

TITLE

DOI: <https://doi.org/10.20378/irb-6345>

Title: The inter-rater reliability of stroboscopy evaluations

Running title: Inter-rater reliability

Authors: Prof. Dr. Tadeus Nawka¹, PD Dr. Uwe Konerding²

¹ Charité Campus Mitte, University Medicine Berlin, Clinic for Audiology and Phoniatics,
Charitéplatz 1, D-10117 Berlin, Germany

² University of Bamberg, Trimberg Research Academy

Requests for reprints should be sent to

Prof. Dr. Tadeus Nawka,

Campus Charité Mitte, University Medicine Berlin,

Department of Audiology and Phoniatics,

Charitéplatz 1,

D-10117 Berlin,

Germany

Tel. : + 49 (0)30-450 555 024

Email: tadeus.nawka@charite.de

The work on this manuscript was funded by a grant from XION GmbH.

SUMMARY

Objectives/Hypotheses. To investigate the interrater reliability of stroboscopy evaluations assessed using Poburka's Stroboscopy Evaluation Rating Form (SERF).

Study Design. Single-factor experiment with repeated measures on the same element.

Methods. Evaluations of nine experts pertaining to 68 stroboscopy recordings and 16 SERF variables were analyzed. For the 14 SERF variables possessing interval scale level, interrater reliability was investigated using the intraclass correlations for absolute agreement (ICC-a) and consistency (ICC-c). ICCs-c were computed for both original values and values standardized with respect to raters' means and standard deviations (ipsative values). For the two nominally scaled SERF variables, "vertical level" and "glottal closure" interrater reliability was investigated using kappa coefficients.

Results. For evaluations of single raters, ICCs-a ranged from 0.32 to 0.71, ICCs-c for original values from 0.41 to 0.72, and ICCs-c for ipsative values from 0.43 to 0.72. For mean evaluations of two raters, the corresponding values were 0.48 to 0.83 for ICCs-a, 0.58 to 0.84 for ICCs-c for original values, and 0.60 to 0.84 for ICCs-c for ipsative values. The interval scale variables with the lowest interrater reliabilities were phase closure, phase symmetry, and regularity. The kappa coefficients for vertical level and glottal closure were 0.15 and 0.38, respectively.

Conclusions. The interrater reliabilities for vertical level, glottal closure, phase closure, phase symmetry, and regularity are so low that these variables should not be assessed via stroboscopy. For the remaining variables, adequate reliability can be obtained by aggregating evaluations from at least two raters.

Key Words: Stroboscopy—Laryngeal examination—Voice diagnostics—Interrater reliability—Intraclass correlation.

INTRODUCTION

In the 19th century, when cadaver larynges were examined during phonation, the investigators noted that it might be fruitful to investigate human phonation by means of stroboscopy [1].

Stroboscopy is a method of imaging the vocal folds during phonation either with fundamental-frequency coupled flashlight or with shutter opening of a video camera. Stroboscopy depicts the epithelium at the free edge of the vocal folds during phonation. By showing images of the vocal folds during motion, different phases of the cycles of vocal fold vibration are captured. Vibrational characteristics are expressed by shape and dislocation of the vocal fold tissue related to its assumed position at the onset of movement. Although the resulting images are a reduced representation of the true vocal fold movement, they can be regarded as a representation of the motion that produces the sound. Stroboscopy reveals small structural changes that may impair the normal function. It has augmented laryngeal diagnostics with exact visualization and has gained a central position in the clinical examination of patients with voice disorders.

A central problem of stroboscopy is that it yields only uninterpreted complex images. However, for clinical diagnostics and, even more so, for clinical research, adequately quantified parameters with a clear meaning are required. With regard to this requirement, several authors [1–7] have developed forms for expert raters to systematically register those laryngeal features that might be important for voice production. At the present time, there is hardly any research concerning the measurement characteristics of these forms and, correspondingly, none of these forms is commonly accepted. The most comprehensive and analytically best-elaborated one of these forms is the Stroboscopy Evaluation Rating Form (SERF) developed by Poburka [5] (Appendix). The parameters evaluated in this form are amplitude, mucosal wave, nonvibrating portion, supraglottic activity, vocal fold edge smoothness, vocal fold edge straightness, vertical level, phase closure, phase symmetry, regularity, and glottal closure pattern. SERF requires the registration of stroboscopic findings in a systematic way and with numeric values. The numeric values are asked for in such a way that, as far as possible,

measurement on interval scale level is supported. This, in turn, is a basic presupposition for the application of many important statistical procedures.

From a mere analytical point of view, SERF is excellent. However, for responsible interpretation of results produced by SERF, empirical information about its measurement characteristics, that is, its reliability and validity, is required. Reliability defines the upper bound of validity because a measurement instrument can only be as valid as it is reliable. For stroboscopy evaluation, the most important component of its reliability is the extent to which different raters judge the same stroboscopy videos in the same way, that is, the interrater reliability. Consequently, empirical research concerning SERF should first focus on this component.

At present, there is only one study addressing the measurement characteristics of the SERF parameters. This study is concerned with SERF's interrater reliability and was reported by Poburka in the context of his presentation of SERF. This study, however, only refers to the results of three raters, including the principal investigator. Moreover, the statistics applied by Poburka do not characterize systematically enough the way in which different raters might deviate from each other. To be specific, raters might differ with respect to the heights of their evaluations, ranges of their evaluations, or other features such as the rank order between the video recordings. These different kinds of deviations play a very different role in the further processing and interpretation of the resulting values. Therefore, the interrater reliability of the SERF parameters should be investigated with statistics that appropriately reflect these different aspects of the interrater reliability and with a broader and more representative sample of expert raters than in Poburka's study.

In the following, a study about the interrater reliability of evaluations of stroboscopic parameters is presented. This study has two aims. The first is to check whether the reliability of the parameters in SERF is sufficient. If reliability is practically zero, there is no sense in pursuing any further investigations into validity, nor is there any justification for measuring the corresponding parameters via stroboscopy. If reliability is higher than zero but still so low that alternative approaches to

measuring the corresponding parameters are more promising, then these alternative approaches should be applied. However, as long as the reliability of these parameters is definitely higher than zero, investigations concerning their validity still make sense. The second aim is to provide information for responsible application, both in clinical diagnostics and in clinical research, of the parameters that remain in SERF.

MATERIALS AND METHODS

Materials

The stroboscopic recordings were selected from a large number of videos recorded in a phoniatic clinic between June 2000 and April 2002. The recordings were made with a Timcke Stroboscope (Timcke/Rehder, Hamburg, Germany), a 90-degree rigid laryngoscope attached to a Panasonic camera (Panasonic Corporation, Osaka, Japan), and a computer database system (rpSzene; Rehder/Partner GmbH, Hamburg, Germany) that converted the analog video into Motion JPEG-Audio Video Interleave (M-JPEG-AVI) format at 25 frames per second. Only videos that showed organic changes of the vocal folds scheduled for phonosurgery were selected. Of these circa 140 cases, those were chosen that fulfilled the following requirements: good illumination of the inner larynx, a sharp image, visibility of the whole length of the vocal folds, a phonation time long enough to allow for registration of a sustained phonation, and at least one complete cycle of vibration, that is, 25 subsequent frames.

The final material for evaluation consisted of 72 stroboscopic video recordings from 42 patients (30 female). Thirty patients (23 female) were recorded twice, the first time before and the second time after phonosurgical intervention. The diagnoses for all 42 patients were Reinke edema 17 (16 female), vocal fold polyp 16 (eight female), vocal fold nodule three (female), papilloma two (one female), contact granuloma one (male), sulcus vocalis one (female), atrophy of the vocalis muscle one (female), and a cyst at the arytenoid cartilage one (male). The whole group had a mean age of 47.7 ± 13.04 years, the female patients 46.2 ± 14.09 years, and the male patients 51.2 ± 12.74 years.

The mean length of the video recordings was 16.5 seconds (standard deviation: 12.0, minimum: 4 seconds, maximum: 92 seconds). The recordings were stored on DVDs, which could be viewed with either the original rpScene system or other media players.

Raters

The raters were nine phoniatricians (six female) who had training and experience in doing stroboscopic examinations regularly for a period ranging between 2 and 37 years. Neither of the authors was among the raters.

Procedure

The principal investigator arranged a meeting in which the material was demonstrated to the raters and three of the video recordings were evaluated in open discussion to familiarize the raters with SERF. Two weeks later, every rater was handed a copy of the recordings as AVI files with M-JPEG compression. This allowed the raters to play the recordings repeatedly or use the shuttle mode and, thereby, to examine the recordings very thoroughly. For practical reasons, three of the SERF parameters were evaluated in a different manner to that originally proposed by Poburka: (1) the nonvibrating portion was assessed in multiples of 10% instead of 5%; (2) for phase closure, the percentage of time the glottis was closed during the vibratory cycle was estimated, not the time when it was open; and (3) the shape of glottal closure was numbered as: 1 for “hourglass,” 2 for “complete,” 3 for “incomplete,” 4 for “irregular,” 5 for “posterior gap,” 6 for “anterior gap,” 7 for “spindle gap,” and 8 for “variable pattern.”

The raters evaluated the stroboscopic recordings independently at their workplaces and sent their evaluation protocols to the principal investigator.

Statistical analyses

Video recordings were included in the analyses when less than 10% of the requested evaluations were missing. To avoid having to omit further parts of the data from the analyses and make certain that all analyses could be performed for all recordings, missing values in these recordings were imputed. The procedure for imputation depended on the scale level of the parameters. With the exception of vertical level and glottal closure, all SERF parameters were assumed to be interval scales. For these 14 interval scale parameters, imputation of missing values was performed using linear regression. To be specific, for each of the 14 parameters, linear regression was performed with the valid values as dependent variables and dummy-coded raters and video recordings as independent variables. The missing values were then imputed from the values predicted by the corresponding regression equation. Predicted values lower than the lower bound of the variable were then replaced by this lower bound and, correspondingly, predicted values higher than the upper bound were replaced by this upper bound. Such imputation procedures using multiple linear regression can be assumed to produce more valid data sets than the removal of cases with incomplete data would do [8]. For the vertical level and glottal closure parameters, which are assumed to be nominal scales, modal values of the evaluations given by the other raters for the same recording were inserted for the missing values.

The different kinds of deviations from perfect interrater reliability as outlined in the introduction apply to interval scale variables only. Hence, these deviations were only investigated for these 14 interval scale SERF parameters. For this purpose, three different specifications of the intraclass correlation (ICC) [8,9] were applied. ICCs describe the concordance of two or more rows of measurements that refer to the same objects and are given on the same scale. In the study presented here, the evaluations given by a single rater with respect to one parameter for all recordings constitute such a row of measurements. There are different specifications of the ICC, which capture different kinds of deviations from absolute concordance. The ICC for absolute agreement (ICC-a) reflects all possible deviations from concordance, that is, differences in height,

range, and all other differences. The ICC for consistency (ICC-c) applied to the original values does not reflect differences in height but only differences in range and all other differences. The ICC-c applied to ipsative values, that is, to measurements that are standardized with respect to the means and the standard deviations of the corresponding rows, does not reflect differences in height and range but only the remaining differences such as differences in rank order.

When ICCs are applied for investigating interrater reliability, a further distinction is made in the literature: ICCs for all raters and ICCs for single raters [9]. An ICC for all raters is an estimate of the reliability for the average values of all raters considered; an ICC for single raters is an estimate of the average reliability for measurements produced by a single rater. The ICC-c for single raters applied to ipsative values is identical with the mean of all product-moment correlations between the raters. In the literature on ICCs [9,10], further distinctions are made. These distinctions take into account that the raters and objects to be rated can be considered either as a sample drawn by chance from a very large if not infinite population or as the population itself. In the first case, they are referred to as random factors and in the second case as fixed factors. A factor must be considered as random if inferences from the sample on a larger population are intended. If, in contrast, only statements about the raters or the objects considered in the study are intended, then the respective factor must be considered as fixed.

In the analysis presented here, the ICC-a, ICC-c for original values, and ICC-c for ipsative values were computed. In all three cases ICCs for single raters were used. Recordings and raters were both considered as random factors. The ICC-a was applied to take into account all possible deviations from perfect reliability, that is, differences in height, range, and any other differences. This coefficient is relevant for judging the reliability of absolute values or means of absolute values of SERF parameters. The ICC-c for original values was applied to focus on differences in range and on other differences than in height and range. When this coefficient is used, differences in height are disregarded. This coefficient is relevant for judging the reliability of differences of SERF parameters expressed in SERF

terms. The ICC-c for ipsative values was applied to focus only on other differences than differences in height and range.

This coefficient is relevant for judging the reliability of all statistics in which differences between SERF parameters are standardized with respect to standard deviations of the same parameters. Examples of such statistics are effect sizes, t and F values computed for testing differences between means, and Pearson correlations.

The analyses presented here focus on ICCs for single raters because clinical diagnostics is usually performed by one rater. However, especially in clinical research, sometimes two or even three raters might be available. The evaluations of these two or three raters could be averaged to produce more reliable values. Depending on the purpose of the data analyses, different values must be averaged. Averaging of original values is optimal for improving the reliability of means of absolute values or of differences in SERF terms. Averaging of ipsative values is optimal for improving the reliability of statistics based on standardized differences. To investigate how the reliability increases by applying these approaches, ICC-a and both kinds of ICC-c were estimated for two and three raters using the Spearman-Brown formula [11–13]. For all three ICCs and all numbers of raters considered, the lower and upper bounds of the 95% confidence interval were computed.

For a better interpretation of the different coefficients, several further computations were performed for the 14 interval scale parameters. To enable a better comparison, these parameters were standardized with zero for the lowest and 100 for the highest possible value on SERF. All parameters that are percentages already correspond to this standardization. For the parameters standardized this way, several statistics were computed: (1) the total mean, (2) the total standard deviation, (3) the standard deviation due to video recordings, (4) the standard deviation due to raters, and (5) the standard deviation of the residuals.

The statistics computed here serve different purposes for interpreting the data. The total mean and total standard deviation characterize distributions of the parameters. The standard deviation due to

video recordings was estimated because all measures of interrater reliability increase with this statistic independently of the true agreement between the raters. Therefore, measures of interrater reliability referring to different parameters should only be compared under consideration of the corresponding standard deviations due to video recordings. The standard deviation due to raters indicates the extent to which there are systematic differences in the heights of the raters' evaluations. Thus, this statistic captures one kind of possible deviation from absolute agreement between the raters. The standard deviation of the residuals, in turn, reflects the remaining kinds of possible deviations from absolute agreement. These encompass systematic differences in the widths of the raters' evaluations and unsystematic errors. The standard deviations due to video recordings and raters were tested for deviation from zero using an F test with the residual mean squares as denominator.

To obtain information about the quality of the different raters' evaluations for the 14 interval scale variables, the ICCs-a, ICCs-c, and the product-moment correlations between these evaluations and the mean evaluations of all raters were calculated as quality measures for each rater. Thus, for each rater and each of the three quality measures 14 coefficients were computed. The three quality measures for the raters correspond to the three different coefficients applied here for investigating the interrater reliability. The ICCs-a reflect all deviations of the single rater evaluations from the mean evaluations, the ICCs-c reflect all deviations apart from deviations in height, and the product-moment correlations reflect neither deviations in height nor deviations in range. To examine whether the raters differ with regard to these three quality measures, an analysis of variance was conducted for each of the three quality measures. In these analyses, the raters were considered as a between-subject factor and SERF parameters as a within-subject factor. Another issue examined was whether experience affects the quality of the stroboscopy evaluations. For this purpose, the raters were divided into two groups, one with 2–5 years and the other with 10–37 years of experience, and both groups were compared with regard to all three quality measures. These comparisons were

accomplished by conducting analyses of variance with experience group considered as a between-subject factor and SERF parameters considered as a within-subject factor.

For the two nominal scale parameters, that is, vertical level and glottal closure, interrater reliability was investigated using the kappa coefficient.¹⁴ The parameter “glottal closure” comprises eight categories of which two, that is, “complete” and “posterior gap,” usually occur in nonpathologic states and the remaining six in pathologic states. For this reason, a kappa coefficient for glottal closure dichotomized according to these two superordinated categories (tending toward nonpathologic and tending toward pathologic glottal closure) was also computed.

RESULTS

For four of the 72 video recordings, more than 10% of the data were missing. These four video recordings were excluded from further analyses. Of the 9792 values requested for the remaining 68 recordings, only 49 (0.5%) were missing. These values were replaced according to the procedures described above. The means of 12 of the 14 interval scale parameters are nearer to the lower end of the scale. The two exceptions are phase symmetry and regularity (Table 1). For these two variables, the upper end of the scale represents a healthy condition, whereas the lower end of the scale represents a healthy condition for all the other parameters except phase closure. For phase closure, the optimal value is in the middle of the scale. The total standard deviations cover large parts of the continuum. The same also holds for the standard deviations due to video recordings (Table 1). All standard deviations differ significantly from zero. Moreover, there are also essential differences between the different standard deviations. The ratio between the largest and the smallest standard deviation for the video recordings is 2.9. The standard deviations due to raters also differ significantly from zero, that is, for each parameter, raters systematically differ with respect to the heights of their evaluations.

Insert Table 1 about here

The standard deviations of the residuals are nearly as high as the standard deviations due to video recordings. The ICCs-a for single raters are not very high. They range from 0.32 for regularity to 0.71 for vocal fold edge straightness left (Table 2). For means of two raters, they range from 0.48 to 0.83 and for means of three raters from 0.58 to 0.88. The single rater ICCs-c for original values are all higher than the corresponding ICCs-a. They range from 0.41 for phase closure to 0.72 for vocal fold edge straightness left (Table 2). For means of two raters, they range from 0.58 to 0.84 and for means of three raters from 0.68 to 0.88. The single rater ICCs-c for ipsative values are still a little higher than for original values. For single raters, these coefficients range from 0.43 to 0.72 and the amounts of increase range from 0.00 for amplitude left, mucosal wave left, and vocal fold edge straightness left to 0.07 for vocal fold edge smoothness right. For means of two raters, the coefficients range from 0.60 to 0.84 and for means of three raters from 0.69 to 0.89.

Insert Table 2 about here

The rank order of the 14 interval scale parameters is very similar for all the coefficients of reliability. The Kendall tau-b between the ICCs-a and the ICCs-c for original values is 0.74, it is 0.80 between the ICCs-a and the ICCs-c for ipsative values, and it is 0.97 between the two ICCs-c. The corresponding product-moment correlations are 0.92, 0.91, and 0.97, respectively. The two values for vocal fold edge straightness are always the highest, followed by values for mucosal wave and nonvibrating portion. Phase closure and regularity always have the lowest coefficients. Considering the standard deviations due to video recordings, the low values for these two parameters are certainly not a statistical artifact. However, the coefficients for phase symmetry might well be overestimated.

Hence, with equal standard deviations due to video recordings, the reliability coefficients for phase symmetry will be about as low as those for phase closure and regularity. The boundaries of the confidence intervals reveal that all coefficients for vocal fold edge straightness are significantly higher than the coefficients for phase closure and regularity. For the comparison with phase symmetry, this holds for four of the six coefficients.

The coefficients reflecting the relation between single rater evaluations and mean evaluations differ among raters (ICCsa averaged for each rater over the 14 parameters mean = 0.70, SD = 0.05, minimum = 0.66, maximum = 0.79; ICCs-c averaged for each rater over the 14 parameters mean = 0.75, SD = 0.04, minimum = 0.71, maximum = 0.78; correlations averaged for each rater over the 14 parameters mean = 0.79, SD = 0.03, minimum = 0.75, maximum = 0.84). According to the analyses of variance, these differences are statistically significant for all three coefficients ($P < 0.001$). However, for raters with less experience in stroboscopy and raters with more experience in stroboscopy, the means of the quality measures are virtually identical (ICCs-a: mean for less experience \approx 0.702, mean for more experience = 0.700, $P = 0.952$; ICCs-c: mean for less experience \approx 0.751, mean for more experience \approx 0.745, $P = 0.818$; correlations: mean for less experience = 0.786, mean for more experience = 0.798, $P \approx$ 0.368).

The kappa coefficients for the two nominal scale variables are poor. The coefficient for vertical level is only 0.15 and not significantly different from zero. This is because displacements of vertical level occur extremely rarely in the sample of video recordings and therefore the percentage of agreements by chance is already 91.0%. Consequently, coefficient kappa is very low although the percentage of agreements is 92.3%. The kappa coefficient for glottal closure is 0.38 and, at least, significantly different from zero ($P < 0.001$). For glottal closure states categorized as “tending toward pathologic” and “tending toward nonpathologic,” it is 0.43 and is significantly different from zero ($P < 0.001$).

DISCUSSION

In the study presented here, originally 72 different video recordings were evaluated by nine different raters with respect to 16 different parameters. Because of missing values, four of the 72 video recordings had to be excluded from the analyses. The remaining 68 video recordings still provide a very good empirical basis for the analyses performed here. A limiting aspect of the study might be that the raters were trained before the rating and that three of the 72 recordings evaluated in the study had been used for this training. These activities might have slightly enhanced the reliabilities. On the other hand, training of judgments before the evaluation will always be reasonable when stroboscopy evaluations are performed for clinical research. Hence, the training has enhanced the ecological validity of the study.

The data show that the interrater reliability for the two nominal scale variables is very poor. For the variable “vertical level,” the kappa coefficient does not even differ significantly from zero, that is, the agreement between the raters is not significantly better than it would have been if the raters had only guessed. This poor result, however, might be partly due to the fact that none of the recordings belonged to a disease pattern for which a displacement of vertical level can be expected. Examples of such disease patterns would be vocal fold atrophy or vocal fold paresis. Because of the lack of actual vertical level displacements, the few displacements reported by the raters are most probably false alarms and these false alarms can hardly be expected to correlate highly between the raters. Hence, the variable “vertical level” might be useful for special examinations of patients who have already been diagnosed via laryngoscopy for a disease that might be associated with vertical displacement. In investigations with a broader scope, in contrast, this variable might be better omitted. The kappa coefficient for the other nominal scale variable, that is, for “glottal closure,” is also quite modest. This coefficient does not even increase much when this variable is dichotomized into pathologic and nonpathologic states. The latter means that there is a large amount of disagreement as to whether the shape and extent of closure to be judged is pathologic. This finding suggests that “glottal closure” should also be removed from SERF.

The interrater reliability of the 14 interval scale parameters is distinctly higher than that of the two nominal scale parameters. Yet it is still not overwhelming. The raters systematically differ with respect to the heights of their evaluations for all 14 parameters. For the ICCs-a (for original values) only four, for the ICCs-c (for original values) six, and for the ICCs-c (for ipsative values) eight of the 14 parameters are at least as high as 0.55, which, according to the classification of Landis and Koch [15], is the lower limit of a coefficient with adequate reliability. Considering not only the ICCs but also the standard deviations due to recordings and the residual standard variations, phase closure, phase symmetry, and regularity are the three of the 14 interval scale parameters with the lowest reliability. This coincides with the fact that stroboscopy registers only the average movement of the vocal folds by merging vibrational phases from consecutive cycles. Consequently, short-term changes of movements that usually have a period length of 2.5–10 milliseconds cannot be determined. Hence, stroboscopy is not the appropriate technique for measuring these three parameters. The detection of real cycle-to-cycle changes would require registration in real time by high-speed videography^{16,17} or videokymography^{18,19} with a high temporal resolution of up to 4000 frames per second. Phase closure, phase symmetry, and regularity should be measured by these techniques instead of being guessed by means of stroboscopy.

The remaining 11 parameters comprise those parameters that have hitherto been judged as most important for clinical diagnostics. The first parameter is vocal fold edge straightness, the characteristic that can be diagnosed best by laryngoscopy and stroboscopy. Inevitably, the examiner assesses the deviation from the straight vocal fold when he or she finds an organic lesion such as a polyp. The second parameter is the mucosal wave, the finest motion of the epithelium, which depends on its pliability. Voice specialists have long been aware that the absence of the mucosal wave is linked to impaired voice quality. The third parameter is the extent of nonvibrating portions of the vocal folds. From the beginning of systematic stroboscopic examination, the detection of nonvibrating vocal folds was one of the most important findings to identify glottal cancer.

The pattern of the reliability coefficients of the 11 interval scale parameters that remain after excluding phase closure, phase symmetry, and regularity from SERF is much more advantageous than for all 14 parameters. For single raters, only three of the 11 ICCs-c for ipsative values fall short of adequate reliability and they do this only slightly. Nevertheless, there are still reliabilities that are not adequate, especially for the ICCs-a and for ICCs-c for original values. Therefore, it is advisable to rely on more than one rater. In clinical diagnostics, this should be done at least in serious cases. In these cases, the evaluation should be validated by a colleague. In clinical research, it should be usual practice to aggregate the evaluations of at least two independent raters. With two raters, all three coefficients are higher than 0.55, that is, higher than the lower bound of adequate reliability, for all 11 parameters. For three raters, even the lower bounds of the 95% confidence intervals of both kinds of ICC-c are higher than this critical value. For the ICCs-a, this holds for all parameters except for supraglottic activity. More than three raters are usually very difficult to recruit, and the gain in reliability decreases with each additional rater. Hence, two or three raters will usually be the optimal compromise between accuracy of measurement and minimization of time and effort.

In this context, a serious warning must be given! The increase in reliability reported here for two or three raters will only be achieved when these raters evaluate in a completely independent manner. The raters have to view the video recordings separately and must not be given information on the evaluations of the other raters. The evaluations must not be discussed afterward to obtain a consensus! Such a procedure would enhance the agreement between the raters and increase all types of ICCs. However, it would also lower the agreement with the evaluations of other experts. In other words, the procedure of finding consensus in one institution would reduce the reliability of stroboscopic assessment in the sense of reproducibility of the results in another institution under equivalent conditions.

The results suggest that the raters differ in their quality. However, for raters who have performed stroboscopy for at least 2 years, differences in experience have no effect on the quality of the

evaluations. This suggests that 2 years of experience are sufficient to secure the maximal quality that can be achieved.

CONCLUSION

The interrater reliability for the two nominal scale variables “vertical level” and “glottal closure” is so low that the corresponding evaluations have practically no informative value at all. Any further considerations concerning the validity of these variables are pointless, and it makes no sense to have these variables assessed via stroboscopy. For all 14 interval scale parameters, the interrater reliability is at least that high that their evaluations possess some informative value. Hence, considerations concerning the validity of these evaluations make sense. However, the three interval scale variables with the poorest reliability, that is, phase closure, phase symmetry, and regularity, can be measured more accurately using high-speed videography or videokymography instead of stroboscopy.

Therefore, there is little sense in continuing to assess these three variables with stroboscopy. The interrater reliability of the remaining 11 parameters is still quite modest when only single raters are applied. Averages of evaluations from two raters, however, possess adequate reliability. For averages of evaluations from three raters, the reliability is even quite good. Therefore, if possible, at least two raters should be applied in clinical research. The validity of the 14 interval scale variables in SERF is still unclear. Further research is required to obtain more information about this.

Acknowledgments

We would like to thank Sylva Bartel-Friedrich, Roswitha Berger, Jörg Flaschka, Michael Fuchs, Holger Hanschmann, Silke Heidemann, Dagmar Kayser, Petra Schelhorn-Neise, and Karin Winter for evaluating the stroboscopic video recordings; Thomas Murry for critically discussing a former version of the manuscript; and Peter Bereza for correcting our English.

REFERENCES

1. Bless DM, Hirano M, Feder RJ. Videostroboscopic evaluation of the larynx. *Ear Nose Throat J* 1987;66:289-296.
2. Boehme G, Gross M. *Stroboscopy*. London and Philadelphia: Whurr, 2005:170.
3. Friedrich G. Qualitätssicherung in der Phoniatrie [Quality assurance in phoniatrics. Recommendation for standardization of clinical voice evaluation]. *HNO* 1996;44:401-416.
4. Hirano M, Bless DM. *Videostroboscopic Examination of the Larynx*. San Diego: Singular Publishing Group, 1993.
5. Poburka BJ. A new stroboscopy rating form. *J Voice* 1999;13:403-413.
6. Rosen CA. Stroboscopy as a research instrument: development of a perceptual evaluation tool. *Laryngoscope* 2005;115:423-428.
7. Stemple JC, Gerdemann BK, Kelchner LN. Instrumental measurement of voice. In: Stemple JCa, ed. *Clinical Voice Pathology*. San Diego: Singular Publishing Group, 1998.
8. Raessler S, Rubin D, Zell E. Incomplete Data in Epidemiology and Medical Statistics. In: Rao C, Miller J, Rao D, eds. *Handbook of Statistics 27: Epidemiology and Medical Statistics*. Munich: Elsevier, 2008:569-601.
9. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 1979;86:420-428.
10. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. *Psychological Methods* 1996;1:20-46.
11. Brown W. Some experimental results in the correlation of mental abilities. *British Journal of Psychology* 1910;3:296-322.

12. Hempel S. Reliability. In: Miles J, Gilbert P, eds. *A handbook of research methods in clinical and health psychology*. Oxford UK: University Press, 2005:193-204.
13. Spearman C. Correlation calculated from faulty data. *British Journal of Psychology* 1910;3:271-295.
14. Fleiss J. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 1971;76:378-382.
15. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159-174.
16. Bonilha HS, Deliyski DD. Period and glottal width irregularities in vocally normal speakers. *J Voice* 2008;22:699-708.
17. Lohscheller J, Eysholdt U. Phonovibrogram visualization of entire vocal fold dynamics. *Laryngoscope* 2008;118:753-758.
18. Svec JG, Sram F, Schutte HK. Videokymography in voice disorders: what to look for? *Ann Otol Rhinol Laryngol* 2007;116:172-180.
19. Svec JG, Sundberg J, Hertegard S. Three registers in an untrained female singer analyzed by videokymography, strobolaryngoscopy and sound spectrography. *J Acoust Soc Am* 2008;123:347-353.

Table 1: Basic statistics for the interval-scale parameters

Parameter	Total mean	Total SD ^a	SD ^a due to video recordings ^b	SD ^a due to raters ^b	SD ^a of the residuals
Amplitude					
Right	26.4	14.6	9.8 ^{***}	5.0 ^{***}	9.8
Left	26.8	15.7	10.4 ^{***}	4.5 ^{***}	11.0
Mucosal wave					
Right	27.9	22.2	15.8 ^{***}	8.6 ^{***}	13.4
Left	27.0	22.5	17.5 ^{***}	7.0 ^{***}	12.7
Non-vibrating portion					
Right	13.3	28.6	21.7 ^{***}	8.2 ^{***}	17.1
Left	14.5	28.4	19.8 ^{***}	9.9 ^{***}	18.3
Supraglottic activity	15.4	21.9	14.5 ^{***}	12.2 ^{***}	11.9
VFE ^c smoothness					
Right	4.7	13.3	9.6 ^{***}	3.0 ^{***}	8.9
Left	5.1	14.0	9.8 ^{***}	2.6 ^{***}	9.7
VFE ^c straightness					
Right	24.1	27.2	21.9 ^{***}	3.4 ^{***}	16.1
Left	25.1	26.9	22.8 ^{***}	2.4 ^{**}	14.3
Phase closure	33.8	27.3	16.2 ^{***}	11.2 ^{***}	19.4
Phase symmetry	56.0	43.0	27.9 ^{***}	17.8 ^{***}	28.3
Regularity	71.7	36.4	20.9 ^{***}	18.9 ^{***}	23.9

^a SD=standard deviation.

^b Statistically significant deviation from zero is marked with * for p<0.05, ** for p<0.01, and *** for p<0.001.

^c VFE = vocal fold edge.

Table 2: Intra-class correlations for the interval-scale parameters^a

Parameter	Number of raters		
	1	2	3
ICC-a			
Amplitude			
Right	0.44 (0.34 - 0.56)	0.62 (0.51 - 0.72)	0.71 (0.61 - 0.79)
Left	0.44 (0.34 - 0.55)	0.61 (0.51 - 0.71)	0.70 (0.61 - 0.78)
Mucosal wave			
Right	0.50 (0.38 - 0.62)	0.66 (0.55 - 0.76)	0.75 (0.65 - 0.83)
Left	0.59 (0.49 - 0.70)	0.74 (0.65 - 0.82)	0.81 (0.74 - 0.87)
Non-vibrating portion			
Right	0.57 (0.46 - 0.67)	0.72 (0.63 - 0.80)	0.80 (0.72 - 0.86)
Left	0.48 (0.37 - 0.59)	0.65 (0.54 - 0.74)	0.73 (0.64 - 0.81)
Supraglottic activity	0.42 (0.27 - 0.57)	0.59 (0.42 - 0.73)	0.69 (0.52 - 0.80)
VFE ^b smoothness			
Right	0.51 (0.41 - 0.61)	0.67 (0.58 - 0.76)	0.76 (0.68 - 0.83)
Left	0.49 (0.40 - 0.60)	0.66 (0.57 - 0.75)	0.74 (0.66 - 0.82)
VFE ^b straightness			
Right	0.64 (0.55 - 0.73)	0.78 (0.71 - 0.84)	0.84 (0.79 - 0.89)
Left	0.71 (0.64 - 0.79)	0.83 (0.78 - 0.88)	0.88 (0.84 - 0.92)
Phase closure	0.34 (0.24 - 0.46)	0.51 (0.39 - 0.63)	0.61 (0.49 - 0.72)
Phase symmetry	0.41 (0.30 - 0.53)	0.58 (0.46 - 0.69)	0.68 (0.56 - 0.77)
Regularity	0.32 (0.21 - 0.45)	0.48 (0.34 - 0.62)	0.58 (0.44 - 0.71)
ICC-c			
Amplitude			
Right	0.50 (0.41 - 0.61)	0.67 (0.58 - 0.75)	0.75 (0.67 - 0.82)
Left	0.48 (0.38 - 0.58)	0.64 (0.55 - 0.73)	0.73 (0.65 - 0.81)
Mucosal wave			
Right	0.58 (0.49 - 0.68)	0.74 (0.66 - 0.81)	0.81 (0.74 - 0.86)
Left	0.66 (0.57 - 0.74)	0.79 (0.73 - 0.85)	0.85 (0.80 - 0.90)
Non-vibrating portion			
Right	0.62 (0.53 - 0.71)	0.76 (0.69 - 0.83)	0.83 (0.77 - 0.88)
Left	0.54 (0.44 - 0.64)	0.70 (0.62 - 0.78)	0.78 (0.71 - 0.84)
Supraglottic activity	0.60 (0.51 - 0.69)	0.75 (0.67 - 0.82)	0.82 (0.76 - 0.87)
VFE ^b smoothness			
Right	0.53 (0.44 - 0.63)	0.70 (0.61 - 0.78)	0.77 (0.70 - 0.84)
Left	0.51 (0.41 - 0.61)	0.67 (0.58 - 0.76)	0.76 (0.68 - 0.82)
VFE ^b straightness			
Right	0.65 (0.56 - 0.73)	0.79 (0.72 - 0.85)	0.85 (0.79 - 0.89)
Left	0.72 (0.64 - 0.79)	0.84 (0.78 - 0.88)	0.88 (0.84 - 0.92)
Phase closure	0.41 (0.32 - 0.52)	0.58 (0.48 - 0.68)	0.68 (0.58 - 0.76)
Phase symmetry	0.49 (0.40 - 0.60)	0.66 (0.57 - 0.75)	0.74 (0.67 - 0.82)
Regularity	0.43 (0.34 - 0.54)	0.60 (0.51 - 0.70)	0.70 (0.61 - 0.78)

ICC-c for ipsative values (mean inter-correlation)			
Amplitude			
Right	0.53 (0.43 - 0.63)	0.69 (0.60 - 0.77)	0.77 (0.70 - 0.83)
Left	0.48 (0.39 - 0.59)	0.65 (0.56 - 0.74)	0.74 (0.66 - 0.81)
Mucosal wave			
Right	0.61 (0.52 - 0.70)	0.75 (0.68 - 0.82)	0.82 (0.76 - 0.87)
Left	0.66 (0.58 - 0.75)	0.80 (0.73 - 0.86)	0.86 (0.81 - 0.90)
Non-vibrating portion			
Right	0.64 (0.56 - 0.73)	0.78 (0.71 - 0.84)	0.84 (0.79 - 0.89)
Left	0.60 (0.51 - 0.70)	0.75 (0.68 - 0.82)	0.82 (0.76 - 0.87)
Supraglottic activity	0.63 (0.54 - 0.72)	0.77 (0.70 - 0.83)	0.83 (0.78 - 0.88)
VFE ^b smoothness			
Right	0.60 (0.51 - 0.70)	0.75 (0.68 - 0.82)	0.82 (0.76 - 0.87)
Left	0.54 (0.45 - 0.64)	0.70 (0.62 - 0.78)	0.78 (0.71 - 0.84)
VFE ^b straightness			
Right	0.67 (0.59 - 0.75)	0.80 (0.74 - 0.86)	0.86 (0.81 - 0.90)
Left	0.72 (0.65 - 0.80)	0.84 (0.79 - 0.89)	0.89 (0.85 - 0.92)
Phase closure	0.44 (0.35 - 0.55)	0.61 (0.52 - 0.71)	0.70 (0.62 - 0.79)
Phase symmetry	0.50 (0.41 - 0.61)	0.67 (0.58 - 0.75)	0.75 (0.67 - 0.82)
Regularity	0.43 (0.34 - 0.54)	0.60 (0.50 - 0.70)	0.69 (0.60 - 0.78)

^a Limits of 95%-confidence-intervals in brackets.

^b VFE = vocal fold edge.

APPENDIX

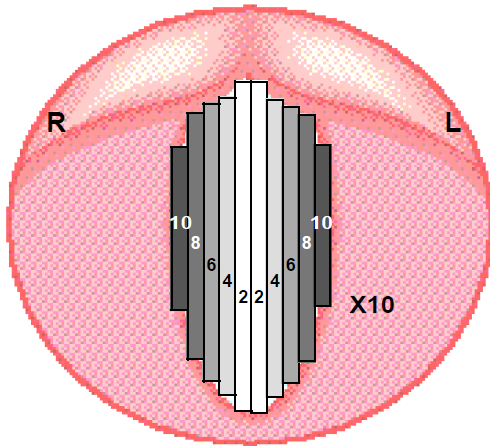
Stroboscopy Evaluation Rating Form

Stroboscopy Evaluation Rating Form (SERF)

Bruce J. Poburka, Ph.D.

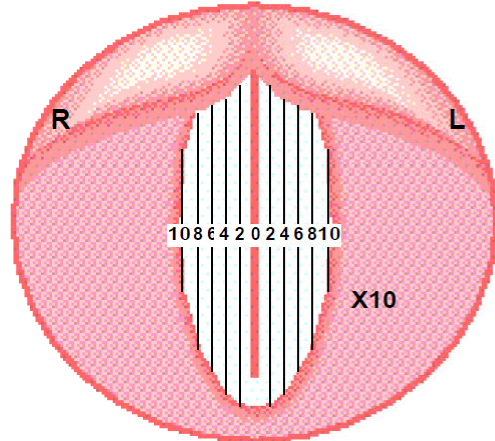
Rater: _____
 Client: _____
 Date: _____

Amplitude (Rate @ normal pitch & loudness)



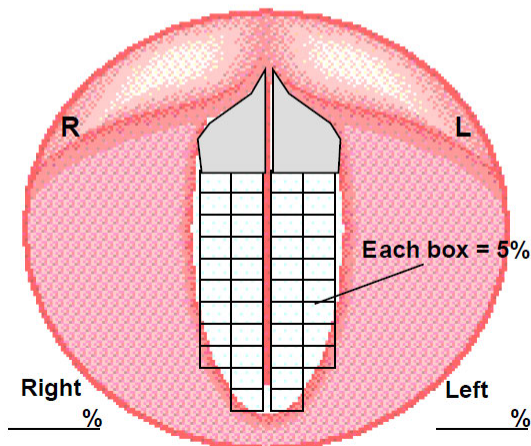
Right: _____% Left: _____%
 Fo: _____

Mucosal Wave (Rate @ normal pitch & loudness)



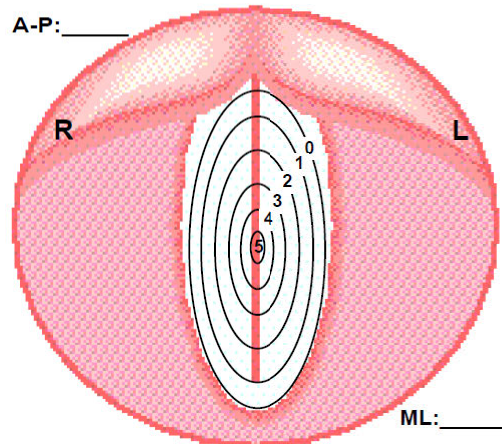
Right: _____% Left: _____%
 Fo: _____

Non-vibrating Portion (shade in affected areas)



Right _____% Left _____%

Supraglottic Activity (Ignore voice onsets)



A-P: _____
 ML: _____



Vocal Fold Edge Smoothness

Right Fold 0 1 2 3 4 5 smooth rough	circle one	Left Fold 0 1 2 3 4 5 smooth rough
---	------------	--

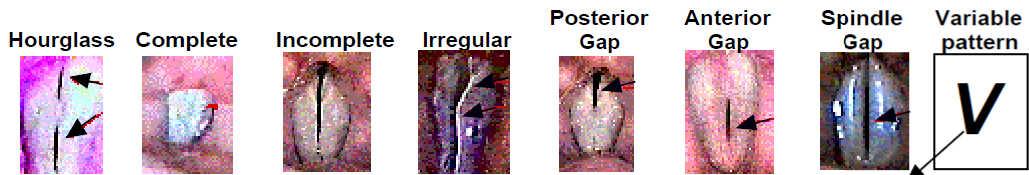
Vocal Fold Edge Straightness

Right Fold 0 1 2 3 4 5 straight irregular	circle one	Left Fold 0 1 2 3 4 5 straight irregular
---	------------	--

Rate @normal pitch & loudness

<u>Vertical Level</u>	<u>Phase Closure</u>	<u>Phase Symmetry</u>	<u>Regularity</u>
circle one cross section view of glottal area.  on-plane  off-plane	Rate @ point of contact % of time open closed +90% <10% 66% 33% "Normal" 33% 66% <10% +90% Frame count: open phase: _____ Closed phase: _____	Rate @ point of contact % of time symmetrical Always assymetrical circle one 0% 20% 40% 60% 80% 100% Always symmetrical	% of time regular Always irregular 0% 20% 40% 60% 80% 100% Always regular Method(s) used: stop phase _____ running phase _____

Glottal Closure



If closure pattern is variable, indicate the predominant closure pattern: _____

Summary/Additional Comments:
