing airway obstruction in a targeted fashion, it may be possible to tailor surgical treatment to a patient's specific pattern of obstruction—improving surgical results and/or minimizing the scope of surgical intervention and the associated risks. OSA surgical outcomes are associated with the pattern of airway obstruction occurring during sleep. Among those undergoing palate surgery with or without tonsillectomy, patients with presumed isolated palatal obstruction have demonstrated better outcomes than those with at least some component of hypopharyngeal obstruction.[1, 2] Treating hypopharyngeal airway obstruction is associated with improved surgical outcomes in the latter group.[3]

In this light, a major goal of surgical upper airway evaluation in OSA is determining the pattern of obstruction. The authors of the largest meta-analysis of surgical treatment of OSA wrote that the failure to identify and treat all levels of airway obstruction was the principal factor in not achieving optimal surgical results.[1] A Cochrane Collaboration review of surgical treatment for OSA indicated that determination of the site of obstruction should be a major focus of research efforts in sleep-disordered breathing.[4] Techniques that are most commonly used to characterize the pattern of obstruction utilize specific components of physical examination,[2] awake fiberoptic examination with Muller maneuver,[5-7] and lateral cephalometry.[6] Because they are all performed on the awake patient and because most also provide a static rather than dynamic evaluation, they may not be ideal methods to assess the behavior of the upper airway during sleep.

Drug-induced sleep endoscopy (DISE) avoids these drawbacks and may provide a more accurate evaluation of the upper airway. First described in 1991, the technique requires the pharmacologic induction of sleep and the placement of a flexible fiberoptic telescope (passed through the nose) to visualize the upper airway.[8, 9] This provides an opportunity to observe directly and characterize the upper airway collapse that occurs during drug-induced sleep. Several centers around the world have shown that DISE is a safe, feasible, and valid assessment of the upper airway but have not studied important qualities of diagnostic tests such as reliability.[10-14] Inter-rater reliability has been studied by our group and was moderate-good.[15] The objective of this study was to examine the test-retest reliability of DISE.

Methods

This prospective cohort study included patients seen at the University of California, San Francisco Department of Otolaryngology—Head and Neck Surgery by one author (EJK). Inclusion criteria encompassed the following: age > 18 years, apnea-hypopnea index (AHI) > 5 diagnosed by overnight sleep study, and unable to tolerate positive airway pressure therapy. Exclusion criteria included pregnant women and allergy to propofol or cross-reaction to components of propofol such as egg lecithin or soybean oil. This study was approved by the University of California, San Francisco institutional review board.

All patients underwent two separate DISE examinations (referred to as DISE Exam 1 and DISE Exam 2). These were carried out in the operating room on different days. The first examination was performed as an isolated diagnostic evaluation, and the second was performed on the day of planned OSA surgical treatment immediately prior to that procedure.

The DISE technique has been described previously.[9] Topical decongestant was applied to both nasal cavities, and a topical anesthetic/decongestant mixture was applied to one nasal cavity. Patients were placed in a supine position on the operating room table with lights dimmed. Oximetry and cardiac rhythms were monitored by the Anesthesia team throughout the procedure, and supplemental oxygen was administered by blow-by facemask or nasal cannula as necessary. The intravenous infusion of propofol was used as the sole agent to achieve drug-induced sleep, with a target level of anesthesia of arousal to loud verbal stimulation (similar to a Modified Ramsay Score of 5 or Observer's Assessment of Alertness/Sedation Score of 3-4). The initial infusion rate of propofol was 50-75 mcg/kg/min, and the rate was adjusted to meet the target level of anesthesia. The infusion rate required for induction of sleep was recorded. With the onset of drug-induced sleep, the flexible fiberoptic laryngoscope was passed through the anesthetized nasal cavity to perform the examination, which was digitally recorded.

The video images of the recorded DISE procedures were later evaluated concurrently by an unblinded (EJK) and blinded (ANG) surgeon. The unblinded surgeon was aware of the patient identity when performing the procedure and reviewing the video images, but the blinded surgeon reviewed the images with only knowledge of whether or not the patient had undergone prior tonsillectomy. The blinded surgeon had no prior knowledge of other history or physical exam findings, sleep study results, or planned procedures. To examine the variability in reviewer assessments, each of the two DISE examinations was reviewed twice by each surgeon (Evaluation 1 and Evaluation 2) on separate days 2-6 weeks apart. This produced a total of eight evaluations per patient: two evaluations of each of two DISE examinations for each of two reviewers.

Each evaluation included three analyses. Analysis I was a global dichotomous assessment of obstruction at each of two levels: the palate and the hypopharynx. Analysis II reflected the degree of palatal and hypopharyngeal obstruction. This was graded separately for each region subjectively and categorized in an ordinal fashion as <50 %, 50-75%, and > 75% obstruction; these assessments did not include a quantitative assessment of airway dimensions but were intended to represent no/mild, moderate, and severe obstruction. Analysis III centered on individual regions of the pharynx and specific structures. It included a determination of which structure at the level of the palate and hypopharynx was the primary factor in airway obstruction, if present, and a dichotomous assessment of whether each of the individual structures contributed to airway obstruction. Structures were grouped as those at the level of the palate (palate, tonsils when present, and lateral pharyngeal walls at the level of the velopharynx) and the hypopharynx (tongue, epiglottis, and lateral pharyngeal walls at the level of the hypopharynx).

Descriptive statistics were calculated for baseline patient characteristics, and results are reported with mean ± standard deviation. Summary statistics for DISE findings were also calculated. Test-retest reliability in this study includes three sources of variation in DISE examination analyses: those attributable to patient variation between DISE examinations, to reviewer variation in evaluation of the same DISE examination, and to reviewer variation in the differential scoring of the same examination. Test-retest reliability was determined using generalized linear latent and mixed model estimation. Generalized linear latent and mixed models are one type of random effects model that, in this case, enables multilevel clustering of these evaluations by patient, examination number, and reviewer. By determining the variance in DISE findings attributable to different sources, it is possible to calculate results analogous to intraclass correlation coefficients for logistic models.[16] Four separate such test-retest intraclass correlation coefficient analogs are generated to describe the test-retest reliability: correlation from Evaluation 1 to Evaluation 2 by the same reviewer and same exam (ICC Evaluation), correlation from blinded to unblinded reviewer for the same exam (ICC Reviewer), correlation from Exam 1 to Exam 2 for the same reviewer (ICC Exam), and

correlation across exams and reviewers (ICC Reviewer-Exam). For all results, point estimates of the intraclass correlation analogs are reported with 95% confidence intervals. Confidence intervals were generated from bootstrapping techniques,[17] which generate more accurate confidence intervals for such nonlinear functions, except for the evaluation of primary structure contributing to airway obstruction in Analysis III, where the delta-method[18] and large sample confidence intervals were used because it was not possible to perform bootstrapping. Statistical analyses were conducted using Stata Version 10.0 (StataCorp LP, College Station, TX), and the gllamm module[19] was used within Stata to fit multilevel models. Because the primary goal of this study was to estimate reliability of assessments and because of the complicated methodologies (multi-level modeling and bootstrapping), accurate sample size calculations were not possible.

Results

Thirty-two patients were enrolled and underwent two separate DISE examinations between 2004 and 2008. Patient age was 42.7 ± 9.7 (range 20 - 66) years, and 9% (3/32) were female. Mean AHI was 40.7 ± 21.1. The lowest oxygen saturation on preoperative sleep study was $79.8 \pm 17.4\%$, and on average patients spent $15.3 \pm 22.4\%$ of the sleep study with oxygen saturations below 90%. Eleven (34%) patients had undergone prior tonsillectomy, and no patients had other previous pharyngeal surgery. The interval between DISE examinations for a given patient was 118 ± 85 days, and there was no statistically significant change in mean body mass index during this time interval.

The propofol infusion rate required to achieve drug-induced sleep was 116 ± 23 (range 75-175) mcg/kg/min, and the difference in infusion rates between the two exams for each patient was 12.5 ± 12.7 mcg/kg/min. In virtually all cases that required a rate greater than 125 mcg/kg/min, the infusion rate was decreased immediately after clinical determination of sleep onset with tolerance of endoscope placement. Total propofol dose was not recorded, as the length of time of evaluation during drug-induced sleep varied widely.

Summary statistics for DISE findings are presented in Table 1. Almost all patients demonstrated evidence of obstruction at the level of the palate, and a large majority of patients also demonstrated obstruction at the level of the hypopharynx (Analyses I and II). There was diversity in the patterns of obstruction, especially in the structures which appeared to play a primary role in obstruction at each level and in whether individual structures contributed to obstruction (Analysis III).

Test-retest reliability results are presented in Table 2. As a group, the ICC analog point estimates were higher for evaluations related to the hypopharynx than for those related to the palate, although there was considerable overlap for the confidence intervals. Although the test-retest reliability for the presence or absence of hypopharyngeal obstruction (Analysis I) was somewhat higher for the palate than the hypopharynx, the reverse was true for the degree of hypopharyngeal obstruction (Analysis II) and both the primary and individual structures contributing to hypopharyngeal obstruction (Analysis III). In particular, the reliability of assessments related to the lateral pharyngeal wall at the level of the velopharynx was lower than for other components of the analyses. For Analysis III, the primary structure contributing to obstruction (or whether there was felt to be no obstruction) at the palate and hypopharynx was higher than for individual structures at each level.

ICC Evaluation values, reflecting the correlation for ratings of each reviewer for repeated evaluations of the same exam, were generally the highest, although results for ICC Reviewer and ICC Exam were roughly similar. ICC Reviewer-Exam values, reflecting the correlation for ratings across reviewers and exams, were the lowest. There was no substantial difference

between the blinded and unblended reviewers for the test-reliability of the primary structures contributing to airway obstruction, the only analysis in which the assessments of individual reviewers was considered separately.

Discussion

DISE has good test-retest reliability, especially in its evaluation of the hypopharyngeal airway. The interpretation of ICC analogs, similar to Cohen's kappa values, is controversial, but one framework has been proposed by Landis and Koch.[20] While the coefficient point estimates and their confidence intervals overlap multiple categories, overall the level of agreement is moderate to substantial.

Test-retest reliability compares results from two distinct tests, and there are four potential comparisons that describe the variation in findings. The lowest (in magnitude) of the four estimates of reliability presented here is generally that for the ICC Reviewer-Exam, comparing findings across reviewers and exams (one reviewer's evaluations of one exam to the other reviewer's evaluations of a distinct exam). Because this incorporates multiple levels of variation, it is not surprising that the ICC analog values were the lowest, although still largely in the range of moderate-substantial agreement. There were no systematic differences in the point estimates and confidence intervals among the other three measures of test-retest reliability. Future studies incorporating more reviewers, evaluations of the same exam, examinations of each patient, or more patients may generate more precise estimates of these aspects of test-retest reliability.

DISE offers a unique structure-based assessment of the airway, compared to other commonly-used evaluation techniques. We present a region- (Analyses I and II) and structure-based (Analysis III) method to serve two major purposes of OSA surgical upper airway evaluation: characterizing the pattern of airway obstruction and selecting among treatments. We believe that identifying the primary structure contributing to obstruction in each segment of the airway serves both purposes, and the test-retest reliability of this specific assessment was higher than that for the involvement of individual structures (including those with primary and secondary roles).

There is wide variation in the DISE classification schemes presented in the literature, and we attempted to balance completeness and simplicity to describe the variation in patterns of upper airway obstruction. The upper airway does not consist of two independent regions (palate and hypopharynx), each containing various structures that can contribute to airway obstruction in isolation; instead, these two regions and the various structures have dynamic interactions that are not understood completely. Any attempt to simplify these relationships will have important deficiencies, and we anticipate revisions to our method over time. In fact, the first modification to our original method was based on the idea that it was important to determine the primary structures contributing to airway obstruction in each region (as in Analysis III).

Because surgical procedures are ultimately directed at specific structures, DISE may improve procedure selection and outcomes. This is especially true for the hypopharyngeal airway, where evaluation of the hypopharyngeal airway—and the choice among treatment options—is often a critical factor in surgical decision making. The three structures that most commonly contribute to hypopharyngeal airway obstruction are the tongue, epiglottis, and lateral pharyngeal walls, and the results for Analysis III indicate that DISE can differentiate their contributions to airway obstruction with good test-retest reliability. The array of surgical and non-surgical treatment options to treat the hypopharyngeal airway may exert differential effects on these various structures. For example, the genioglossus advancement and tongue radiofrequency procedures likely produce greater changes in tongue position during sleep than

in the lateral pharyngeal walls. The hyoid suspension may have less effect on tongue position but may alter the behavior of the epiglottis and/or lateral pharyngeal walls during sleep. Because there is choice among these procedures for treatment of the hypopharynx, a diagnostic test may be most valuable if it not only determines whether hypopharyngeal obstruction is present but also which structures contribute most to that obstruction.

For surgical treatment of the palatal airway, DISE may not differentiate palate vs. velopharynx lateral pharyngeal wall obstruction as well as for the hypopharyngeal structures, based on the lower point estimates and wider confidence intervals. The implications are unclear. The most common surgical treatment for palatal obstruction in previously-untreated OSA patients is uvulopalatopharyngoplasty, with tonsillectomy in most patients without previous tonsillectomy. Because a similar surgical approach is used for patients regardless of whether the soft palate or velopharynx lateral pharyngeal walls contribute more to obstruction, the question of whether a patient has palate-level obstruction or not (as in Analysis I) may be more important than determining whether specific structures contribute to collapse (Analysis III). Because almost all patients in this study demonstrated palatal obstruction, the confidence intervals for Analysis I for the palate were wide, suggesting that the test-retest reliability cannot be determined precisely by this sample. Differentiating palate vs. velopharynx-level lateral pharyngeal wall obstruction based on DISE (Analysis III) appears more challenging. Again, the importance of this distinction is unclear; with the adoption of a wider variety of first-line palate procedure, this may prove more important.

This study is not without limitations. First, the confidence intervals for many ICC analog estimates were somewhat wide. We believe that the pattern of the estimates and confidence intervals is more important than any single result and that the pattern suggests that the test-retest reliability of DISE is good (moderate-substantial according to one framework). Larger studies could encompass a broader population of OSA patients and generate estimates with narrower confidence intervals, however.

DISE as a diagnostic procedure has important logistical drawbacks. There are costs and risks (allergic reaction and airway obstruction) that must be balanced against the benefits of the procedure, and ultimately specific subgroups of patients may benefit most from the procedure.

Although DISE has demonstrated validity compared to a gold standard of polysomnography, [10-13] the ideal fiberoptic evaluation of the airway would occur with natural sleep. Previous researchers have shown that this is cumbersome and problematic, in part due to activation of airway reflexes with instrumentation. DISE requires drug-induced sleep, and the differences from natural sleep have not been elucidated completely. Because there is likely heterogeneity in the anatomical factors that produce airway obstruction in OSA, it is reassuring that patients in this study demonstrated a diversity of obstruction patterns during DISE. Further research can compare upper airway mechanics and physiology during drug-induced and natural sleep.

The final limitation of our study was that both reviewers are experienced sleep surgeons. We examined four types of test-retest reliability and found that the correlation for ratings for two different reviewers was similar to the correlation for different evaluations and exams. The generalizability of the findings can be explored with larger studies that include more reviewers.

Conclusion

The test-retest reliability of DISE is good, particularly in the evaluation of the hypopharyngeal airway. This is particular important for an evaluation that is inherently subjective, not inexpensive, and technique-sensitive such as DISE.

Acknowledgments

References

1. Sher AE, Schechtman KB, Piccirillo JF. The efficacy of surgical modifications of the upper airway in adults with obstructive sleep apnea syndrome. Sleep 1996;19:156–77. [PubMed: 8855039]

2. Friedman M, Ibrahim H, Bass L. Clinical staging for sleep-disordered breathing. Otolaryngol Head Neck Surg 2002;127:13–21. [PubMed: 12161725]

3. Kezirian EJ, Goldberg AN. Hypopharyngeal surgery in obstructive sleep apnea: an evidence-based medicine review. Arch Otolaryngol Head Neck Surg 2006;132:1–8.

4. Sundaram S, Bridgman SA, Lim J, et al. Surgery for obstructive sleep apnoea. Cochrane Database Syst Rev 2005:CD001004. [PubMed: 16235277]

5. Sher AE, Thorpy MJ, Shprintzen RJ, et al. Predictive value of Muller maneuver in selection of patients for uvulopalatopharyngoplasty. Laryngoscope 1985;95:1483–7. [PubMed: 4068867]

6. Petri N, Suadicani P, Wildschiodtz G, et al. Predictive value of Muller maneuver, cephalometry and clinical features for the outcome of uvulopalatopharyngoplasty. Evaluation of predictive factors using discriminate analysis in 30 sleep apnea patients. Acta Otolaryngol 1994;114:565–71. [PubMed: 7825443]

7. Doghramji K, Jabourian ZH, Pilla M, et al. Predictors of outcome for uvulopalatopharyngoplasty. Laryngoscope 1995;105:311–4. [PubMed: 7877422]

8. Croft CB, Pringle M. Sleep nasendoscopy: a technique of assessment in snoring and obstructive sleep apnoea. Clin Otolaryngol Allied Sci 1991;16:504–9. [PubMed: 1742903]

9. Kezirian EJ. Drug-induced sleep endoscopy. Op Tech Otolaryngol 2006;17:230–2.

10. Marais J. The value of sedation nasendoscopy: a comparison between snoring and non-snoring patients. Clin Otolaryngol Allied Sci 1998;23:74–6. [PubMed: 9563671]

11. Sadaoka T, Kakitsuba N, Fujiwara Y, et al. The value of sleep nasendoscopy in the evaluation of patients with suspected sleep-related breathing disorders. Clin Otolaryngol Allied Sci 1996;21:485–9. [PubMed: 9118566]

12. Steinhart H, Kuhn-Lohmann J, Gewalt K, et al. Upper airway collapsibility in habitual snorers and sleep apneics: evaluation with drug-induced sleep endoscopy. Acta Otolaryngol 2000;120:990–4. [PubMed: 11200597]

13. Berry S, Roblin G, Williams A, et al. Validity of sleep nasendoscopy in the investigation of sleep related breathing disorders. Laryngoscope 2005;115:538–40. [PubMed: 15744173]

14. Hohenhorst, W. Personal communication. San Francisco, CA: 2005.

15. Kezirian, EJ.; Goldberg, AN. Propofol sleep endoscopy for evaluation in obstructive sleep apnea; AAO-HNS Annual Meeting; Los Angeles, CA. 2005.

16. Browne WJ, Subramanian SV, Jones K, et al. Variance partitioning in multilevel logistic models that exhibit overdispersion. J R Statist Soc A 2005;168:599–613.

17. Davison, AC.; Hinkley, DV. Bootstrap methods and their application. Cambridge University Press; 1997.

18. Oehlert GW. A note on the delta method. Am Stat 1992;46:27–9.

19. Rabe-Hesketh S, Skrondal A, Pickles A. Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. J Econometrics 2005;128:301–23.

20. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74. [PubMed: 843571]

Table 1

Summary of DISE examination findings.

	DISE Exam 1		DISE Exam 2	
	Reviewer 1 (Unblinded)	Reviewer 2 (Blinded)	Reviewer 1 (Unblinded)	Reviewer 2 (Blinded)
	# (%)	# (%)	# (%)	# (%)
Analysis I. Global assessment of level(s) of obstruction				
Palate	60 (94)	60 (94)	59 (92)	61 (95)
Hypopharynx	52 (81)	52 (81)	56 (88)	54 (84)
Analysis II. Degree of obstruction				
Palate				
<50%	3 (5)	4 (6)	5 (8)	4 (6)
50-75%	10 (16)	6 (9)	11 (17)	3 (5)
>75%	51 (80)	54 (84)	48 (75)	57 (89)
Hypopharynx				
<50%	11 (17)	9 (14)	7 (11)	7 (11)
50-75%	18 (28)	22 (34)	22 (34)	21 (33)
>75%	35 (55)	33 (52)	35 (55)	36 (56)
Analysis III. Specific structures.				
Primary structure contributing to obstruction at level of the palate				
None	4 (6)	4 (6)	5 (8)	3 (5)
Palate	35 (55)	36 (56)	28 (44)	30 (47)
Tonsils[*]	24 (38)	24 (38)	28 (44)	31 (48)
LPW—Velopharynx	1 (2)	0 (0)	3 (5)	1 (2)
Primary structure contributing to obstruction at the level of the hypopharynx				
None	11 (14)	10 (16)	7 (9)	7 (12)
Tongue	36 (56)	35 (55)	37 (52)	35 (55)
Epiglottis	10 (19)	11 (17)	12 (31)	11 (19)
LPW—Hypopharynx	7 (11)	8 (12)	8 (8)	9 (14)
Specific structures contributing to obstruction[**]				
Palate	58 (91)	56 (88)	58 (91)	56 (88)
Tonsils[*]	38 (90)	40 (95)	38 (90)	38 (90)
LPW—Velopharynx	29 (45)	41 (64)	33 (52)	36 (56)
Tongue	47 (73)	44 (69)	47 (73)	46 (72)
Epiglottis	27 (42)	26 (41)	32 (50)	29 (45)
LPW—Hypopharynx	9 (14)	11 (17)	10 (16)	16 (25)

[*] Eleven patients had undergone previous tonsillectomy.

[**] Each structure is considered separately, with percentages expressed as a fraction of total DISE evaluations. Percentages sum to greater than 100% because it is possible for a patient to have more than one structure contributing to airway obstruction.

Table 2

Intraclass correlation coefficients analogs.

	ICC Evaluation	ICC Reviewer	ICC Exam	ICC Reviewer-Exam
Analysis I. Global assessment of level(s) of obstruction				
Palate	0.89 (0.61, 1.0)	0.89 (0.61, 1.0)	0.89 (0.61, 1.0)	0.89 (0.61, 1.0)
Hypopharynx	0.84 (0.72, 1.0)	0.84 (0.72, 1.0)	0.84 (0.72, 1.0)	0.84 (0.72, 1.0)
Analysis II. Degree of obstruction				
Palate	0.74 (0.69, 0.91)	0.60 (0.42, 0.89)	0.62 (0.30, 0.94)	0.48 (0.14, 0.82)
Hypopharynx	0.77 (0.68, 0.86)	0.77 (0.62, 0.92)	0.68 (0.57, 0.89)	0.66 (0.51, 0.87)
Analysis III. Specific structures.				
Primary structure contributing to obstruction at level of the palate				
Primary structure at palate	Unblinded 0.63 (0.55, 0.78) Blinded 0.64 (0.57, 0.79)	0.60 (0.51, 0.86)	Unblinded 0.77 (0.58, 0.91) Blinded 0.76 (0.55, 0.88)	0.57 (0.44, 0.76)
Primary structure at hypopharynx	Unblinded 0.68 (0.44, 0.93) Blinded 0.69 (0.47, 0.99)	0.80 (0.55, 1.0)	Unblinded 0.63 (0.41, 1.00) Blinded 0.56 (0.40, 1.00)	0.60 (0.45, 0.84)
Specific structures contributing to obstruction				
Palate	0.68 (0.58, 0.97)	0.54 (0.30, 1.00)	0.68 (0.58, 0.97)	0.54 (0.30, 1.00)
Tonsils[†]	0.98	0.96	0.96	0.93
LPW— Velopharynx	0.57 (0.35, 0.80)	0.51 (0.33, 0.70)	0.47 (0.13, 0.80)	0.41 (0.18, 0.63)
Tongue	0.66 (0.62, 1.00)	0.66 (0.62, 1.00)	0.59 (0.49, 0.88)	0.57 (0.47, 0.88)
Epiglottis	0.74 (0.70, 0.99)	0.79 (0.67, 1.00)	0.73 (0.56, 0.88)	0.67 (0.50, 1.00)

	ICC Evaluation	ICC Reviewer	ICC Exam	ICC Reviewer-Exam
LPW— Hypopharynx	0.78 (0.63, 1.00)	0.78 (0.63, 1.00)	0.78 (0.63, 1.00)	0.74 (0.53, 1.00)

[†] It was not possible to calculate ICC analogs for the assessment of tonsillar contribution to airway obstruction due to the skewed underlying distribution. As such, the raw agreement is shown for the various comparisons.