Published in final edited form as:

Laryngoscope. 2017 February ; 127(2): 411–416. doi:10.1002/lary.26034.

The Effect of Anchors on Reliability of Endoscopic Tremor Ratings

Soren Y. Lowell, PhD, CCC-SLP, **Richard T. Kelley, MD**, **Lauren Busekroos, MD**, **Ramani Voleti, MS, CCC-SLP**, **Carly J. Hosbach-Cannon, MS, CCC-SLP**, **Raymond H. Colton, PhD**, and **Dragos Mihaila, MD**

Department of Communication Sciences and Disorders (S.Y.L., R.V., C.J.H.-C., R.H.C.), Syracuse University, Syracuse, New York; Department of Otolaryngology & Communication Sciences (R.T.K., L.B.), SUNY Upstate Medical University, Syracuse, New York; and the Department of Neurology (D.M.), SUNY Upstate Medical University, Syracuse, New York, U.S.A

Abstract

Objectives/Hypothesis—The purpose of this study was to determine the effects of anchors and training on intrarater and inter-rater reliability for visual-perceptual, endoscopic tremor ratings.

Study Design—Prospective cohort study.

Methods—Nasoendoscopy recordings of 10 participants with a diagnosis of essential voice tremor were evaluated by five voice specialists using the Vocal Tremor Scoring System. Ratings were performed before, immediately after, and 4 weeks after implementation of a training program with anchor stimuli. Immediate and long-term post-training ratings were performed with simultaneous use of anchor samples for each rating.

Results—Intrarater reliability showed significant improvement from pretraining to immediate and long-term post-training. Mean correlation coefficients (Spearman's rho) increased from 0.71 at pretraining to 0.84 and 0.90 at immediate and long-term post-training, respectively. Inter-rater reliability was not affected by training with anchors, with mean correlation coefficients ranging from 0.62 at pretraining to 0.58 and 0.64 at immediate and long-term post-training, respectively.

Conclusions—Consistent, reproducible ratings are critical for the interpretation and comparison of endoscopic tremor data. Reliability findings from this study indicate that the use of anchor samples as referents for making ordinal judgments about the severity of tremor in oropharyngeal and laryngeal regions was helpful for improving internal standards and consistency but less useful for calibrating across different raters.

Keywords

Reliability; endoscopic ratings; laryngoscopy; tremor; voice

Send correspondence to Soren Y. Lowell, 621 Skytop Road, Suite 1200, Department of Communication Sciences & Disorders, Syracuse University, Syracuse, NY 13244. slowell@syr.edu.

Level of Evidence: 4

The authors have no other funding, financial relationships, or conflicts of interest to disclose.

INTRODUCTION

Essential voice tremor (EVT) is estimated to affect between 5 and 10 million people in the United States alone.[1] Up to 31% of patients diagnosed with overall essential tremor may present with voice tremor,[2] and prevalence rates can underestimate actual occurrence due to mild cases often not being reported.[3] In EVT, centrally generated oscillations affect multiple speech-related muscles, including intrinsic and extrinsic laryngeal muscles,[4,5] oropharyngeal muscles,[3,6] and respiratory muscles.[7] Endoscopic determination of task-specific tremor effects on oral, pharyngeal, and laryngeal regions is considered critical for assessing severity and response to treatment.[3] Whereas several validated tools are available for the visual-perceptual assessment of limb tremor,[8–10] the Vocal Tremor Scoring System (VTSS)[6] is the only standardized and validated rating scale for voice tremor. The VTSS implements a 0 to 3 scale to assess the degree of tremor in the palate, base of tongue, pharyngeal walls, global larynx, supraglottis, and true vocal folds.[6]

Consistent and reproducible ratings are fundamental to the use of laryngeal endoscopic data derived from the VTSS for diagnostic assessment and treatment outcomes, yet reliability for endoscopic ratings can be problematic.[11] Dynamic laryngeal parameters, such as vibratory amplitude,[12] glottic closure,[13,14] and phase symmetry,[14,15] often show the lowest reliability. Ratings of action-based tremor are likewise dynamic and can show insufficient reliability.[16] In endoscopic tremor ratings using the VTSS, several challenges may affect reliability. More than 90% of patients with EVT will show tremor in multiple oropharyngeal and laryngeal regions.[3] The nasoendoscopy recording typically shows several regions simultaneously that are variably affected in severity, making it difficult to differentially score each region. In the validation study for the VTSS,[6] high intra-rater reliability with more variable inter-rater reliability levels between individual raters and the expert consensus ratings were reported. However, five expert raters were included who all had contributed to the development of the VTSS and had extensive exposure to endoscopy tremor recordings, providing a level of familiarity with the tool that most other researchers and clinicians lack.

Methods for improving the reliability of endoscopy ratings, particularly for dynamic parameters that show poor reliability, have been minimally investigated. Multiple studies have shown that auditory-perceptual rating reliability can be substantially improved when external comparison stimuli (anchors) are provided as referents and training is implemented prior to ratings being performed.[17–20] Factors that affect reliability of auditory-perceptual ratings, such as a lack of stable internal standards within the rater, rater experience, and drift in severity standards over time[21–23] may also affect visual-perceptual ratings. Therefore, a visual-perceptual rater training program that includes anchors may improve the consistency and agreement of endoscopic ratings. The purpose of this study was to determine the effects of anchors and training on intrarater and inter-rater reliability for endoscopic tremor ratings.

MATERIALS AND METHODS

Participants

Endoscopic recordings were collected from 10 participants with EVT who provided informed consent and were paid for their participation. All participants were diagnosed with

EVT by an otolaryngologist or neurologist, did not carry a diagnosis of spasmodic dysphonia, had no history of laryngeal surgery, showed acoustically measured frequency or amplitude tremor and auditory-perceptual tremor characteristics, were not receiving voice therapy, had no Botox injection for ≥6 months, and were not taking any tremor-reducing pharmaceuticals. The sample size in this study, though small, is similar to other studies addressing the EVT population.[6,24,25]

Endoscopy raters for this study included two otolaryngologists and three certified speech-language pathologists (SLPs) (two MS and one PhD level), all with a specialty interest and experience in voice disorders. Among the otolaryngologists, one was board certified in otolaryngology and the other in her final year of otolaryngology residency. Years of clinical experience in managing voice disorders ranged from 3 to 25 years. All raters had prior experience in evaluating laryngoscopic examinations, although level of experience in rating tremor varied across raters.

Nasoendoscopic Recordings

The nasoendoscopy examinations were performed by a laryngologist during a comprehensive initial voice assessment. For the VTSS, visualization of all oropharyngeal and laryngeal regions was elicited during sustained /i/ phonation. Additional tasks were included to assess abductory/adductory movements and resting status. Experimental samples collected for this study represent a subset of an ongoing long-term study addressing a different study purpose.

For the anchor samples and training samples, deidentified endoscopy recordings from a clinical database were used, which included tremor at various severity levels. From a large initial pool of recordings, samples were selected to represent the 0 to 3 VTSS severity levels for each anatomic region. A laryngologist and SLP initially reviewed and selected several samples for each region for which they agreed on the VTSS severity score. Next, a laryngologist, SLP, and speech scientist reviewed and selected the anchor samples for each severity level and region if all three investigators agreed on the score and felt that it adequately represented the target region. This generated four anchor samples for each of the anatomic regions except the palate, which had fewer available recordings with adequate representation and was limited to an anchor representing severity ratings of 0 and 3. Finally, additional practice and test training samples that contained adequate representation of all anatomic regions were chosen from the initial database.

To avoid any confounding effects, audio signals were removed from all experimental, anchor, and training samples. Video samples were edited to include only periods of sustained /i/ phonation. For the anchor samples only, when the edited sample was short, looping was used to repeat the sample within the anchor recording. The mean durations of video samples were 12.2 seconds for the anchor samples and 55.5 seconds for the experimental samples.

Visual-Perceptual Endoscopy Ratings

Five raters used the VTSS to score each anatomical region for the experimental samples. Recorded samples were duplicated and randomized so that each rater performed ratings of

20 recordings. These ratings were performed individually at three time points: prior to the training session and use of anchors (pretraining), within 48 hours after the training session was completed (post-training), and approximately 4 weeks after the training session was completed (long-term post-training). For the pretraining ratings, raters were provided with the verbatim written definitions from the VTSS, which defined regions and the substructures included.[6]

Raters performed their ratings in a quiet environment using a standardized screen display size for images. Randomized samples were numbered 1 through 20, and raters performed the ratings sequentially. Raters were blinded to all patient characteristics when performing ratings and blinded to repeat status of videos. Raters were instructed to rate each anatomic region/parameter in isolation, and to rate the most severe period represented for each region. If a parameter included more than one structure (e.g., supraglottis region included the epiglottis, false vocal cords, ventricles, aryepiglottic folds, and supraglottic portion of the arytenoids), raters were instructed to rate the structure within the parameter that was most severely affected. For the post-training and long-term post-training ratings, raters first reviewed the written definitions and anchor samples (representing four different severity levels) for each region and then performed the experimental ratings. Raters were instructed to review at least two anchor stimuli for each parameter that they rated on each experimental sample.

Anchor Training

A 2-hour training session was held after the pretraining ratings were completed. For each anatomic region/parameter, anchor samples were reviewed and discussed sequentially after initial demonstration of the 0 and 3 extremes. Next, several training samples were reviewed and discussed until a consensus score was reached. Two or more anchor samples were reviewed for each training sample and region to demonstrate the use and applicability of the anchors. Finally, raters individually rated four training test samples, subsequent scoring was compared, and discrepancies were discussed.

Statistical Analyses

SPSS version 19.0 software (IBM, Armonk, NY) was used for all statistical analysis. Normality testing (Shapiro-Wilk) for the distributions of the raw data across all time points, regions, and raters showed non-normal distributions for 89% of the data. Therefore, nonparametric Spearman’s rho correlation coefficients were used for all subsequent reliability assessments rather than parametric statistics such as intraclass correlation coefficients.

To determine intrarater reliability, Spearman’s rho correlation coefficients were first computed for each rater based on their ratings of the 10 participants for each of the six anatomic regions and the total score. Due to the non-normal distributions of the intrarater correlation coefficients at each of the three time points, a Friedman’s analysis of variance (ANOVA) was used to test for overall differences in intrarater reliability across the three related samples for the six rated regions. Subsequent Wilcoxon signed ranks tests were used to determine which time point comparisons were significantly different. An alpha level of .

025 (Bonferroni correction) was considered significant for the follow-up Wilcoxon (paired sample) signed ranks tests. To test for statistically significant differences in intrarater percent agreement between the pretraining, post-training, and long-term post-training time points, related-samples Cochran's Q tests were used due to the bivariate nature of these data. To determine inter-rater reliability, Spearman's rho correlation coefficients were first computed for each possible rater combination (10 combinations for the five raters), for each region, and the total score. Due to the non-normal distributions of the inter-rater correlation coefficients at each of the three time points, a Friedman's ANOVA was performed on the correlation coefficients to determine differences across the three related samples for the six rated regions.

RESULTS

Intrarater Reliability

Descriptive summary statistics for the overall intra-rater correlation coefficients for the five raters, across the six VTSS regions were as follows: 1) pretraining mean = 0.705, standard deviation (SD) = 0.25; 2) post-training mean = 0.840, SD = 0.15; and 3) long-term post-training mean = 0.899, SD = 0.08. Mean intrarater correlations across the five raters are shown in Table I for each of the six anatomic regions and the total score. Mean intrarater correlations for each rater at the three time points are shown in Table II.

The Friedman's ANOVA showed a significant omnibus difference across the three time points (χ^2 =15.34, df = 2, P < .001 for two-tailed exact significance). Follow-up Wilcoxon signed rank tests showed that reliability at immediate post-training (z = 2.67, P = .007 for two-tailed exact significance) and at long-term post-training (z = 3.75, P < .001 for two-tailed exact significance) were significantly greater than pretraining reliability.

Percent of exact intrarater agreement was also computed for each of the three time points across each of the six anatomic region scores, with summary data provided in Table III. Overall, across all raters and anatomic regions, there was exact intrarater agreement for 63.9% of pretraining ratings, 84.4% of post-training ratings, and 88.1% of long-term post-training ratings. There were five missing data points for intrarater pretraining reliability that were not included in the percent agreement computations, which were items that were missed/not scored by raters. The related-samples Cochran's Q tests (Table III) showed a significant increase in agreement from pretraining to post-training or long-term post-training for the regions of base of tongue, pharyngeal walls, larynx (global), and the true vocal folds.

Inter-rater Reliability

Descriptive summary statistics for the overall inter-rater correlation coefficients across the six VTSS regions were as follows: 1) pretraining mean = 0.623, SD = 0.20; 2) post-training mean = 0.582, SD = 0.18; and 3) long-term post-training mean = 0.639, SD = 0.18. Mean inter-rater correlations across the five raters are shown in Table IV for each of the six regions and the total score. Means represent the average Spearman's rho correlation coefficients computed for every combination of the five raters. The Friedman's ANOVA showed no

significant difference across the three time points ($\chi^2 = 2.43$, $df = 2$, $P = .297$ for exact significance).

DISCUSSION

Despite the importance of reliability for interpreting and comparing research and clinical data, few studies report on laryngeal endoscopic reliability or include procedures necessary for reducing bias and establishing integrity of ratings.[11] The current study investigated whether the use of anchors and training improved the reliability of tremor ratings using a validated tremor rating scale, while including multiple procedures to maximize the integrity and interpretability of our reliability results. Video recordings were randomized prior to blinded review, with 100% of samples duplicated for intrarater reliability, and removal of the audio track to avoid possible bias. None of the experimental samples were included in the training or anchor video samples. Rating parameters were defined with additional directions provided to maximize consistency and specificity of ratings. Raters with a range in laryngoscopic and tremor assessment experience were included who did not have extensive experience with the VTSS, allowing the determination of training effects for raters who might represent typical voice labs or clinic environments.

Applying guidelines for interpreting strength of rater reliability proposed by Landis and Koch[26] to the present study, mean intrarater reliability was substantial at pretraining but improved to a level of almost perfect after training. The two regions with lowest intrarater reliability before training were the true vocal folds and pharyngeal walls, although by 4 weeks post-training these regions reached reliability levels of ≥0.90. Increases in mean correlation coefficients of 0.14 and 0.19 from pretraining to immediate and long-term post-training, along with a notable 20% to 24% increase in percent exact agreement, indicate that the anchors substantially improved the internal consistency of raters. The continued improvement evidenced in the long-term post-training ratings for four of the five raters suggests that the intrarater benefit from the use of anchors may increase with time and anchor use.

No benefit of training and anchors was evidenced for reliability between the five raters for the mean correlation coefficients across all VTSS regions. Mean inter-rater reliability was substantial at pretraining and long-term post-training. The two regions with the lowest inter-rater reliability at both post-training time points were the palate and true vocal folds. Leaving these two regions out of the computation, mean inter-rater reliability at long-term post-training reached a level of 0.70, in comparison to the pretraining mean of 0.64 for these four regions. Of note, although the total score inter-rater reliability was between 0.76 and 0.89 across all time points, this did not accurately reflect the reliability of the individual regions. Apparently, raters were consistent between each other on overall tremor severity being higher or lower for the subject samples, but disagreed on which specific regions reflected that severity.

Consistent with our findings, other studies that have reported on both intra- and inter-rater laryngoscopic reliability generally show higher intrarater reliability as compared to inter-rater reliability.[12,27,28] Even when implementing 1 hour of training prior to experimental

laryngostroboscopic ratings, Steward et al.[28] achieved high intrarater reliability (0.78) but poor inter-rater reliability (0.17). Several factors may have contributed to the lack of significant improvement in inter-rater reliability with training and anchors in the present study. Longer exam durations are needed for nasoendoscopic tremor assessments to fully visualize all VTSS regions during phonation, with mean duration of about 1 minute in this study. Over the length of each sample, variation in severity of one or more regions and frequent severity discrepancies between the left and ride sides made it difficult for raters to select a single severity level for each region. Raters may have been able to use the anchors to set an internal standard for gauging experimental sample severity, but this would not necessarily produce similar gauging between raters. These gauging differences may be more likely to occur for certain regions such as the palate and true vocal folds.

Results of some prior studies addressing laryngoscopic ratings suggest that provision of a normal, calibrating video prior to ratings,[12] use of standardized parameter definitions,[29] and multiple group discussions of practice ratings[29] may help some raters achieve higher inter-rater reliability. Louis and colleagues[30] developed a training videotape to increase reliability of limb/body tremor ratings on a validated tremor scale, and achieved excellent levels of inter-rater reliability after its implementation. To improve reliability in the training development phase, these researchers modified the referent samples based on initial scoring discrepancies, and included more practice items for score ranges which they found to be the hardest to distinguish. Refinement of anchor samples based on initial practice scoring, more time for self-assessment and feedback, and greater representation of hard-to-distinguish scores during practice items might be effective methods for improving laryngoscopic inter-rater reliability in future studies involving training.

CONCLUSION

This study investigated the effects of a training program with anchor video samples on reliability of visual-perceptual, nasoendoscopic ratings of voice tremor using the VTSS. The anchor training significantly improved intrarater reliability and exact agreement when comparing pretraining to immediate and long-term post-training ratings. In contrast, inter-rater reliability did not improve with the use of anchors and training. Our findings suggest that the use of anchor samples as referents for making ordinal judgments about the severity of tremor in oropharyngeal and laryngeal regions is helpful for improving internal standards and consistency but less useful for calibrating across different raters. These findings indicate that with refinement of the training procedures and anchor stimuli, anchor techniques show promise as methods for improving reliability of laryngoscopic ratings.

Acknowledgments

Funding for this study was provided by the National Institute on Deafness and Other Communication Disorders, National Institutes of Health, R03DC012429.

BIBLIOGRAPHY

1. Pahwa R, Lyons KE. Essential tremor: differential diagnosis and current therapy. Am J Med. 2003; 115:134–142. [PubMed: 12893400]

2. Diaz NL, Louis ED. Survey of medication usage patterns among essential tremor patients: movement disorder specialists vs.general neurologists. Parkinsonism Relat Disord. 2010; 16:604–607. [PubMed: 20691629]

3. Sulica L, Louis ED. Clinical characteristics of essential voice tremor: a study of 34 cases. Laryngoscope. 2010; 120:516–528. [PubMed: 20066728]

4. Koda J, Ludlow CL. An evaluation of laryngeal muscle activation in patients with voice tremor. Otolaryngol Head Neck Surg. 1992; 107:684–696. [PubMed: 1437206]

5. Hertegard S, Granqvist S, Lindestad PA. Botulinum toxin injections for essential voice tremor. Ann Otol Rhinol Laryngol. 2000; 109:204–209. [PubMed: 10685574]

6. Bove M, Daamen N, Rosen C, Wang CC, Sulica L, Gartner-Schmidt J. Development and validation of the vocal tremor scoring system. Laryngoscope. 2006; 116:1662–1667. [PubMed: 16955000]

7. Tomoda H, Shibasaki H, Kuroda Y, Shin T. Voice tremor: dysregulation of voluntary expiratory muscles. Neurology. 1987; 37:117–122. [PubMed: 3796827]

8. Louis ED, Wendt KJ, Albert SM, Pullman SL, Yu Q, Andrews H. Validity of a performance-based test of function in essential tremor. Arch Neurol. 1999; 56:841–846. [PubMed: 10404986]

9. Fahn, S., Tolosa, ES., Marin, C. Clinical rating scale for tremor. In: Jankovic, J., Tolosa, ES., editors. Parkinson's Disease and Movement Disorders. Baltimore, MD: Williams & Wilkens; 1993. p. 271-280.

10. Elble R, Comella C, Fahn S, et al. Reliability of a new scale for essential tremor. Mov Disord. 2012; 27:1567–1569. [PubMed: 23032792]

11. Bonilha HS, Focht KL, Martin-Harris B. Rater methodology for stroboscopy: a systematic review. J Voice. 2015; 29:101–108. [PubMed: 25261957]

12. Yiu EM, Lau VC, Ma EP, Chan KM, Barrett E. Reliability of laryngostroboscopic evaluation on lesion size and glottal configuration: a revisit. Laryngoscope. 2014; 124:1638–1644. [PubMed: 24222186]

13. Hillel AT, Johns MM III, Hapner ER, Shah M, Wise JC, Klein AM. Voice outcomes from subligamentous cordectomy for early glottic cancer. Ann Otol Rhinol Laryngol. 2013; 122:190–196. [PubMed: 23577572]

14. Nawka T, Konerding U. The interrater reliability of stroboscopy evaluations. J Voice. 2012; 26:812.e811–e810.

15. van Gogh CD, Verdonck-de Leeuw IM, Boon-Kamma BA, et al. The efficacy of voice therapy in patients after treatment for early glottic carcinoma. Cancer. 2006; 106:95–105. [PubMed: 16323175]

16. Martinez-Martin P, Gil-Nagel A, Gracia LM, Gomez JB, Martinez-Sarries J, Bermejo F. Unified Parkinson's Disease Rating Scale characteristics and structure. The Cooperative Multicentric Group. Mov Disord. 1994; 9:76–83. [PubMed: 8139608]

17. Eadie TL, Kapsner-Smith M. The effect of listener experience and anchors on judgments of dysphonia. J Speech Lang Hear Res. 2011; 54:430–447. [PubMed: 20884782]

18. Gerratt BR, Kreiman J, Antonanzas-Barroso N, Berke GS. Comparing internal and external standards in voice quality judgments. J Speech Hear Res. 1993; 36:14–20. [PubMed: 8450655]

19. Awan SN, Lawson LL. The effect of anchor modality on the reliability of vocal severity ratings. J Voice. 2009; 23:341–352. [PubMed: 18346869]

20. Chan KM, Yiu EM. The effect of anchors and training on the reliability of perceptual voice evaluation. J Speech Lang Hear Res. 2002; 45:111–126. [PubMed: 14748643]

21. Kreiman J, Gerratt BR, Ito M. When and why listeners disagree in voice quality assessment tasks. J Acoust Soc Am. 2007; 122:2354–2364. [PubMed: 17902870]

22. Kreiman J, Gerratt BR. Validity of rating scale measures of voice quality. J Acoust Soc Am. 1998; 104:1598–1608. [PubMed: 9745743]

23. Kreiman J, Gerratt BR, Precoda K. Listener experience and perception of voice quality. J Speech Hear Res. 1990; 33:103–115. [PubMed: 2314068]

24. Warrick P, Dromey C, Irish JC, Durkin L, Pakiam A, Lang A. Botulinum toxin for essential tremor of the voice with multiple anatomical sites of tremor: a crossover design study of unilateral versus bilateral injection. Laryngoscope. 2000; 110:1366–1374. [PubMed: 10942143]

25. Adler CH, Bansberg SF, Hentz JG, et al. Botulinum toxin type A for treating voice tremor. Arch Neurol. 2004; 61:1416–1420. [PubMed: 15364688]

26. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977; 33:159–174. [PubMed: 843571]

27. Rosow DE, Sulica L. Laryngoscopy of vocal fold paralysis: evaluation of consistency of clinical findings. Laryngoscope. 2010; 120:1376–1382. [PubMed: 20564722]

28. Steward DL, Wilson KM, Kelly DH, et al. Proton pump inhibitor therapy for chronic laryngopharyngitis: a randomized placebo-control trial. Otolaryngol Head Neck Surg. 2004; 131:342–350. [PubMed: 15467597]

29. Smith ME, Ramig LO, Dromey C, Perez KS, Samandari R. Intensive voice treatment in Parkinson disease: laryngostroboscopic findings. J Voice. 1995; 9:453–459. [PubMed: 8574314]

30. Louis ED, Barnes L, Wendt KJ, et al. A teaching videotape for the assessment of essential tremor. Mov Disord. 2001; 16:89–93. [PubMed: 11215599]

TABLE I

Mean Intrarater Spearman's Rho Correlation Coefficients and Standard Deviations for Each of the Six Regions Rated in the Vocal Tremor Scoring System at the Three Rating Time Points.

	Intrarater Reliability		
VTSS Region/Parameter Rated	**Pretraining SRCC**	**Post-training SRCC**	**Long-term Post-training SRCC**
Palate	0.716 (0.31)	0.794 (0.08)	0.895 (0.10)
Base of tongue	0.737 (0.19)	0.713 (0.29)	0.871 (0.08)
Pharyngeal walls	0.684 (0.35)	0.881 (0.05)	0.918 (0.06)
Larynx (global)	0.768 (0.12)	0.897 (0.10)	0.870 (0.07)
Supraglottis	0.778 (0.12)	0.834 (0.16)	0.888 (0.11)
True vocal folds	0.548 (0.34)	0.923 (0.06)	0.952 (0.06)
Total score	0.790 (0.16)	0.929 (0.04)	0.954 (0.02)

SRCC = Spearman's rho correlation coefficients; VTSS = Vocal Tremor Scoring System.

TABLE II

Mean Intrarater Spearman's Rho Correlation Coefficients and Standard Deviations for Each of the Five Raters at the Three Rating Time Points Including the Six Vocal Tremor Scoring System Regions.

	Intrarater Reliability by Rater			
Rater	**Pretraining SRCC**	**Post-training SRCC**	**Long-term Post-training SRCC**	**Mean for All Time Points**
Rater 1	0.777 (0.18)	0.713 (0.22)	0.936 (0.08)	0.809 (0.11)
Rater 2	0.870 (0.05)	0.905 (0.10)	0.826 (0.05)	0.867 (0.04)
Rater 3	0.858 (0.07)	0.886 (0.08)	0.918 (0.05)	0.887 (0.03)
Rater 4	0.607 (0.21)	0.915 (0.08)	0.938 (0.07)	0.820 (0.18)
Rater 5	0.414 (0.31)	0.782 (0.18)	0.878 (0.09)	0.691 (0.24)

SRCC = Spearman's rho correlation coefficients.

TABLE III

Mean Percent Agreement for Each of the Six Regions Rated in the Vocal Tremor Scoring System at the Three Rating Time Points and the Cochran's Q Statistical Results for Differences Across the Three Time Points.

	Intrarater Percent Agreement (Exact)			
VTSS Region/Parameter Rated	**Pretraining %**	**Post-training %**	**Long-term Post-training %**	**Cochran's Q Test for Significant Difference Across Three Time Points**
Palate	68.9	82.2	84.4	0.167
Base of tongue	67.3	84.0	89.8	0.008*
Pharyngeal walls	64.6	82.0	90.0	0.008*
Larynx (global)	57.4	86.0	86.0	0.002*
Supraglottis	70.0	84.0	86.0	0.102
True vocal folds	55.1	88.0	92.0	< 0.001*

VTSS = Vocal Tremor Scoring System.

TABLE IV

Mean Inter-rater Spearman's Rho Correlation Coefficients and Standard Deviations for Each of the Six Regions Rated in the Vocal Tremor Scoring System, at the Three Rating Time Points.

	Inter-rater Reliability		
VTSS Region/Parameter Rated	**Pretraining SRCC**	**Post-training SRCC**	**Long-term Post-training SRCC**
Palate	0.582 (0.32)	0.553 (0.22)	0.477 (0.34)
Base of tongue	0.522 (0.29)	0.465 (0.29)	0.668 (0.14)
Pharyngeal walls	0.578 (0.25)	0.671 (0.09)	0.685 (0.12)
Larynx (global)	0.802 (0.07)	0.691 (0.13)	0.659 (0.15)
Supraglottis	0.643 (0.11)	0.630 (0.16)	0.793 (0.11)
True vocal folds	0.614 (0.14)	0.480 (0.18)	0.550 (0.22)
Total score	0.837 (0.08)	0.759 (0.06)	0.888 (0.07)

SRCC = Spearman's rho correlation coefficients; VTSS = Vocal Tremor Scoring System.