Published in final edited form as:

J Voice. 2015 January ; 29(1): 101–108. doi:10.1016/j.jvoice.2014.06.014.

Rater Methodology for Stroboscopy: A Systematic Review

Heather Shaw Bonilha, Ph.D., CCC-SLP^{1,2}, Kendrea L. Focht, C.Sc.D., CCC-SLP¹, and Bonnie Martin-Harris, Ph.D., CCC-SLP^{1,2}

¹Department of Health Science and Research, Medical University of South Carolina, Charleston, SC

²Department of Otolaryngology - Head and Neck Surgery, Medical University of South Carolina, South Carolina

SUMMARY

Objectives—Laryngeal endoscopy with stroboscopy (LES) remains the clinical gold standard for assessing vocal fold function. LES is used to evaluate the efficacy of voice treatments in research studies and clinical practice. LES as a voice treatment outcome tool is only as good as the clinician interpreting the recordings. Research using LES as a treatment outcome measure should be evaluated based on rater methodology and reliability. The purpose of this literature review was to evaluate the rater-related methodology from studies that use stroboscopic findings as voice treatment outcome measures.

Study Design—Systematic Literature Review

Methods—Computerized journal databases were searched for relevant articles using terms: stroboscopy and treatment. Eligible articles were categorized and evaluated for the use of rater-related methodology, reporting of number of raters, types of raters, blinding and rater reliability.

Results—Of the 738 articles reviewed, 80 articles met inclusion criteria. More than one-third of the studies included in the review did not report the number of raters who participated in the study. Eleven studies reported results of rater reliability analysis with only two studies reporting good inter- and intra-rater reliability.

Conclusion—The comparability and use of results from treatment studies that employ LES are limited by a lack of rigor in rater methodology and variable, mostly poor, inter- and intra-rater reliability. To improve our ability to evaluate and use the findings from voice treatment studies that employ LES features as outcome measures, greater consistency of reporting rater methodology characteristics across studies and improved rater reliability is needed.

© 2014 The Voice Foundation. All rights reserved.

Address for Correspondence: Heather Shaw Bonilha, PhD, Department of Health Science and Research, Medical University of South Carolina, 77 President St, Charleston, SC 29425, Tel: (843) 792-2527, bonilhah@muscc.edu.

Portions of this study were presented at the 43rd Symposium of The Voice Foundation: Care of the Professional Voice. Philadelphia, Pennsylvania, May 2014.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Keywords

voice; stroboscopy; reliability; rater

INTRODUCTION

It has been estimated that at any given time between 6.6% to 7.5% people in the United States have a voice disorder.¹ The cost of voice disorders are estimated to be between 0.7 to 4.9 billion dollars.² There are several types of treatments for voice disorders; broad categories of treatments are pharmaceutical, surgical and behavioral. Treatment outcome measures are used to evaluate the efficacy of these treatments in research studies and clinical practice. There are several different types of outcome measures available to evaluate the effectiveness of a treatment for voice disorders including: patient report, perceptual assessment, acoustic analysis, aerodynamic measures, and laryngeal imaging.^{3, 4} This paper will focus on methodology related to using laryngeal imaging, specifically stroboscopy, as a voice treatment outcome measure.

Three studies epitomize the importance of laryngeal endoscopy with stroboscopy (LES). Sataloff et al.⁵ studied the clinical value of LES, beyond that provided by clinical and laryngeal mirror examination, and found that LES added diagnostically-relevant information in 47% of the cases and that clinically significant findings were detected only through LES in 32.4% of patient cases. These results led the authors to conclude that “stroboscopy is invaluable in daily practice and essential for valid, reliable diagnosis of voice disorders”. Similarly, Remacle⁶ found that in 732 patients LES findings were considered useful in 92% of cases. Behrman⁷ found that 94% of speech-language pathologists who treated patients with voice disorders considered LES important for defining overall therapy goals. In addition, they found that 89% considered LES informative for outcomes assessment, and 81% considered LES important for educating patients about voice production. While there are some limitations to LES, such as its reliance on pitch tracking and temporal resolution, it remains invaluable for assessing vocal fold structure and function.

LES as a voice treatment outcome tool is only as good as the clinician interpreting the LES recordings. The interpretation and value of stroboscopic findings are directly linked to training and skills of the operator. Thus, studies using LES as outcomes measures rely heavily on their raters. There are a number of rater-related characteristics that should be considered when using LES as an outcome measure including: number of raters, profession of raters (ENT/SLP), blinding, training, use of randomization, inter-rater reliability and intra-rater reliability. Given the value of LES and the importance of reporting consistent voice treatment outcome measures, we undertook a literature review on these topics. The purpose of this literature review was to evaluate the rater-related methodology from studies that use stroboscopic findings as voice treatment outcome measures.

METHOD

Search Strategy

This review was conducted using the PRISMA statement standards⁸. A search strategy was developed and implemented in three computerized journal databases (Pubmed, Ovid and Cochrane) to identify all English language studies where LES was used as an outcome measure for the treatment of a voice disorder. The following search terms were used: “laryngostroboscopy”, “stroboscopy”, “stroboscopy”, “stroboscopy”, “stroboscopy”, “videostroboscopy”, and “videostroboscopy”. Each of the search terms was combined with “treatment”. Specifically, the Pubmed search was: ((laryngostroboscopy OR stroboscopy OR stroboscopy OR stroboscopy OR videostroboscopy OR videostroboscopy) AND treatment). All studies published from database inception (Pubmed and Ovid electronic 1946, Cochrane 1993) to our last search on November 21, 2013 were reviewed for eligibility. Unpublished reports were not considered for this review. Authors were not contacted.

Inclusion Criteria

One reviewer (either KF or HB) assessed each study based on the following *inclusion* criteria in the following order: English language; original article; human study; perceptual judgments of stroboscopic findings reported both pre- and post-treatment; and aggregate data reported for five or more participants. Duplicate results were deleted. Then, a second reviewer (either KF or HB) assessed each study identified by the first reviewer for inclusion criteria.

Assessment of Evidence

Eligible articles included in this review were categorized and evaluated for the use of rater-related methodology, reporting of number of raters, types of raters, blinding and rater reliability by HB and KF.

Data Synthesis

The analysis was descriptive in nature because the heterogeneity of rater methodologies and data used to report rater reliability precluded a robust statistical analysis (i.e., metaanalysis).

RESULTS

Assessment of Evidence

Of the 738 articles reviewed, 80 articles met inclusion criteria (see Figure 1). Eligible studies are summarized in Table 1.

Number and Type of Raters—More than one-third of the studies included in the review (30/80, 38%) did not report the number of raters who participated in the study. Of the 50 studies that did report the number of raters, 9 (18%) employed 1 rater, 20 (40%) employed 2 raters, 16 (32%) employed 3 raters, 4 (8%) employed 4 raters, and 1 (2%) employed 6 raters. Forty-six of the 80 (58%) papers specified the profession of the raters. The raters were ENTs

(otolaryngologists, phoniatricians and laryngologists) in 32 papers, both ENTs and SLPs in 12 papers, and SLPs only in 2 papers.

Rater Blinding and Recording Randomization—Twenty-five of the 80 studies (31%) reported using raters blinded to treatment status. Sixteen out of the 80 studies (20%) reported randomizing the LES recordings for rating.

Rater Reliability Reporting—Eleven studies (14%) reported results of rater reliability analysis. Six papers reported results for inter- and intra-rater reliability, 4 reported only intra-rater reliability and 1 reported only inter-rater reliability (Table 2). Fifteen of the 80 papers reported the use of consensus rating. One paper, Galletti,⁹ reported highly concordant results between raters but did not report the methods used to test rater reliability or results.

Intra-rater reliability results: In the 10 studies that reported intra-rater reliability, 5 studies reported correlation coefficients, 2 reported exact percent agreement, 2 reported kappa values, and 1 study reported correlation and regression slope values. Correlation coefficients ranged between 0.17 and 0.93. Percent agreement values ranged from 0 to 100, and kappa values ranged from 0.098 to 1.0.

Five of the 10 studies (50%) reported good intra-rater reliability (per study-specific criteria in Table 2). Lam's¹⁰ study included data from 82 patients who underwent LES exams at 4 time points: 1 before, 2 during and 1 after a treatment/placebo period. Lam reported the highest intra-rater reliability at above 0.90, although, this was based solely on one examiner re-rating 8 recordings. Furthermore, the features rated in Lam were from the Reflux Finding Score¹¹, which are all anatomical and stationary. Karpenko¹² reported high intra-rater reliability (89%) for ratings of supraglottic activity, mucosal wave, and glottal competency using a 5-point scale. However, only four LES recordings were rated to obtain intra-rater reliability data in Karpenko's study. Since good intra-rater reliability was found for four consecutive LES recordings, only one rater scored the remaining six recordings. Wang¹³ reported good intra-rater reliability (0.80) from 10% of their recordings re-rated simultaneously by two otolaryngologists using a consensus method. This was the only study that used a consensus scoring approach and reported reliability results. Beaver¹⁴ reported correlation coefficients indicating a good level of agreement, although their data were based on rating elements such as, edema or erythema using a 4-point scale rather than rating vocal fold vibratory features. Beaver also used still images from the LES exam and only one rater to re-assess images. Smith¹⁵ obtained intra-rater reliability data for the following features: glottal configuration, degree of glottal incompetence and laryngeal hyperfunction. Each feature was rated using feature-specific 5-point scales with anatomical descriptions for each point on each scale. Data were not provided for each feature separately, so it is unknown which features obtained better intra-rater reliability.

The remaining five studies reported poor intra-rater reliability for at least one feature or one rater. Steward¹⁶ reported good intra-rater reliability for 3 raters and questionable reliability for one rater. Steward's study uniquely included a one-hour training session for how to use the 5-point Likert scale with scale anchors (present to absent or none to severe) to score the following features: vocal cord edema and erythema; arytenoid edema and erythema;

pachydermia; and mucosal wave. Van Gogh's¹⁷ study reported data from two raters re-assessing 14 features from 10 recordings using a 16-item rating scale adapted from Bless et al.¹⁸. The intra-rater reliability was reported to be moderate to good, although kappa values per feature were not reported. Thus, it is unknown which features had moderate and which had good intra-rater reliability. Phase symmetry was found to have poor intra-rater reliability and was subsequently excluded from analysis. Chang¹⁹ uniquely excluded data that did not meet the requirement of having a correlation coefficient above 0.50 or a regression line slope below 0.75 or above 1.25. Only vertical level ratings met the cut-off for intra-rater reliability, with mucosal wave and glottal closure ratings having the worse intra-rater reliability. Hillel²⁰ reported intra-rater reliability pre- and post-treatment for mucosal wave and glottis closure, judged on a 4-point and 3-point scale, respectively. Intra-rater reliability for mucosal wave ranged from kappa values of 0.63–1.00, while glottis closure kappa values ranged from 0.36–1.00. The study by Morgan and colleagues²¹ used the Stroboscopy Research Instrument²² to re-rate nine recordings. Morgan reported data for each feature in a figure ranging from 0–1 with a mean of 0.55, values given here and in the table are extracted from the figure. Judgments of symmetry, right amplitude and right periodicity were found to have the poorest intra-rater reliability. It should be noted that the data reported here for Morgan reflects the removal of an outlier.

Inter-rater reliability results: Seven studies reported inter-rater reliability, with 3 reporting correlation coefficients, 2 reporting kappa values, 1 reporting exact percent agreement, and one reporting both kappa and agreement. Correlation coefficients reported ranged between 0.17 and 1.0. Percent agreement reported ranged from 0.5 to 0.77, and kappa values reported ranged from 0.09 to 0.90.

Two of the seven studies reported good inter-rater reliability, per study-specific criteria in Table 2. Karpenko¹² reported good inter-rater reliability of 77%. Smith¹⁵ also reported good inter-rater reliability between 0.87 and 1.0. (See previous section on intra-rater reliability for further study details.)

The remaining five studies reported poor inter-rater reliability for at least one LES feature. Beaver¹⁴ reported kappa values ranging from 0.097 – 0.766 with only 16 out of 72 comparisons having significant agreement at a kappa of 0.3. Beaver noted that the features with the highest inter-rater reliability were: presence/absence of leukoplakia, nodules or prenodules and contact granuloma; and severity of pre-treatment vocal fold and subglottis edema. Chang¹⁹ found that only inter-rater reliability data for mucosal wave, not vertical level or glottal closure, was high enough to meet their kappa and agreement cut-offs of below 0.4 and below 0.65, respectively. Dejonckere³ found inter-rater correlation coefficients of 0.68, 0.60, and 0.66 for glottal closure, regularity, and mucosal wave, respectively. Hillel's²⁰ inter-rater reliability data for mucosal wave ranged from kappa values of 0.38–0.90, while glottis closure kappa values ranged from 0.09–0.2. Steward¹⁶ reported poor inter-rater reliability (correlation coefficient 0.17), with the highest correlation between two raters at 0.43.

DISCUSSION

This literature review revealed that the majority of studies that used stroboscopic findings as behavioral, medical or surgical outcome measure(s) did not describe the method used to test rater reliability. Use and reporting of rater methodology is essential for accurately interpreting results of the assessment of laryngeal characteristics from LES. The strength and applicability of the reported results would seemingly be greater for videostroboscopic findings characterized by high versus low agreement between raters. From the 80 articles reviewed using LES as a voice treatment outcome measure, only 7.5% reported both inter- and intra-rater reliability, with only half reporting the number of raters employed.

Only two studies reported good intra- and inter-rater reliability, per study-specific criteria. Overall, ratings of anatomical/stationary features were more reliable than ratings of functional/temporally dependent features, with phase symmetry being the feature with the poorest rater reliability. Due to the reported poor rater reliability,^{22–24} it is necessary that researchers report the rater reliability from their studies so that the treatment results can be accurately interpreted. Moreover, researchers should consider implementing methodology of a priori reliability cut-offs and excluding data that does not meet those criteria similar to Chang et al.¹⁹ Such standards would allow for increased confidence in study results which would, in turn, improve the usefulness of the study for evidence-based practice decisions. Details of the rater training procedures on the specific scales used for the study should also be reported.

Approaches that attempted to improve rater reliability were used in only two studies. In Steward, the use of training appears to have led to good intra-rater results in three out of four raters, but the inter-rater reliability results were poor. Smith's study provided clear anatomical definitions for every level of each scale and provided feature-specific scales, which resulted in good intra- and inter-rater reliability. Smith's results provide insight into the possibility of honing this approach to improve rater reliability. These data are consistent with published studies that have sought to improve rater reliability.^{22–24} Future research should focus on creating a standardized rating system that has unambiguous, standardized definitions for each level of the scale and has feature-specific scales. It is likely that such a system in combination with rater training to improve intra-rater reliability would result in a more reliable method of interpreting vocal fold structure and function from LES.

While the results of this literature review are overall not favorable regarding the reliability of LES, we believe that LES can be a powerful voice treatment outcome measure for research and clinical use. Poor rater reliability has been common in many aspects of medical care that rely on visual judgments²⁵. The voice community can harness the methodology and knowledge of researchers who have improved the reliability of other medical diagnostic tools by creating standardized protocols, developing rater training programs and better understanding the types and reasons for rater errors. This knowledge should encourage us to improve our rater methodology and standardize the LES interpretation to optimally use this important assessment. Currently, there are few published studies with adequate rigor that use LES as an outcome measure²⁶. With these improvements, LES should become a better voice treatment outcome measure that is used and reported with scientific rigor.

There are limitations to this review that possibly influenced the results. The search was limited to articles written in English. There was a lack of robust statistical analysis due to the variability in rater reliability reporting. The search terms may not have detected all relevant studies. This review did not include the many variables that may influence rater reliability including differences in imaging systems. Lastly, an assessment of study bias was not conducted because the rater-specific methodology, not the study result, was the target of this review.

CONCLUSION

LES is commonly used as a voice treatment outcome measure in the peer-reviewed literature and in clinical practice⁷. The comparability and use of results from treatment studies that use LES are limited by a lack of rigor in rater methodology and variable, mostly poor, inter- and intra-rater reliability. To improve our ability to evaluate and use the findings from voice treatment outcome studies that use LES features as outcome measures, greater consistency of reporting rater methodology characteristics across studies is needed. We strongly encourage the development of systematic reporting of LES results and encourage journal authors, reviewers and editors to support and enforce this reporting. Ultimately, standardized, reliable and valid methods to train, interpret and report features of vocal fold structure and function from LES are needed.

Acknowledgments

This publication was supported by the South Carolina Clinical & Translational Research (SCTR) Institute, with an academic home at the Medical University of South Carolina, through NIH Grant Numbers KL2 RR029880 and KL2 TR000060.

References

1. Roy N, Merrill RM, Gray SD, Smith EM. Voice disorders in the general population: Prevalence, risk factors, and occupational impact. *Laryngoscope*. 2005; 115:1988–1995. [PubMed: 16319611]
2. Cohen SM, Kim J, Roy N, Asche C, Courey M. Direct health care costs of laryngeal diseases and disorders. *Laryngoscope*. 2012; 122:1582–1588. [PubMed: 22544473]
3. Dejonckere PH. Clinical implementation of a multidimensional basic protocol for assessing functional results of voice therapy. A preliminary study. *Rev Laryngol Otol Rhinol (Bord)*. 2000; 121:311–313. [PubMed: 11387655]
4. DeJonckere PH, Crevier-Buchman L, Marie JP, Moerman M, Remacle M, Woisard V, et al. Implementation of the european laryngological society (els) basic protocol for assessing voice treatment effect. *Rev Laryngol Otol Rhinol (Bord)*. 2003; 124:279–283. [PubMed: 15144022]
5. Sataloff RT, Spiegel JR, Hawkshaw MJ. Stroboscovideolaryngoscopy: Results and clinical value. *Ann Otol Rhinol Laryngol*. 1991; 100:725–727. [PubMed: 1952664]
6. Remacle M. The contribution of videostroboscopy in daily ENT practice. *Acta Otorhinolaryngol Belg*. 1996; 50:265–281. [PubMed: 9001636]
7. Behrman A. Common practices of voice therapists in the evaluation of patients. *J Voice*. 2005; 19:454–469. [PubMed: 16102671]
8. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: The prisma statement. *BMJ*. 2009; 339:b2535. [PubMed: 19622551]
9. Galletti B, Freni F, Cammaroto G, Catalano N, Gangemi G, Galletti F. Vocal outcome after CO₂ laser cordectomy performed on patients affected by early glottic carcinoma. *J Voice*. 2012; 26:801–805. [PubMed: 22717493]

10. Lam PK, Ng ML, Cheung TK, Wong BY, Tan VP, Fong DY, et al. Rabeprazole is effective in treating laryngopharyngeal reflux in a randomized placebo-controlled trial. *Clin Gastroenterol Hepatol*. 2010; 8:770–776. [PubMed: 20303417]
11. Belafsky PC, Postma GN, Koufman JA. The validity and reliability of the reflux finding score (rfs). *Laryngoscope*. 2001; 111:1313–1317. [PubMed: 11568561]
12. Karpenko AN, Dworkin JP, Meleca RJ, Stachler RJ. Cymetra injection for unilateral vocal fold paralysis. *Ann Otol Rhinol Laryngol*. 2003; 112:927–934. [PubMed: 14653360]
13. Wang CT, Lai MS, Liao LJ, Lo WC, Cheng PW. Transnasal endoscopic steroid injection: A practical and effective alternative treatment for benign vocal fold disorders. *Laryngoscope*. 2013; 123:1464–1468. [PubMed: 23494523]
14. Beaver ME, Stasney CR, Weitzel E, Stewart MG, Donovan DT, Parke RB, et al. Diagnosis of laryngopharyngeal reflux disease with digital imaging. *Otolaryngol Head Neck Surg*. 2003; 128:103–108. [PubMed: 12574767]
15. 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]
16. Steward DL, Wilson KM, Kelly DH, Patil MS, Schwartzbauer HR, Long JD, et al. Proton pump inhibitor therapy for chronic laryngo-pharyngitis: A randomized placebo-control trial. *Otolaryngol Head Neck Surg*. 2004; 131:342–350. [PubMed: 15467597]
17. van Gogh CD, Verdonck-de Leeuw IM, Boon-Kamma BA, Rinkel RN, de Bruin MD, Langendijk JA, et al. The efficacy of voice therapy in patients after treatment for early glottic carcinoma. *Cancer*. 2006; 106:95–105. [PubMed: 16323175]
18. Bless DM, Hirano M, Feder RJ. Videostroboscopic evaluation of the larynx. *Ear Nose Throat J*. 1987; 66:289–296. [PubMed: 3622324]
19. Chang J, Fang TJ, Yung K, van Zante A, Miller T, Al-Jurf S, et al. Clinical and histologic predictors of voice and disease outcome in patients with early glottic cancer. *Laryngoscope*. 2012; 122:2240–2247. [PubMed: 22956472]
20. Hillel AT, Johns MM 3rd, 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]
21. Morgan JE, Zraick RI, Griffin AW, Bowen TL, Johnson FL. Injection versus medialization laryngoplasty for the treatment of unilateral vocal fold paralysis. *Laryngoscope*. 2007; 117:2068–2074. [PubMed: 17828043]
22. Rosen CA. Stroboscopy as a research instrument: Development of a perceptual evaluation tool. *Laryngoscope*. 2005; 115:423–428. [PubMed: 15744150]
23. Poburka BJ, Bless DM. A multi-media, computer-based method for stroboscopy rating training. *J Voice*. 1998; 12:513–526. [PubMed: 9988038]
24. Poburka BJ. A new stroboscopy rating form. *J Voice*. 1999; 13:403–413. [PubMed: 10498056]
25. Milstein, CF. Reliability in the interpretation of laryngeal imaging. In: Leonard, KKaR, editor. *Laryngeal evaluation: Indirect laryngoscopy to high-speed digital imaging*. New York, NY: Thieme Medical Publishers, Inc; 2011.
26. Roy N, Barkmeier-Kraemer J, Eadie T, Sivasankar MP, Mehta D, Paul D, et al. Evidence-based clinical voice assessment: A systematic review. *Am J Speech Lang Pathol*. 2013; 22:212–226. [PubMed: 23184134]

Figure 1.

Table 1

First Author	Year Published	Reported Reliability (yes/no)	Consensus Rating (yes/no)	Number of Raters	Profession of Raters	Blinding Reported ?	Randomization of Recordings Reported?
Allegra	2011	No	No	3	ENTs	Yes	No
Bajaj	2011	No	No	N/R	N/R	No	No
Beaver	2003	Yes	No	3	ENTs	Yes	Yes
Biacabe	2001	No	No	2	ENT & SLP	No	No
Bielanowicz	2000	No	No	N/R	N/R	No	No
Chang	2003	No	Yes	2	ENTs	No	No
Chang	2012	Yes	No	2	ENTs	Yes	No
Cheng	2013	No	No	N/R	N/R	No	No
Chhetri	2011	No	Yes	2	ENT & SLP	No	No
Choi	2012	No	Yes	4	2 ENTs & 2 SLPs	No	No
Crumley	1991	No	No	N/R	N/R	No	No
Dedo	2004	No	No	N/R	N/R	No	No
DeJonckere	2000	Yes	No	2	ENTs	No	No
DeJonckere	2003	No	No	2	ENT & SLP	No	No
Dixon	1999	No	No	1	ENT	Yes	No
Fass	2010	No	No	2	N/R	No	Yes
Finck	2005	No	No	N/R	N/R	No	No
Finck	2010	No	Yes	3	2 ENTs & 1 SLP	No	No
Ford	1986	No	No	N/R	N/R	No	No
Ford	1992	No	No	N/R	N/R	No	No
Ford	1993	No	No	N/R	N/R	No	No
Galletti	2012	No	No	3	ENTs	No	No
Geyer	2010	No	Yes	3	ENTs & SLP	No	No
Halawa	2013	No	No	N/R	N/R	No	No
Hallén	2001	No	No	3	ENT	Yes	Yes
Hillel	2013	Yes	No	2	ENTs	Yes	Yes

First Author	Year Published	Reported Reliability (yes/no)	Consensus Rating (yes/no)	Number of Raters	Profession of Raters	Blinding Reported ?	Randomization of Recordings Reported?
Hirano	2012	No	No	N/R	N/R	No	No
Hwang	2013	No	No	2	ENTs	Yes	No
Iloabachie	2007	No	No	N/R	N/R	No	No
Jensen	2013	No	No	N/R	N/R	No	No
Karpenko	2003	Yes	No	3	2 SLPs & 1 ENT	Yes	No
Keilmann	2011	No	No	3	ENTs	Yes	Yes
Lam	2010	Yes	No	1	ENT	Yes	Yes
Lan	2010	No	Yes	2	ENTs	No	No
Law	2012	No	No	1	ENT	Yes	Yes
Ledda	2006	No	No	N/R	N/R	No	No
Lee	2007	No	Yes	3	2 SLPs & 1 ENT	Yes	Yes
Li	2011	No	No	3	ENTs	Yes	Yes
Maia	2012	No	No	3	SLPs	Yes	No
Maronian	2003	No	No	3	ENTs	Yes	Yes
Mesallam	2011	No	No	N/R	N/R	No	No
Milstein	2005	No	No	3	N/R	Yes	No
Monini	2006	No	No	N/R	N/R	No	No
Mora	2010	No	No	1	ENT	No	No
Morgan	2007	Yes	Yes	2	ENTs	Yes	No
Mortensen	2006	No	No	N/R	N/R	No	No
Nakagawa	2012	No	No	N/R	N/R	No	No
Nam	2013	No	No	N/R	N/R	No	No
Okur	2013	No	No	N/R	N/R	No	No
Paniello	2008	No	No	N/R	N/R	No	No
Pedersen	2004	No	No	N/R	N/R	No	No
Pitman	2012	No	No	N/R	N/R	No	No
Reiter	2012	No	No	1	ENTs	No	No
Rihkanen	2004	No	Yes	6	ENTs	No	No

First Author	Year Published	Reported Reliability (yes/no)	Consensus Rating (yes/no)	Number of Raters	Profession of Raters	Blinding Reported ?	Randomization of Recordings Reported?
Roh	2005	No	No	2	ENTs	Yes	No
Rohde	2012	No	No	1	SLP	No	No
Rontal	2003	No	No	N/R	N/R	No	No
Schindler	2013	No	No	2	ENTs	Yes	No
Schneider (a)	2003	No	No	N/R	N/R	No	No
Schneider (b)	2003	No	No	N/R	N/R	No	No
Shi	2011	No	No	3	ENTs	No	No
Silbergleit	2013	No	No	2	ENT & SLP	Yes	Yes
Smith	1995	Yes	No	4	3 SLPs & 1 ENT	No	Yes
Steward	2004	Yes	No	4	3 ENTs & 1 SLP	Yes	Yes
Storck	2007	No	No	N/R	N/R	No	No
Storck	2010	No	No	4	2 ENTs & 2 SLPs	No	No
Su (a)	2005	No	No	2	N/R	No	No
Su (b)	2005	No	No	2	N/R	Yes	No
Tsunoda	2005	No	No	N/R	N/R	No	No
van Gogh	2006	Yes	Yes	2	ENT	Yes	Yes
Wang (a)	2011	No	Yes	3	ENTs	Yes	Yes
Wang (b)	2011	No	Yes	3	ENTs	Yes	Yes
Wang	2012	No	Yes	2	ENTs	No	No
Wang (a)	2013	Yes	Yes	2	ENTs	No	No
Wang (b)	2013	No	Yes	2	ENTs	No	No
Wiskirka-Woznica	2011	No	No	N/R	N/R	No	No
Woo	2013	No	No	1	ENT	No	No
Yilmaz (a)	2012	No	No	1	ENT	No	No
Yilmaz (b)	2012	No	No	N/R	N/R	No	No
Zhang	2011	No	No	1	ENT	No	No

N/R=Not reported; ENT=laryngologist/otolaryngologist/otorhinolaryngologist/phonomatrician; SLP=Speech-Language Pathologist

Table 2

Studies reporting rater reliability

First Author	Statistic Reported	Study-Specific Criteria	Inter-Rater Results	Intra-Rater Results	Intra-Rater Methodology
Beaver	Inter-rater = Kappa test; Intra-rater = Spearman correlation coefficient	Significance was set at Kappa values of >0.30 and correlation coefficient of 0.40	Kappa values ranged from 0.097 to 0.766, with only 16/72 combinations reaching values at or above 0.30	Correlation coefficients were between 0.419 and 0.778	18.5% of the photographs were re-rated
Chang	Inter-rater = Quadratically weighted kappa & proportion of agreement; Intra-rater = Pearson's coefficient	"Parameters with poor inter- or intrarater values were excluded from analysis; this includes ICC coefficients below 0.4, Kappa values below 0.4, proportion of agreement below 0.65, Pearson's or Spearman's coefficients below 0.5, and slope values below 0.75 or above 1.25." Values for 'low' and 'moderate' qualifiers were not reported.	"Interrater reliability for the vibratory parameters, mucosal wave, vertical phase, and glottic closure ranged from 0.56 to 0.74 for kappa, from 0.59 to 0.76 for composite proportion of agreement, and from 0.50 to 0.72 for non-normal proportion of agreement. Only mucosal wave met the kappa and proportion of agreement cutoffs of 0.4 and 0.65 respectively."	"Spearman's rho values for vibratory characteristics ranged from 0.17 to 0.96 and slope values ranged from 0.17 to 0.95. One of the vibratory characteristics, vertical phase exceeded the intrarater reliability cutoff values of 0.5 for correlation and below 0.75 or above 1.25 for slope."	10% of the videos were graded twice
Dejonckere	Spearman's correlation coefficients	Not explicitly reported. "These values are in agreement with previous findings" which were noted to "show sufficient reliability"	0.68 for glottal closure, 0.60 for regularity and 0.66 for mucosal wave	N/R	N/A
Hillel	Cohen's kappa	Used qualifiers (excellent, fair, poor) in association with reporting numerical results. A priori cut-off values were not provided.	For mucosal wave, "The interrater reliability was excellent at the preoperative rating session ($\kappa = 0.90$) and poor at the postoperative session ($\kappa = 0.38$). For glottic closure, "The preoperative interrater reliability was very poor ($\kappa = 0.09$), as was the postoperative interrater reliability ($\kappa = 0.20$)."	"The intrarater reliability for mucosal wave ratings was excellent at the preoperative time point (rater 1 $\kappa = 1.00$; rater 2 $\kappa = 0.80$); and "fair at the postoperative time point (rater 1 $\kappa = 0.63$; rater 2 $\kappa = 0.68$). For the glottic closure ratings, "the preoperative intrarater reliability was poor for rater 1 ($\kappa = 0.36$) and fair for rater 2 ($\kappa = 0.59$). For rater 2, the intrarater reliability was excellent at the postoperative time point ($\kappa = 1.00$)."	100% of recordings were rated twice by each of the 2 raters
Karpenko	Percent agreement	Qualified values reported as 'high'.	77%	89%	3 raters re-rated 12 recordings (4 patients at 3 times)
Lam	Pearson's correlation	No qualifier used.	N/A	0.93	2.4% of recordings were re-rated
Morgan	Kappa statistics	Moderate agreement = 0.4–0.59 =; Substantial agreement = 0.6–0.79	N/A	* Mean 0.55; Symmetry 0; R Amplitude 0.32; L Amplitude 1;	24% of examinations were rated twice

First Author	Statistic Reported	Study-Specific Criteria	Inter-Rater Results	Intra-Rater Results	Intra-Rater Methodology
				R Periodicity 0.4; L Periodicity 0.4; R Periodicity 0.4; L Periodicity 0.8; R Periodicity 0.4; L Periodicity 0.6; R Periodicity 0.4; L Periodicity 0.6;	D Duration of Closure 0.8; Degree of D Duration of Closure 0.8; Degree of D Duration of Closure 0.8; Degree of D Duration of Closure 0.8; Degree of
Smith	Pearson's correlation	Qualified intra-rater reliability results as 'good' when reported with the values.	Ranged from 0.87 to 1.0	Ranged from 0.78 to 0.94	25% re-rated
Steward	Pearson's correlation	Used qualifiers 'good', 'reasonable', 'poor' and 'questionable' when reporting values. A priori cut-off values were not provided.	Poor 0.17, two correlated raters had reasonable reliability at 0.43	Good for 3 raters: 0.83, 0.90, 0.90 and questionable for one rater 0.49	Each of the 4 raters rerated 10% of the recordings
van Gogh	Weighted Kappa values	Used qualifiers 'moderate to good' and 'poorly weighted' when reporting values. A priori cut-off values were not provided.	N/A	0.46 to 0.88 for 14 items, 0.098 for phase symmetry	21% of recordings were rated twice by 2 raters
Wang (2013a)	Pearson's correlation	A P value <0.05 was considered significant. Results were considered "substantial and significant".	N/A	0.80	Re-evaluated 10% of recordings

N/R = Not Reported; N/A = Not Applicable; R = Right; L = Left

* Estimates from figure in article