Original Paper

Folia Phoniatrica et Logopaedica

Folia Phoniatr Logop 2008;60:86–90
DOI: 10.1159/000114650

Published online: January 31, 2008

The Interobserver and Test-Retest Variability of the Dysphonia Severity Index

Marieke M. Hakkesteegt Marjan H. Wieringa Michael P. Brocaar
Paul G.H. Mulder Louw Feenstra

Department of Otorhinolaryngology, Erasmus MC – University Medical Center Rotterdam, The Netherlands

Key Words
Dysphonia Severity Index · Voice quality · Interobserver variability · Test-retest variability

Abstract
Objective: The purpose of this study was to investigate the interobserver variability and the test-retest variability of the Dysphonia Severity Index (DSI), a multiparametric instrument to assess voice quality. ***Methods:*** The DSI was measured in 30 nonsmoking volunteers without voice complaints or voice disorders by two speech pathologists. The subjects were measured on 3 different days, with an interval of 1 week. ***Results:*** The difference in DSI between two observers (interobserver difference) was not significant. The intraclass correlation coefficient for the DSI was 0.79. The standard deviation of the difference between two duplicate measurements by different observers was 1.27. ***Conclusion:*** Differences in measurements between different observers were not significant. The intraclass correlation coefficient of the DSI was 0.79, which is to be considered good. Differences in DSI within one patient need to be larger than 2.49 to be significant.

Introduction

Speech pathologists, as well as other clinicians, are more and more stimulated to practice 'evidence-based' treatment. Therefore, measurements are needed to assess results of intervention. Voice disorders are multidimensional, and the assessment of voice disorders should be multidimensional as well, consisting of (video)laryngostroboscopy, assessment of voice quality and subjective self-evaluation of the voice by the patient [1]. For the assessment of voice quality, perceptual as well as objective measures are used. Although there is no consensus yet on what objective measures to use, it seems that multiparametric measures are better at assessing voice quality than single-parameter measures. The Dysphonia Severity Index (DSI) [2] is such a multiparametric measure, and has been used for assessment of voice quality for different groups of patients [3–11]. The DSI is derived from a multivariate analysis of 387 subjects with the goal to describe the perceived voice quality, based on objective measures. The classification of the severity of dysphonia was based on the perceptual assessment, which was scored for grade on the GRBAS scale [12]. The parameters used for the DSI are the highest fundamental frequency (F_0-high in Hz),

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

1021–7762/08/0602–0086$24.50/0

Accessible online at:
www.karger.com/fpl

Marieke M. Hakkesteegt
Department of Otorhinolaryngology
Erasmus MC – University Medical Center Rotterdam
PO Box 2040, NL–3000 CA Rotterdam (The Netherlands)
Tel. +31 10 463 3516, Fax +31 10 463 4240, E-Mail m.hakkesteegt@erasmusmc.nl

lowest intensity (I-low in dB SPL), maximum phonation time (MPT in s) and jitter (%). The DSI is constructed as DSI = 0.13 × MPT + 0.0053 × F_0-high – 0.26 × I-low – 1.18 × jitter (%) + 12.4. It is constructed such that a perceptually normal voice (grade 0) corresponds with a DSI of +5; a severely dysphonic voice (grade 3) corresponds with a DSI of –5. Also scores beyond this range are possible (higher than +5 or lower than –5). An advantage of the DSI is that the parameters can be obtained relatively quickly and easily by speech pathologists in daily clinical practice.

When using an instrument to assess the effects of intervention on voice quality, it is important to know the variability and the measurement accuracy of that instrument, to be able to interpret differences in measurements, for example before and after therapy [13]. The variability of several single objective measures has been investigated [13–22]. The results of these studies are rather diverse for the different measures. Therefore, the variability of a multiparametric measurement such as the DSI cannot be predicted from those results.

The purpose of this study was to test the interobserver variability and to investigate the test-retest variability of the DSI.

Methods

Subjects

Thirty nonsmoking adult volunteers (19 female, 11 male) without voice complaints participated in this study, performed at our Department of Otorhinolaryngology. They were recruited from employees and medical trainees of the hospital. The mean age of the subjects was 26 years (standard deviation, SD, 3.3 years, range 20–35 years). The subjects had no history of voice disorders or voice therapy. A speech therapist scored their voices perceptually as grade 0 on the GRBAS scale [12].

Equipment

Intensity and frequency measurements were obtained with an automatically recording phonetograph (Pabon/Laryngograph 1997). A Sennheiser microphone (BG 2.0 dyn) was used. The distance between mouth and microphone was 30 cm. The MultiSpeech program (Kay Elemetrics) was used for calculating jitter. Audio recordings were made with a sampling rate of 11,025 Hz and 16 bits quantization. A stopwatch was used for measuring the MPT. Data recording took place in a room with 'living room acoustics' [23].

Measurements

From all subjects, measurements for the following four parameters of the DSI were obtained: highest fundamental frequency, lowest intensity, MPT and jitter. Subsequently, the DSI was calculated for each subject.

Frequency and Intensity Measurements

The subjects were asked to phonate an /a/ as softly as possible at a comfortable pitch. After that, they were asked to produce an /a/, starting at a comfortable pitch going up to the highest and down to the lowest pitch. This instruction was accompanied by a demonstration by the speech pathologist. Frequency was measured in Hertz, intensity in dB SPL.

Maximum Phonation Time

The subjects were asked to inhale deeply and sustain an /a/ for as long as possible at a comfortable pitch and loudness. This was recorded 3 times; the longest measured phonation time in seconds was used.

Jitter

The subjects phonated 3 times an /a/ at a comfortable pitch and loudness during approximately 3 s. The jitter was calculated on a sample of 1 s, starting half a second after the voice onset. The lowest result of the three calculations was used.

Measurement Schedules

The subjects were measured three times, with a time interval of approximately 1 week. Measurements were performed by two speech pathologists in two schedules. Schedule 1: measurement 1 and 2 by speech pathologist 1, measurement 3 by speech pathologist 2. Schedule 2: measurement 1 by speech pathologist 2, measurement 2 and 3 by speech pathologist 1.

The subjects were randomly assigned to one of the two schedules. To each schedule 15 patients were assigned. After the first measurement, subjects were explicitly told not to practice the tasks at home.

Statistics

For general interpretation of the reproducibility, a Bland-Altman plot was made for the first and third measurement. For analysis, the statistical program SAS was used. A variance component analysis in a random effect model was performed. Since in daily clinical practice the observer will vary, the analysis was performed with the observer and the subject as random variables and the time of measurement (1st, 2nd, 3rd) as fixed effect. To determine which part of the variability of the measurements is attributable to the differences between subjects, the Intraclass Correlation Coefficient (ICC) was calculated. The ICC is defined as the intersubject variance divided by the total variance. The other part of the difference between measurements is explained by differences between observers (interobserver) and the residual error (intraobserver and intrasubject). The standard error of measurement (σ_{error}) is defined as the square root of the variance of the error (interobserver variance + residual variance). The SD of the difference between two duplicate measurements to the same subject equals $\sigma_{error} \cdot \sqrt{2}$.

Results

Of all 30 subjects, 22 completed 3 measurements (13 females, 9 males) and 8 subjects completed 2 measurements (6 females, 2 males). Five of those dropouts were

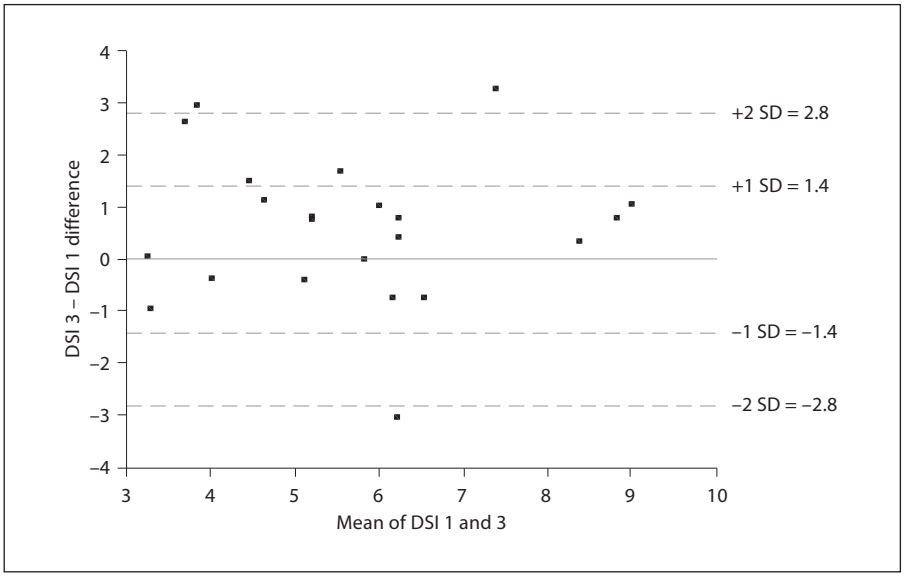

Fig. 1. Bland-Altman plot: the difference between the first and the third measurement (DSI 3 – DSI 1) plotted against the mean of the first and the third measurement (mean of DSI 1 and 3), with the SD of the difference between DSI 1 and 3.

Table 1. Mean values of DSI, F_0-high, I-low, MPT and jitter on measurements 1, 2 and 3

Measurement	DSI	F_0-high	I-low	MPT	Jitter
1	5.6	896	54	23	0.54
2	6.0	953	55	26	0.55
3	6.0	938	54	25	0.64

measured twice by the same speech pathologist (schedule 1), 3 were measured by two speech pathologists (schedule 2).

The mean time interval between measurements 1 and 2 was 9 days (SD 6 days) and between measurements 2 and 3 it was also 9 days (SD 5 days). The mean time interval between measurements 1 and 3 was 18 days (SD 9). In table 1, the mean values of the DSI and all separate parameters for the three different measurements are shown.

Figure 1 shows a Bland-Altman plot of the first and the third DSI measurement. A Bland-Altman plot shows the difference between two measurements against their mean. In figure 1 the y-axis represents the difference between the first and the third measurement (DSI 3 – DSI 1), with the SD of the difference between DSI 1 and DSI 3. The x-axis shows the mean of DSI 1 and DSI 3. The plot shows that a large number of the subjects have a higher DSI the third time than the first time. The fixed effect of the time of measurement ('practice effect') on the DSI was +0.6 from 1st to 2nd measurement and +0.06 from 2nd to 3rd (overall $p = 0.022$).

The total variance of the DSI was 3.92. The intersubject variance was 3.11, the interobserver variance was 0.21 and the residual variance (intraobserver and intrasubject) was 0.60. The ICC was 0.79 (3.11/3.92). For the separate parameters, we found the following ICC values: F_0-high 0.87, I-low 0.57, MPT 0.84 and jitter 0.49.

The measurement error was calculated as follows: variance of the measurement $\sigma^2_{error} = 0.21$ (interobserver variance) + 0.60 (residual variance) = 0.81.

Standard error of measurement $\sigma_{error} = \sqrt{0.81} = 0.90$. The SD of the difference between two duplicate measurements by different observers is $\sigma_{error} \cdot \sqrt{2} = 0.90 \cdot \sqrt{2} = 1.27$.

Discussion

In this study, the interobserver and the test-retest variability of the DSI were investigated. When using an instrument for measuring changes in voice quality (for example before and after therapy), it is important to know the variability and the measurement accuracy of that instrument for the interpretation of the measurements in clinical practice. A test-retest study was done to analyze the relative contribution of various factors that result in differences between repeated measurements of the DSI. To be able to compare the various factors, the ICC was

calculated. The measurement error was calculated to make it possible to determine whether a difference in DSI within one patient is significant, for example before and after therapy.

The DSI was measured in a group of healthy people three times with a 1-week interval, by two speech pathologists. Eight subjects did not complete the 3 measurements planned. Dropout was mainly caused by subjects transferring to another work location. However, since those subjects were equally distributed over both measurement schedules, they do not affect the results.

The Bland-Altman plot shows that there is no relationship between the magnitude of the DSI score and the difference between the two measurements. The plot shows that a large number of the subjects have a higher DSI the third time than the first time; this is possibly due to a 'practice effect', although they were explicitly told not to practice the tasks at home. It is possible that the results were different on the second test because subjects were more familiar with the tasks. The largest contribution to this effect comes from the parameters highest fundamental frequency and MPT. The effect in these healthy subjects was 0.6 between the first and the second measurement. The effect is much smaller between the second and the third measurement. The overall effect is taken into account in the further analysis. It is, however, not clear whether this effect might change with the length of the time interval, and could be smaller or disappear with longer time intervals. Neither is it clear whether a similar effect will be present in patients. We chose a time interval of 1 week in this study because longer time intervals increase the likelihood that individual circumstances change and alter a subject's voice quality. In clinical practice, most time intervals will be much longer than 1 week, and usually will be at least 3 months or more. It is possible that the 'practice effect' may weaken or completely disappear over longer periods of time.

The differences in DSI between the different measurements are caused by 3 components: the intersubject variance, the interobserver variance and the combination of the intraobserver and the intrasubject variance (the residual variance). A reliable measure will be one where the intersubject variance provides the greatest contribution to overall variance. The ICC is 0.79, which means that the variance between subjects (intersubject) is indeed the largest part (79%) of the differences between measurements. The ICC of 0.79 is to be considered 'excellent' [24]. Of the separate parameters, the ICC values of F_0-high and MPT are higher than of I-low and jitter, and of the DSI.

Although the measures used to calculate the DSI are objective, they are obtained from human performances and therefore dependent on cooperation of the subject and stimulation by the observer. Consequently, it is possible that there are differences between observers. The interobserver variance was 0.21. This means that only a small part (5%) of the differences between measurements is due to differences between different observers. This observer effect is not significant ($p < 0.05$). In clinical practice, this means that it does not matter which observer is performing the measurements. This further suggests that studies of different institutes are comparable, assuming that measurements are made in the same way.

Studies of test-retest variability of objective measures of voice quality are sparse, as concluded by Carding et al. [13] as well. We did not find any reports on the test-retest variability of the DSI, or on other multiparametric measures. Furthermore, existing studies on single parameters use different statistical methods to calculate variability, which makes comparisons difficult. Several studies reported an ICC only for 'jitter'. Our results of the ICC of 'jitter' are comparable to the results of Carding et al. [13] and Bough et al. [14]. They found ICCs of 0.46 and 0.31, respectively, for 'jitter'; we found an ICC of 0.49. Also in other studies, 'jitter' is found to be quite variable [15, 17, 25, 26]. The only report we found about test-retest of 'highest fundamental frequency' [22] reported only differences in semitones. These differences were not significant. This is in concordance with the ICC of 0.87 we found. For the 'lowest intensity', it is found that test-retest results remain within about 3-dB differences [20], and that the SD of the differences between two measures is 3 dB [19, 21]. We also found an SD of the difference between the first and third measurement of 3 dB and an ICC of 0.57. For the 'maximum phonation time', Lee et al. [18] reported consistent results for two different measurements. This is in concordance with the ICC of 0.84 in our study.

The measurement error of the DSI was 1.27. In clinical practice, this means that a difference in DSI between two measurements within the same subject is significant ($p > 0.05$) when it is 2.49 (1.96 · 1.27) or more. According to Wuyts et al. [2], the range of scores of the DSI is between –5 and +5. In our clinical experience with quite a large group of patients with a wide range of severity of dysphonia, the range of scores is approximately between –8 and +8. A significant difference in DSI of 2.49 within one patient seems therefore to represent a relatively large difference. When the change in voice quality is quite clear, a larger difference will easily be found. However, in more

subtle voice changes it is very well possible that a measured difference in DSI will not be significant. This significant difference in DSI of 2.49 is applicable to individual patients, but not when comparing groups of patients. The usefulness of the DSI in clinical practice, for example in measuring results of therapy, needs further investigation.

Conclusion

In repeated measurements of the DSI, the variability between subjects is the largest part. The ICC of 0.79 is to be considered good. The differences in measurements between different observers are not significant. Differences in DSI within one patient need to be larger than 2.49 to be significant.

References

1 Dejonckere PH, Bradley P, Clemente P, et al: A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques. Guideline elaborated by the Committee on Phoniatrics of the European Laryngological Society (ELS). Eur Arch Otorhinolaryngol 2001;258:77–82.
2 Wuyts FL, De Bodt MS, Molenberghs G, et al: The dysphonia severity index: an objective measure of vocal quality based on a multiparameter approach. J Speech Lang Hear Res 2000;43:796–809.
3 Timmermans B, De Bodt MS, Wuyts FL, et al: Poor voice quality in future elite vocal performers and professional voice users. J Voice 2002;16:372–382.
4 Timmermans B, De Bodt MS, Wuyts FL, Van de Heyning PH: Training outcome in future professional voice users after 18 months of voice training. Folia Phoniatr Logop 2004; 56:120–129.
5 Timmermans B, De Bodt M, Wuyts F, Van de Heyning P: Voice quality change in future professional voice users after 9 months of voice training. Eur Arch Otorhinolaryngol 2004;261:1–5.
6 Van Lierde KM, Vinck BM, Baudonck N, De Vel E, Dhooge I: Comparison of the overall intelligibility, articulation, resonance, and voice characteristics between children using cochlear implants and those using bilateral hearing aids: a pilot study. Int J Audiol 2005; 44:452–465.
7 Van Lierde KM, Vinck B, De Ley S, Clement G, Van Cauwenberge P: Genetics of vocal quality characteristics in monozygotic twins: a multiparameter approach. J Voice 2005;19: 511–518.
8 Van Lierde KM, Claeys S, De Bodt M, Van Cauwenberge P: Vocal quality characteristics in children with cleft palate: a multiparameter approach. J Voice 2004;18:354–362.
9 Kooijman PG, de Jong FI, Oudes MJ, Huinck W, van Acht H, Graamans K: Muscular tension and body posture in relation to voice handicap and voice quality in teachers with persistent voice complaints. Folia Phoniatr Logop 2005;57:134–147.
10 Van Lierde KM, De Ley S, Clement G, De Bodt M, Van Cauwenberge P: Outcome of laryngeal manual therapy in four Dutch adults with persistent moderate-to-severe vocal hyperfunction: a pilot study. J Voice 2004;18: 467–474.
11 Van Lierde KM, Claeys S, De Bodt M, van Cauwenberge P: Long-term outcome of hyperfunctional voice disorders based on a multiparameter approach. J Voice 2007;21: 179–188.
12 Hirano M: Clinical Examination of Voice. Wien, Springer, 1981.
13 Carding PN, Steen IN, Webb A, MacKenzie K, Deary IJ, Wilson JA: The reliability and sensitivity to change of acoustic measures of voice quality. Clin Otolaryngol 2004;29: 538–544.
14 Bough ID, Heuer RJ, Sataloff RT, Hills JR, Cater JR: Intrasubject variability of objective voice measures. J Voice 1996;10:166–174.
15 Dwire A, McCauley R: Repeated measures of vocal fundamental frequency perturbation obtained using the Visi-Pitch. J Voice 1995; 9:156–162.
16 Stone RE Jr, Rainey CL: Intra- and intersubject variability in acoustic measures of normal voice. J Voice 1991;5:189–196.
17 Higgins MB, Saxman JH: A comparison of intrasubject variation across sessions of three vocal frequency perturbation indices. J Acoust Soc Am 1989;86:911–916.
18 Lee L, Stemple JC, Kizer M: Consistency of acoustic and aerodynamic measures of voice production over 28 days under various testing conditions. J Voice 1999;13:477–483.
19 Gramming P, Sundberg J, Akerlund L: Variability of phonetograms. Folia Phoniatr (Basel) 1991;43:79–92.
20 Stone RE Jr, Ferch PA: Intra-subject variability in FO-SPLmin voice profiles. J Speech Hear Disord 1982;47:134–137.
21 Sihvo M, Laippala P, Sala E: A study of repeated measures of softest and loudest phonations. J Voice 2000;14:161–169.
22 Gelfer MP: Stability in phonational frequency range. J Commun Disord 1989;22:181–192.
23 Schutte HK, Seidner W: Recommendation by the Union of European Phoniatricians (UEP): standardizing voice area measurement/phonetography. Folia Phoniatr (Basel) 1983;35:286–288.
24 Cicchetti DV: Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol Assess 1994;6:284–290.
25 Gonzalez J, Cervera T, Miralles JL: Acoustic voice analysis: reliability of a set of multidimensional parameters (in Spanish). Acta Otorrinolaringol Esp 2002;53:256–268.
26 Speyer R, Wieneke GH, Dejonckere PH: The use of acoustic parameters for the evaluation of voice therapy for dysphonic patients. Acta Acustica United Acustica 2004;90:520–527.