EVALUATION OF VOICE DISORDERS

Dysphonia Severity Index and Voice Handicap Index

Marieke Hakkesteegt

Evaluation of voice disorders, Voice Handicap Index and Dysphonia Severity Index

Thesis: Erasmus University Rotterdam, The Netherlands

Cover: Christy Renard & Marieke Hakkesteegt
Printed by Ipskamp Drukkers, Enschede (www.ipskampdrukkers.nl)
ISBN 978-90-9024566-9

EVALUATION OF VOICE DISORDERS
Dysphonia Severity Index and Voice Handicap Index

EVALUATIE VAN STEMSTOORNISSEN
Dysphonia Severity Index en Voice Handicap Index

Proefschrift

ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam
op gezag van de rector magnificus
Prof.dr. H.G. Schmidt
en volgens besluit van het College voor Promoties

De openbare verdediging zal plaatsvinden op

woensdag 7 oktober 2009 om 13.30 uur

door

Maria Margreet Hakkesteegt

geboren te Gouda

Promotiecommissie

Promotor:	Prof.dr. L. Feenstra
Overige leden:	Dr. T. van der Cammen
	Prof.dr. H.K. Schutte
	Prof.dr.ir. A.F.W. van der Steen
Copromotor:	Dr. M.H. van den Brink – Wieringa

CONTENTS

CHAPTER 1

General introduction

The voice is arguable still the most important tool of communication despite the growing importance of e-mails and text messaging (SMS) in daily contact. Indeed in modern society people are probably even more dependent on their voice than in the rural societies of old. Approximately one third of the working population nowadays need their voice to earn money[1]. And it goes without saying that most people need their voice for daily social activities. Any impairment of the voice therefore will have large impact both on daily work and social activities for many people. Diagnosis and treatment of voice disorders is by that of great importance.

The voice has a variety of characteristics and so have any of its disorders. The complaints of the patient are the starting point in any diagnosis. These complaints usually consist of 'my voice is hoarse'. They may also concern the consequences of professional activities, like a teacher whose voice does not reach the back of the classroom, especially in a noisy one. The sound of the voice may have changed and, finally, we have to pay attention to the cause of the problems, the aetiology of the voice disorder.

All these aspects need to be assessed in a description of any voice disorder. The complaints of the patient may be collected with a self-evaluation form by the patient him/herself. The character of the sound of the voice (voice quality) can be assessed subjectively with the listening ear of the diagnostician and objectively by instruments. The aetiology is determined by the thorough history taking and physical examination. The form and function of the larynx is examined with, for example, (video)laryngostroboscopy. The outcome of these examinations together leads to a diagnosis, which is the basis for decisions about any modality of treatment including intervention like surgical procedures, voice therapy, medication or any combination of these.

The results of these interventions need to be evaluated. Naturally both the patient and the therapist will be interested in the results of any type of intervention. In daily practice it is desirable to be able to compare the results of different measurements of one patient over time (as for example before and after intervention), between different investigators and between different speech centres. Moreover, for reasons of research results of different types of intervention need to be compared.

This thesis will focus on the evaluation of intervention for treatment of benign voice disorders. This evaluation is done from the point of view of a voice therapist, therefore evaluation of the anatomy and function of the larynx is excluded. We limited ourselves to voice quality measurements and self-evaluation by the patient, using existing tools.

For the subjective perceptual evaluation of voice quality, the hearing of the investigator is used. Normal hearing is perfectly capable to register different aspects of voice quality, like hoarsenss, roughness, and breathiness; however, a standardized system is needed to make judgments comparable between investigators. Several systems for perceptual evaluation are developed, like GRBAS[2], Buffalo Voice Screening Profile[3], the Darley

Rating System[4], Cape-V[5]. Of those systems the GRBAS-scale is probably the most widely used system. We used the score on G, representing the overall grade of hoarseness, which appears to be the most reliable parameter of this scale[6-8].
For objective evaluation of voice quality, several methods are used. Fundamental frequency and intensity of the voice can be measured, and the extreme possibilities of fundamental frequency and intensity together can be registered in a phonetogram (or voice range profile)[9,10]. Also several acoustic and aerodynamic measurements are used for objective evaluation, like jitter, shimmer, harmonics to noise ratio, maximum phonation time, etcetera. Albeit no general accepted way to objectively assess the voice quality as yet exists. It appears that multiparametric measurements, combining several objective parameters, are better to assess the voice quality than single parameter measurements[11-18]. We used the Dysphonia Severity Index (DSI)[17], since it is such a multi-parameter objective measurement. Another advantage of the DSI is that the parameters can be obtained relatively quickly and easily by speech pathologists, which makes it applicable in daily practice.
For the self-evaluation by the patient several questionnaires are available[19-21]. Most of those are in English. For this thesis the Voice Handicap Index (VHI)[22] was chosen, as it is worldwide frequently used in research and a Dutch version of the VHI was already available[23]. The VHI can be completed easily by most patients and is therefore applicable in daily practice.

GRBAS

The 'GRBAS scale' was introduced by Hirano in 1981[2]. The parameters of this scale are Grade (overall grade of hoarseness), Roughness, Breathiness, Asthenia and Strain. For each parameter, a four-point scale is used to indicate the severity: '0' equals normal, '1' slight, '2' moderate and '3' severe.

DYSPHONIA SEVERITY INDEX (DSI)

The DSI was developed by Wuyts et al. in 2000[17]. The DSI is derived from a multivariate analysis of 387 subjects with the very goal to describe voice quality within objective terms after instrumental analysis. The classification of the severity of dysphonia was based on the perceptual assessment, which was scored on Grade from the GRBAS-scale[2]. The parameters used for the DSI are:

- Highest fundamental frequency ('F_0-high' in hertz)
- Lowest intensity ('I-low' in decibel Sound Pressure Level)
- Maximum phonation time ('MPT' in seconds)
- Jitter (percentage short-term variability in fundamental frequency)

The DSI is constructed as:
DSI = 0.13 x MPT + 0.0053 x F_0-High - 0.26 x I-low - 1.18 x Jitter (%) + 12.4.
It is constructed in such a way 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).

VOICE HANDICAP INDEX (VHI)

The VHI was introduced by Jacobson in 1997[22]. It consists of thirty items, divided into three subscales of ten items each. The subscales are “Functional” (F), “Emotional” (E) and “Physical” (P). The score for each item ranges from 0-4, resulting in a maximum total score of 120. The higher the score is, the more serious the voice related problem is. Patients can complete the VHI-form within approximately 5 minutes.

THE OVERALL SCOPE OF THIS THESIS

To analyse the clinical applicability of two existing voice measurements:

- The Dysphonia Severity Index: an objective voice quality measurement
- The Voice Handicap Index: a patient based questionnaire

In order to answer these questions several studies were performed.
Chapter two describes a literature search about the knowledge of the intra-subject reproducibility of voice measurements. The types of outcome measurements were the subjective perceptual evaluation, the objective voice quality measurements and patient-based questionnaires (self-evaluation).
Chapter three describes the inter-observer variability and the intra-subject reproducibility of the DSI.
Chapter four describes the reproducibility of the Dutch version of the VHI.
Chapter five investigates the relationship between the perceptual evaluation of voice quality and the DSI, since the perceptual evaluation is often viewed upon as the ‘gold standard’ and since perceptual and objective evaluations have the same end.
Chapter six investigates the effect of gender and aging on the DSI. The parameters of DSI might be different between males and females - especially the highest frequencies and the maximum phonation time - and also age has its effects on the voice.
Chapter seven investigates the applicability of the DSI and the VHI for evaluating the effects of intervention.
Chapter eight finally discusses the results of all studies taken together.

REFERENCES

1. Vilkman E. Voice problems at work: A challenge for occupational safety and health arrangement. *Folia Phoniatr Logop* 2000;52:120-125.
2. Hirano M. *Clinical examination of voice*. Wien; New York: Springer, 1981.
3. Wilson DK. *Voice problems of children*. Baltimore: Williams & Wilkins, 1987.
4. Darley FL, Aronson AE, Brown JR. Differential diagnostic patterns of dysarthria. *J Speech Hear Res* 1969;12:246-269.
5. Kempster GB, Gerratt BR, Verdolini Abbott K, Barkmeier-Kraemer J, Hillman RE. Consensus auditory-perceptual evaluation of voice: development of a standardized clinical protocol. *Am J Speech Lang Pathol* 2009;18:124-132.
6. De Bodt MS, Wuyts FL, Van de Heyning PH, Croux C. Test-retest study of the GRBAS scale: influence of experience and professional background on perceptual rating of voice quality. *J Voice* 1997;11:74-80.
7. Dejonckere PH, Obbens C, de Moor GM, Wieneke GH. Perceptual evaluation of dysphonia: reliability and relevance. *Folia Phoniatr (Basel)* 1993;45:76-83.
8. Kreiman J, Gerratt BR. Validity of rating scale measures of voice quality. *J Acoust Soc Am* 1998;104:1598-1608.
9. Damste PH. The phonetogram. *Practica oto-rhino-laryngologica* 1970;32:185-187.
10. Schutte HK, Seidner W. Recommendation by the Union of European Phoniatricians (UEP): standardizing voice area measurement/phonetography. *Folia Phoniatr (Basel)* 1983;35:286-288.
11. Klein S, Piccirillo JF, Painter C. Student Research Award 1999: comparative contrast of voice measurements. *Otolaryngol Head Neck Surg* 2000;123:164-169.
12. Michaelis D, Frohlich M, Strube HW. Selection and combination of acoustic features for the description of pathologic voices. *J Acoust Soc Am* 1998;103:1628-1639.
13. Piccirillo JF, Painter C, Fuller D, Fredrickson JM. Multivariate analysis of objective vocal function. *Ann Otol Rhinol Laryngol* 1998;107:107-112.
14. Piccirillo JF, Painter C, Fuller D, Haiduk A, Fredrickson JM. Assessment of two objective voice function indices. *Ann Otol Rhinol Laryngol* 1998;107:396-400.
15. Yu P, Ouaknine M, Revis J, Giovanni A. Objective voice analysis for dysphonic patients: a multiparametric protocol including acoustic and aerodynamic measurements. *J Voice* 2001;15:529-542.

16. Yu P, Revis J, Wuyts FL, Zanaret M, Giovanni A. Correlation of instrumental voice evaluation with perceptual voice analysis using a modified visual analog scale. *Folia Phoniatr Logop* 2002;54:271-281.

17. Wuyts FL, De Bodt MS, Molenberghs Get 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.

18. Hartl DA, Hans S, Vaissiere J, Brasnu DA. Objective acoustic and aerodynamic measures of breathiness in paralytic dysphonia. *Eur Arch Otorhinolaryngol* 2003;260:175-182.

19. Hogikyan ND, Sethuraman G. Validation of an instrument to measure voice-related quality of life (V-RQOL). *J Voice* 1999;13:557-569.

20. Gliklich RE, Glovsky RM, Montgomery WW. Validation of a voice outcome survey for unilateral vocal cord paralysis. *Otolaryngol Head Neck Surg* 1999;120:153-158.

21. Wilson JA, Webb A, Carding PN, Steen IN, MacKenzie K, Deary IJ. The Voice Symptom Scale (VoiSS) and the Vocal Handicap Index (VHI): a comparison of structure and content. *Clin Otolaryngol* 2004;29:169-174.

22. Jacobson BH, Johnson A, Grywalski Cet al. The Voice Handicap Index (VHI): Development and Validation. *American Journal of Speech-Language Pathology* 1997;6:66-70.

23. De Bodt M, Jacobson B, Musschoot Set al. De Voice Handicap Index, een instrument voor het kwantificeren van de psychosociale consequenties van stemstoornissen. *Logopedie* 2000;13:29-33.

CHAPTER 2

Reproducibility of voice measurements: a review of the literature

Hakkesteegt MM, Wieringa MH, Brocaar MP, Feenstra L

Submitted

INTRODUCTION

Evidence based practice within voice therapy is still young. So far only a limited number of studies about the effects of voice therapy has been published[1]. However in today's healthcare evidence based practice has been introduced and became even standard occasionally.

To evaluate voice over time the following can be studied:

- The effects of different types of voice therapy.
- The effects of one specific therapy between two (or more) different speech centres.
- The effects of some therapy for an individual patient by the speech therapist in daily practice. Both the therapist and the patient will be interested in the effects of therapy on the quality of the voice.

As the results of voice therapy generally take a long time and daily changes are therefore not discernable, a reliable and consistent evaluation is highly desirable.

For voice disorders there is as yet no consensus about the exact outcome parameters to be used, apart of the consensus that voice measurements should be multidimensional. Clinical assessment for diagnosis and evaluation of voice disorders should consist of:

- Examination of the form and function of the vocal folds (with (video)laryngostroboscopy)
- Assessment of the characteristics of the sound of the voice (the voice quality) by means of both perceptual voice assessment and objective measurements (acoustic analysis, aerodynamic measurements)
- Evaluation of the complaints of the patient.

In this review of the literature we have chosen to focus on the evaluation of the patient's complaints and on the voice quality. The complaints of the patient may be taken together through a self-evaluation questionnaire to be filled in by the patient personally. For investigation of the voice quality both a perceptual (subjective) evaluation and an objective voice quality measurement system may be used.

Applicability of the various outcome measurements depends on the reproducibility of the investigated parameters. It is important to know the reproducibility of the various parameters that are studied. This reproducibility depends on the following:

- Inter-observer variability, i.e. the variablitiy in the judgement of different observers.
- Intra-observer variability, i.e. the variability of one observer over time.
- Intra-subject variability, i.e. the variability in a patient's voice quality at different points in time. This may be due to emotional fluctuations, fatigue et cetera, resulting in different outcomes in subjective parameters, in objective voice quality evaluations and in questionnaires taken at different moments in time.

This chapter gives an overview of published literature on the reproducibility of the following voice measurements: subjective perceptual evaluation, objective voice quality measurements and patient-based questionnaires (self-evaluation).

METHOD

A search in PubMed was performed (1950 - august 2008) with the search terms
(voice disorders OR voice quality) AND (Reproducibility of Results OR retest OR Variability OR variation OR repeatability OR intrasubject OR intra-subject)
In PubMed the search is 'translated' (the 'query translation' as is shown in the 'details' field). In this translation the search terms were extended beyond our search goal. Therefore the terms in the details field were adapted to our search goal, resulting in:
(("voice disorders"[MeSH Terms] OR "voice disorders"[All Fields]) OR ("voice quality"[MeSH Terms] OR "voice quality"[All Fields])) AND (("reproducibility of results"[MeSH Terms] OR "reproducibility"[All Fields] OR "reproducibility of results"[All Fields]) OR retest[All Fields] OR Variability[All Fields] OR repeatability[All Fields] OR intrasubject[All Fields] OR intra-subject[All Fields])

Inclusion criteria:

- Article in English
- Concerning voice quality measurements in humans (perceptual or objective), or patient-based questionnaires regarding voice.
- Articles concerning test-retest research. Test-retest was defined as: the test was repeated at two or more different moments in time (multiple sessions).

Exclusion criteria:

- No data collection (comment, review).
- Measurements repeated within one session
- Publication of the authors of this review

RESULTS

By the search we received 368 references. These references were checked for inclusion and exclusion criteria.

Number of included articles:	N=	60
Concerning subjective perceptual evaluation	N=	38
Concerning objective voice quality measurements	N=	9
Concerning patient-based questionnaires	N=	12
Concerning subjective evaluation AND patient-based questionnaires	N=	1

The references of the included articles concerning objective voice quality measurements and concerning patient-based questionnaires were checked for articles fulfilling the inclusion criteria, which were not found in the PubMed search.
This produced 6 more articles:

Concerning objective voice quality measurements	N= 3
Concerning patient-based questionnaires	N= 3
Total number of included articles:	N= 66
Total subjective perceptual evaluation (table 1)	N= 38
Total objective voice quality measurements (table 2)	N= 12
Total patient-based questionnaires (table 3)	N= 15
Total subjective evaluation AND patient-based questionnaires (tables 1 and 3)	N= 1

SUBJECTIVE PERCEPTUAL EVALUATION:

A total of 39 references published in a period of 23 years (1986 – 2008) about reliability of perceptual evaluation met the inclusion and exclusion criteria (table 1). However, in none of the studies the intra-subject variability was investigated. All studies investigated the inter-observer reliability, and in 32 studies the intra-observer reliability was investigated as well.
To improve the inter- and intra-observer reliability, in 10 studies the voice samples were judged in comparison with another sample. This comparison was sometimes with a standard reference sample, referred to as 'anchor sample'. This 'anchor sample' could be of a selected natural voice[2-4], or a synthesized sample[2,5-7]. In some studies the different samples were compared with each other[5,8-10]. In one study two different samples of one subject were compared[11].
Many different scales were investigated. The Grade-Roughness-Breathiness-Asthenia-Strain (GRBAS) scale (or parts of it), as introduced in 1981 by Hirano[12], was investigated in 24 studies. The original scoring on a four-point categorical scale was used in seven studies[13-19]. The (partial) GRBAS on a categorical scale with five points or more was used in twelve studies[4,6-10,15,20-24]. A Visual Analogue scale was used in ten studies[2,4-6,22,23,25-28]. In five studies different scoring systems for the (partial) GRBAS were compared [4-6,22,23].
In nineteen studies all subjects had voice disorders, in seventeen studies both subjects with and without voice disorders participated, in one study only subjects without voice disorders participated[29]. In two studies the samples were synthetic[30,31].
There is a large variation in the selected instruments. In some cases, a new instrument was developed because the existing scales appeared not suitable for specific groups of patients (for example spasmodic dysphonia[26,32]).

OBJECTIVE VOICE QUALITY MEASUREMENTS:

The twelve references that met the inclusion and exclusion criteria were published in a period of sixteen years (1989 – 2004, table 2). Apparently, objective voice quality measurements have been studied since at least approximately twenty years; however, only twelve studies regarding test-retest reliability could be retrieved.
The studied parameters are quite diverse, although perturbation measures were studied relatively often.
Most studies had fairly general conclusions concerning the measurement ('reliable' or 'good correlation between two measurements' for example)[33-42]. Two studies made conclusions about the used equipment or method[39,43]. In one study a conclusion regarding the significant difference for clinical application and interpretation was drawn[44]. Influence of different observers/investigators was investigated in one study[36].
In most studies the subjects were people without voice disorders, in only one study subjects with voice disorders were studied[43], and in one study both patients and subjects without voice disorders were studied[33].

PATIENT-BASED QUESTIONNAIRES:

A total of sixteen references published in a period of eleven years (1997 – 2007) about reliability of patient-based questionnaire met the inclusion and exclusion criteria (table 3).
In those sixteen studies thirteen different questionnaires were studied; sometimes several questionnaires were investigated in one study. The Voice Handicap Index (VHI), introduced by Jacobson in 1997[45] was used most of the times: in five studies the original version of the VHI was used[45-49], in two an adapted version of the VHI was used (singing VHI[50] and paediatric VHI[51]).
In most studies the subjects were patients, but in two studies the subjects were both patients and controls[49,52]. In one study, only people without voice complaints participated as subjects[53].
In the results of the test-retest studies, usually correlations were provided. The conclusions in these studies were fairly general, like 'the questionnaire is reliable'.
In three studies the significant intra-subject difference was computed (in two studies for the VHI[45,46] and one study for the singing-VHI [50]).

DISCUSSION

For evaluation of therapy, measurements are taken on two or more points in time, generally before and after therapy. When differences between those two measurements are found, it may be the result of the therapy, but it can also be due to the variability of the measurements.

The test-retest reliability of voice quality measurements is influenced by three different aspects: inter-observer, intra-observer and intra-subject variability (as described in the introduction, vide supra).
Only when the influence of all these factors is known, reliable conclusions about the results of some kind of therapy are possible.
Reproducibility of voice quality measurements was generally investigated for perceptual evaluation. For several perceptual evaluation systems intra- and inter- observer reliability was examined. The intra-subject reliability was never verified.
Within the studies about objective voice quality evaluation, reliability has been searched for several instruments; however this was not always specified as intra-subject or intra-observer variability and inter-observer variability was hardly ever checked.
For patient-based questionnaires, the intra-subject reliability was investigated in all test-retest studies. Of course intra- and inter- observer reliability are not applicable. However, the conclusions of the test-retest studies about questionnaires were mostly quite general about the reliability, and the intra-subject variation (standard deviation) was not determined.

SUBJECTIVE PERCEPTUAL EVALUATION

Apart of the inter-observer, intra-observer and intra-subject variability, also the used evaluation system, the used scoring system, and the voice sample (reading, spontaneous speech, vowel) might influence the reproducibility of perceptual evaluation.
In literature, many different evaluation systems and many different types of scoring scales are used. Comparison between these studies is not really possible.
Most studies on subjective perceptual evaluation study the reliability of some scale, both inter-observer and intra-observer. The reliability is often moderate at best. Attempts are occasionally made to improve the reliability by using other scoring systems. Since each observer has his own ‘internal standard’, it is logical to replace this by an external standard (an ’anchor’ voice sample as a standard reference). These investigations using such a standard reference conclude that the reliability thereby improves. Intra-subject variability was never investigated.

OBJECTIVE VOICE QUALITY MEASUREMENTS

Only twelve test-retest studies for objective voice quality measurements were found. The low number of this kind of studies is probably due to the lack of consensus on the measurements that are suitable to define the overall voice quality. This lack of consensus could also be the reason that many different parameters have been used. The conclusions about the test-retest reliability were diverse; some measurements were considered applicable, others found the reliability to be moderate or poor.

Not much attention has been paid to inter-observer differences, probably because the measurements are considered as relatively objective. However, they are dependent on the interaction between the observer and the patient, therefore inter-observer differences might occur. In daily practice two measurements on two different points of time in one patient will often be performed by two different observers. Therefore it is important to know the inter-observer difference.

PATIENT-BASED QUESTIONNAIRES

In the sixteen articles, thirteen different instruments were studied; all of them concluded that their instrument is reliable. Usually only correlation coefficients were computed. A correlation is the relation between two variables; it gives no information about the difference between two measurements on two different points of time in one patient caused by daily fluctuations. Such information is especially needed for the interpretation of the results within one patient, for example before and after therapy. In only three studies such intra-subject difference was computed. It is remarkable that this is not computed more often, since it is relatively easy to do for questionnaires.

CONCLUSION

There is only a limited amount of available literature on reproducibility of voice measurements and even less on intra-subject variability.
More attention should be paid to the intra-subject variability in studies of voice measurements.

Table 1 Subjective perceptual evaluation of the voice

Reference	Measurement	Score	Intra-subject	Intra-observer	Inter-observer	Conclusion in article (regular my entry, italics the author)
Comparative						
Chan, K.M. and E.M. Yiu, 2002 [2]	Roughness Breathiness	VAS[a]	-	+	+	About method: *the anchor and training effects shown in the present study support the notion that internal standards of pathological voice qualities are not stable and that both training and the provision of anchors are required to stabilize these internal standards. The study also shows that synthesized signals are useful anchors.*
Eadie, T.L. and P.C. Doyle, 2002 [3]	Naturalness Severity	9 pt DME[b]	-	+	+	Reliability: *The present data suggest that naturalness and overall severity, if assessed appropriately, may provide reliable and valid clinical measures of TE speakers.*
Granqvist, S., 2003 [5]	Breathiness Hoarseness High-frequency component	VAS[a] VSR[c]	-	+	+	Reliability: *the VSR method gave better reliability of the subjects' ratings in the single-parameter tests: Pearson and Spearman correlation coefficients were significantly higher for the VSR method than for the other methods. For the multi-parameter, intra-subject test, significantly higher Pearson correlation coefficients were found for the VSR method than for the VAS on paper.*
Kreiman, J., B.R. Gerratt, and G.S. Berke, 1994 [8]	Breathiness Roughness	7 pt	-	+	+	Reliability: *Differences in how listeners focus their attention on the different aspects of multidimensional perceptual qualities apparently are a significant source of interrater unreliability (noise) in voice quality ratings.*
Shrivastav, R., 2006 [9]	Breathiness	7 pt 5 pt	-	+	+	Reliability: *Good interjudge reliability for the perceptual ratings.* *Individual differences in the perception of voice quality are not as large as previously thought, a model of voice quality perception for an "average" listener may be a good representation for the general population.*
Webb, M., C.D. Starr, and K. Moller, 1992 [11]	Voice Quality Nasality	11 pt	-	+	+	Reliability: *The perceptual rating procedures used in this study did not produce highly reliable measurements*
Wolfe, V.I., D.P. Martin, and C.I. Palmer, 2000 [10]	Abnormality Roughness Breathiness	2 pt 7 pt	-	+	+	Reliability: *Adequate interlistener reliability for each of the three listening tasks.* *Intralistener: These values were comparable to those obtained in other studies*
Yiu EM, and Ng CY. 2004 [4]	Roughness Breathiness	11pt VAS[a]	-	+	+	Reliability: *Intra-rater agreement moderate, significantly higher in EAI, inter-rater similar in EAI and VAS*

Table 1 Subjective perceptual evaluation of the voice

Reference	Measurement	Score	Intra-subject	Intra-observer	Inter-observer	Conclusion in article (regular my entry, italics the author)
Kreiman, J., B.R. Gerratt, and M. Ito, 2007 [6]	Breathiness	VAS[a] 8 pt	-	-	+	About method: *Interrater variability is an issue of task design, not of listener unreliability.*
Yiu, E.M., K.M. Chan, and R.S. Mok, 2007 [7]	Roughness Breathiness	8 pt 8 pt	-	-	+	About method: *These results suggest that the paired comparison paradigm may be used as an alternative perceptual voice quality evaluation tool.*
Other						
Bassich, C.J. and C.L. Ludlow, 1986 [54]	Darley rating system	7 pt	-	+	+	Reliability: *Despite the extensive training procedures used, our reliability data were not comparable to those reported when highly experienced judges have been used, suggesting that the task of perceptually rating voice quality is difficult and requires extensive professional experience.*
Bele, I.V., 2005 [29]	16 parameters	VAS [a]	-	+	+	Reliability: *The results indicated a high interrater reliability for most perceptual characteristics. Experienced listeners tended to be more consistent in their ratings than did the student raters.*
Bergan, C.C., I.R. Titze, and B. Story, 2004 [30]	Ring Pressed	10 pt	-	+	+	*Reliability:* Although musicians appeared to be a little better in judging ring quality than pressed quality, there was no overall significant difference between musicians and nonmusicians with regard to intrasubject reliability or consistency of their rating of these voice qualities.
Chhetri, D.K., et al., 2008 [55]	Severity ADSD[d]	VAS [a]	-	+	+	Reliability: *voice experts are reliably able to judge and agree on the severity of ADSD.*
Damrose, J.F., et al., 2004 [56]	Overall Symptom Severity	7 pt	-	+	+	Reliability: *Overall, these results suggest good reliability among the judges regardless of the level of their experience*
De Bodt, M.S., et al., 1997 [19]	GRBAS[e]	4 pt	-	+	+	Reliability: *The results of this study show that test-retest reliability of the GRBAS scale is moderate. The G parameter seems to be the most reliable.*
de Krom, G., 1994 [20]	GIRBAS[f]	10 pt	-	+	+	Reliability: *The results of this experiment imply that grade, roughness, and breathiness ratings of connected speech fragments are not necessarily more consistent or more reliable than ratings of vowel-type fragments, even if these are as short as 200 msec.*

Table 1 Subjective perceptual evaluation of the voice

Reference	Measurement	Score	Intra-subject	Intra-observer	Inter-observer	Conclusion in article (regular my entry, italics the author)
Dejonckere, P.H., et al., 1996 [25]	GIRBAS[6]	VAS [a]	-	+	+	Reliability: *The reliability of the GIRBAS appears as very satisfactory, is recommended for widespread clinical use. Concordance is highest for G.*
Erickson, M.L., 2003 [32]	ADSD[4] signs	Count	-	+	+	Reliability: *For the ADSD group: interrater reliability was 92%, and intrarater 90%. For the control group, interrater reliability was 99% and intrarater 97%.*
Eskenazi, L., D.G. Childers, and D.M. Hicks, 1990 [21]	Overall severity, Hoarseness, Breathiness Roughness, Vocal fry Excellence of normal voice	7 pt	-	+	+	Reliability: *We believe this indicated a good intrajudge consistency, given the difficulty of the listening tasks. The listeners tended to agree for the pathological listening tasks and disagreed on the evaluation of "normal" voices.*
Feijoo, S. and C. Hernandez, 1990 [18]	Grade Breathiness	4 pt	-	+	+	Reliability: *Well-trained listeners are quite capable of performing reliable evaluation of voice quality.*
Halberstam, B., 2004 [57]	Severity of hoarseness	7 pt	-	+	+	About method: *The present study indicates that perception of hoarseness may be more valid for continuous speech than sustained vowels.*
Heuer, R., et al., 2000 [58]	Hard glottal attack	2 pt	-	+	+	Reliability: *Because of its reliability, readability, and ease of use, we feel that the Towne-Heuer passage is a valuable tool for the assessment of voice.*
Karnell, M.P., et al., 2007 [17]	Grade CAPE-V[g]	4 pt VAS[a]	-	+	+	Reliability: *Clinician's perceptions of dysphonia appeared to be reliable and unaffected by rating tool (…). TheCAPE-V system appeared to be more sensitive to small differences within and among patients than the GRBAS system.*
Kreiman J, et al., 1993 [23]	Roughness	7 pt VAS[a]	-	+	+	Reliability: *Our results suggest that traditional voice rating methods may never generate ratings that consistently meet strict standards for reliability. However, new rating protocols may be developed to control some of the sources of variability in listeners' perceptions of vocal quality.*
Kreiman J, Gerratt BR. 1998 [22]	Roughness severity, similarity	7 pt VAS[a]	-	+	+	Applicability: *The low levels of listener agreement reported here indicate that traditional protocols for assessing qualities like breathiness and roughness are not useful for measuring perceived vocal quality.*

Table 1 Subjective perceptual evaluation of the voice

Reference	Measurement	Score	Intra-subject	Intra-observer	Inter-observer	Conclusion in article (regular my entry, italics the author)
Langeveld, T.P., et al., 2000 [26]	GRBAS[5] + Aphonia, Diplophonia, Staccato, Tremor, Falsetto, Vocal Fry	VAS[a]	-	+	+	Applicability: *The extended GRBAS system can be used for accurate and reproducible perceptual characterization of ADSD.*
Lee, M., et al., 2005 [16]	GRBAS[5]	4 pt	-	+	+ [h]	Reliability: *Patients appear to have good validity and consistency using GRBAS as a self-perception tool. Validity measured in terms of agreement with clinician ratings is poor.*
Ma, E.P. and E.M. Yiu, 2001 [27] *	Grade Roughness Breathiness	VAS[a]	-	+	+	Reliability: *One intrajudge correlation coefficient for breathiness rating did not reach a statistically significant level. All the other coefficients were significant).*
Morsomme, D., et al., 2001 [28]	GIRBAS[6]	VAS[a]	-	+	+	Reliability: *The intra- and inter- judge agreements were fairly good, at least for control subjects.*
Munoz, J., et al., 2002 [59]	Buffalo voice screening profile	5 pt	-	+	+	Reliability: *It may be concluded that the BuffaloVSP is a reliable protocol for evaluation of normal and pathological voice for both types of speech sample for Spanish speakers.*
Titze IR et al. 2003 [31]	"twang" and "yawn"	10 pt	-	+	+	About judges: *None regarding test-retest (non-musicians greater variability than musicians)*
van der Torn, M., et al., 2002 [60]	Communicative suitability Voice quality	10 pt 7 pt	-	+	+	Reliability: *The concept of measuring listener judgments of "communicative suitability of voice" appears to be basically sound. Ratings are reliable and raters can discriminate between groups of normal and pathological voices.*
Webb, A.L., et al., 2004 [15]	Buffalo Voice Profile VPA[i] GRBAS[5]	5 pt 3 pt 4 pt	-	+	+	Reliability: *BuffaloVP: only overall grade reliable, VPA: reliability poor to moderate. GRBAS: reliable across all parameters except S. GRBAS is a simple reliable measure for clinical use.*
Martens, J.W., et al., 2007 [14]	GIRBAS[6] "visible speech"	4 pt	-	-	+	Reliability: *This study shows that the addition of visible speech clearly increases the interrater reproducibility for the 3 main parameters grade, roughness, and breathiness. Consequently, it enhances the reliability and relevance of perceptual evaluation, justifying widespread use.*

* Also in table 3, patient based questionnaires

Table 1 Subjective perceptual evaluation of the voice

Reference	Measurement	Score	Intra-subject	Intra-observer	Inter-observer	Conclusion in article (regular my entry, italics the author)
Moerman, M., et al., 2006 [61]	INFVo[j]	VAS[a]	-	-	+	Reliability: *The INFVo seems to constitute a reliable tool for the perceptual assessment of substitution voices. Interjudge agreement is good for semi-professionals and excellent for professionals.*
Murry, T., et al., 2004 [13]	GRBAS[5]	4 pt	-	-	+	Reliability: *The two clinicians were found to be reliable in their ratings of severity using the GRBAS scale*
Shrivastav R, et al., 2005 [24]	Breathiness	5 pt	-	-	+	About method: *In general, the probability of exact agreement and the interjudge reliability can be improved by (a) averaging multiple ratings of each stimulus by each listener and (b) using standardized scores instead of actual ratings, particularly when comparing data across listeners.*
Stewart, C.F., et al., 1997 [62]	USDRS[k]	7 pt 2 pt	-	-	+	Applicability: *The USDRS is a good tool for quantifying the voice symptoms of patients with adductor spasmodic dysphonia.*

[a] VAS: Visual Analogue Scale
[b] DME: Direct Magnitude Estimation
[c] VSR: Visual Sort and Rate
[d] ADSD: Adductor Spasmodic Dysphonia
[e] GRBAS: Grade Roughness Breathiness Asthenic Strained
[f] GIRBAS: Grade Instability Roughness Breathiness Asthenic Strained
[g] CAPE-V: Consensus Auditory Perceptual Evaluation-Voice
[h] Inter-judge: between patient and clinician.
[i] VPA: Vocal Profile Analysis scheme
[j] INFVo: Impression Intelligibility Noise Fluency Voicing
[k] USDRS: Unified Spasmodic Dysphonia Rating Scale

Table 2 Objective voice measurements

Reference	N Repeated P=Patients C=Controls	Time Interval between measurements	Measurements	Statistical analysis	Results	Conclusion about test-retest (regular my entry, italics the author)
Bough, I.D., Jr., et al., 1996 [43]	P: 14	15 samples in 15 days	Fo[a] Jitter Shimmer HNR[b] Perturbation	Intraclass Correlation Coefficient	Intra-day 0.041 – 0.999 Inter-day 0.106 – 0.994	About reliability of equipment: *The usefulness of objective voice measures depends upon the reliability and validity of the instruments used to obtain them.*
Carding, P.N., et al., 2004 [33]	P: 50 C: 50	2 hours	Jitter Shimmer NHR[c]	Intraclass Correlation Coefficient	P: 0.33 – 0.46 C: 0.68 – 0.73	Measurement is not clinical applicable *Poor to moderate reliability.*
Dwire A, McCauley R. 1995 [34]	C: 49	1 week	RAP[d]	Pearson correlation coefficient	Male: 0.83-0.90 Female 0.36-0.58	About variability of measurements: *The relationship of between session values for individual subjects was quite marked for men, but only moderate for women.*
Gelfer MP. 1989 [35]	C: 20	3 hours/ 2 months	Phonational frequency range	Analysis of Variance Mean individual variation with standard deviation	Variation: F0_low: 2.15 st F0_high: 2.5 st F0_range: 3.6 st	About variability of measurements: *All measures varied by at least several semitones. Further research is needed in order to determine the sources of this intrasubject variability before these measures can be applied to clinical populations.*
Gramming, P., et al., 1991 [36]	C: 2	15x in 3 weeks (2 investigators)	Phonetogram	Mean standard deviation (soft and loud) Correlation coefficient	F: 2.7 and 2.4 dB M: 3.0 and 3.4 dB Soft: 0.998, Loud: 0.985	Measurement is clinical applicable *The influence of 2 different investigators was found to be very small. The day-to-day or long-term voice SPL variation was approximately 3 dB in soft and loud phonation. In view of these comparatively small effects the phonetogram seems to be a useful tool for clinical purposes*
Higgins MB, Saxman JH. 1989 [37]	C: 15	17x in 33 days	JF[e], PPQ[f] , DPF[g]	Coefficient of variation	Male: 11-46 Female: 31-62	About variability of measurements: *Considarable variation across sessions.*

Table 2 Objective voice measurements

Reference	N Repeated P=Patients C=Controls	Time Interval between measurements	Measurements	Statistical analysis	Results	Conclusion about test-retest (regular my entry, italics the author)
Higgins, M.B., et al., 1994 [44]	C: 21	4x in 2 weeks	Airflow, intra-oral air pressure EGG signal[h] etc.	Coefficient of variation	Intra-subject average: 2 – 22	Conclusion about significant clinical differences for repeated measurements: *Cutoff values that represent the amount of variability that would be considered within normal limits are given.*
Kania RE, et al., 2004 [38]	C: 5	not mentioned	GCQ[i]	Analysis of variance for repeated measures	effect of time of measurement not significant	About variability of measurements: *The variability of GCQ with Fo and intensity contributes to reliability and standardization problems with GCQ as a measure of glottic closure.*
Lee L, et al., 1999 [39]	C: 70	28 days	Fo[a], jitter, MPT[j], phonation volume, flow rate	Analysis of Variance	Results are presented for all parameters	About influences on variability of measurements: *Maintaining a constant intensity and frequency level across repeated tests helps ensure similarities in measurement condition. When these variables are held constant, changes in acoustic and aerodynamic measures over time can be more confidently attributed to management of the underlying disorder.*
Mendoza E, et al., 1996 [40]	C: 17	5x in 2 wks	LTAS[k]	Unifactorial Analysis of Variance	Absolute values: diff between sessions significant Relative values: not sign	About variability of measurements: *The fact that the differences between sessions disappear when employing relative measures may indicate that the utilization of these measures eliminates the sources of systematic or aleatoric error can be introduced during a recording or in the period of time between two consecutive recording sessions.*

Table 2 Objective voice measurements

Reference	N Repeated P=Patients C=Controls	Time Interval between measurements	Measurements	Statistical analysis	Results	Conclusion about test-retest (regular my entry, italics the author)
Stassen HH, 1991 [41]	C: 187	14 days	Spectral patterns (overtone distribution)	Reproducibility	Location and intensity of the first 3 formants: exactly reproduced Bandwith and height of formants: highly reproducible.	About variability of measurements: *The reference values derived of the 187 healthy subjects of this study enabled us to distinguish between "natural" and "significant" fluctuations.*
Stone RE, Jr., Rainey CL. 1991 [42]	C: 8	8 wks	Fo[a] + standard deviation dBA[l] + standard deviation Jitter	Correlation coefficient	0.13 – 0.82	About variability of measurements: *The variability shown by our subjects across days is discouraging. Refining the tools of vocal testing should precede their application for evaluation of intervention.*

[a] Fo Fundamental frequency
[b] HNR Harmonics-to-Noise Ratio
[c] NHR Noise-to-Harmonics Ratio
[d] RAP Relative Average Perturbation
[e] JF Jitter Factor
[f] PPQ Pitch Perturbation Quotient
[g] DPF Directional Perturbation Factor
[h] EGG Electroglottograph signal
[i] GCQ Glottal Closed Quotient
[j] MPT Maximum Phonation Time
[k] LTAS Long-term Average Spectrum
[l] dBA Decibel A-weighting

Table 3 Patient based questionnaires

Reference	N Repeated P=Patients C=Controls	Time Interval between measurements	Measurements	Statistical analysis	Results	Conclusion about test-retest (regular my entry, italics the author)
Bach, K.K., et al., 2005 [49]	P: 40 C: 40	3 mnd (P) 2 wks (C)	VHI[a] GFI[b]	Pearson Correlation Coefficient	0.56	About reliability of instrument: *The GFI demonstrates excellent test-retest reliability, comparable to that of the VHI*
Boseley, M.E., et al., 2006 [52]	P+C: 70	2 wks	P-VRQOL[c]	Weighted kappa Cronbach's α (internal consistency)	0.80 0.96	About reliability of instrument: *P-VRQOL is a valid instrument.*
Cohen, S.M., et al., 2007 [50]	P: 66	17.3 days	S-VHI[d]	Spearman Correlation Coefficient Cronbach's α (internal consistency) Critical difference: 95%confidence interval for difference of mean SVHI.	0.92 0.97 14 points	About significant clinical difference for repeated measures: *The SVHI is a reliable and valid tool.* *Critical difference is 14 points (in results).*
Fang, T.J., et al., 2007 [63]	P: 40	2-4 wks	VOS[e] (Chinese version)	Spearman Correlation Coefficient	0.59	About reliability of instrument: *The moderate reliability of CVOS reaches the level for a valid survey and is comparable with original VOS validation data*
Gliklich, R.E., et al., 1999 [64]	P: 61	2 wks	VOS[e]	Correlation coefficient	0.87	About reliability of instrument: *The test-retest reliability for the total VOS score was excellent*
Hartnick, et al., 2003 [53]	C: 75	2 wks	P-VOS[f]	Weighted kappa	0.89	About reliability of instrument: *The PVOS represents a valid and reliable instrument*
Hogikyan ND, Sethuraman G. 1999 [65]	P: 68	7 days (mean, sd 6)	V-RQOL[g]	Pearson Correlation Coefficient	0.93	About reliability of instrument: *Both measures of reliability, the internal consistency and reproducibility, are strong for the domain or total scores.*

Table 3 Patient based questionnaires

Reference	N Repeated P=Patients C=Controls	Time Interval between measurements	Measurements	Statistical analysis	Results	Conclusion about test-retest (regular my entry, italics the author)
Jacobson BH, et al. 1997 [45]	P: 63	29 days (mean)	VHI[a]	Pearson Correlation Coefficient 95% Confidence interval (critical difference)	0.92 18 points	About significant clinical difference for repeated measures: *A shift in the total score of 18 points or greater is required in order to be certain that the change is not due to unexplained variability inherent in the VHI.*
Kazi R, 2006 [66]	P: 29	4 wks	VPQ[h]	Reliability Proportion test	median of 77.5%	About reliability of instrument: *>75% of patients had a score on retest that was within one point of their original score, indicating that reliability was therefore high*
Lam, P.K., et al., 2006 [48]	P: 68	2 wks – 2 months	VHI[a]	Pearson Correlation Coefficient Cronbach's α (internal consistency)	VHI-30: 0.84 VHI-10: 0.83 VHI-30: 0.98 VHI-10: 0.95	About reliability of instrument: *High test-retest reliability and internal consistency for both Chinese VHI-30 and VHI-10.*
Ma, E.P. and E.M. Yiu, 2001 [27] *	P: 20	2 wks	VAPP[i]	Pearson Correlation Coefficient	0.86	About reliability of instrument: *The test-retest reliability of the VAPP is high enough to be considered a reliable tool for measuring such changes as the result of an intervention.*
Ma, E.P. and E.M. Yiu, 2007 [67]	P: 18	2 wks	Voice AL[j] Voice PR[k]	Pearson Correlation Coefficient Mean item score difference	EAI 0.74, VAS 0.76 EAI 0.06, sd 2.40 VAS –0.13, sd 2.09	About reliability of scoring system: *Either EAI or VA scaling would be considered as appropriate for measuring voice AL and PR.*
van Gogh CD, et al. 2005 [68]	P: 104	5.3 months	5-item screening questionnaire	Intraclass Correlation Coefficient	0.67 – 0.76	About reliability of instrument: *Reliability proved to be good.*

* Also in table 1, subjective evaluation

Table 3 Patient based questionnaires

Reference	N Repeated P=Patients C=Contros	Time Interval between measurements	Measurements	Statistical analysis	Results	Conclusion about test-retest (regular my entry, italics the author)
Van Gogh, C.D., et al., 2007 [46]	P: 30	3.5 months	VHI[a]	Spearman Correlation Coefficient Individuals: maximum differences between test and retest scores For groups: based on Effect Size of 0.80	0.95 Individual difference: -9 – 10 points 15 points	About significant clinical difference for repeated measures: *Clinical relevant difference for individuals 10 points, for groups 15 points.*
Webb, A.L., et al., 2007 [47]	P: 50	1 wk	VHI[a] VPQ[l] VoiSS[m]	Intraclass Correlation Coefficient	VHI : 0.83 VPQ: 0.75 Voiss: 0.63	About reliability of instrument: *VHI, VPQ and VoiSS are reliable and valid instruments*
Zur, K.B., et al., 2007 [51]	P: 10	1 wk	P-VHI[n]	Pearson Correlation Coefficient	0.82	About reliability of instrument: *The pVHI provided a high internal consistency and test-retest reliability.*

[a] VHI Voice Handicap Index
[b] GFI Glottal Function Index
[c] P-VRQOL Pediatric Voice-Related Quality-of-Life survey
[d] S-VHI Singing Voice Handicap Index
[e] VOS Voice Outcome Survey
[f] P-VOS Pediatric Voice Outcome Survey
[g] V-RQOL Voice Related Quality-of-Life survey
[h] VPQ Voice Prosthesis Questionnaire
[i] VAPP Voice Activity and Participation Profile
[j] Voice AL Voice Activity Limitation
[k] Voice PR Voice Participation Restriction
[l] VPQ Vocal Performance Questionnaire
[m] VoiSS Voice Symptom Scale
[n] P-VHI Pediatric Voice Handicap Index

REFERENCES

1. Speyer R. Effects of Voice Therapy: A Systematic Review. *J Voice* 2007.
2. Chan KM, Yiu EM. The effect of anchors and training on the reliability of perceptual voice evaluation. *J Speech Lang Hear Res* 2002;45:111-126.
3. Eadie TL, Doyle PC. Direct magnitude estimation and interval scaling of naturalness and severity in tracheoesophageal (TE) speakers. *J Speech Lang Hear Res* 2002;45:1088-1096.
4. Yiu EM, Ng CY. Equal appearing interval and visual analogue scaling of perceptual roughness and breathiness. *Clin Linguist Phon* 2004;18:211-229.
5. Granqvist S. The visual sort and rate method for perceptual evaluation in listening tests. *Logoped Phoniatr Vocol* 2003;28:109-116.
6. Kreiman J, Gerratt BR, Ito M. When and why listeners disagree in voice quality assessment tasks. *J Acoust Soc Am* 2007;122:2354-2364.
7. Yiu EM, Chan KM, Mok RS. Reliability and confidence in using a paired comparison paradigm in perceptual voice quality evaluation. *Clin Linguist Phon* 2007;21:129-145.
8. Kreiman J, Gerratt BR, Berke GS. The multidimensional nature of pathologic vocal quality. *J Acoust Soc Am* 1994;96:1291-1302.
9. Shrivastav R. Multidimensional scaling of breathy voice quality: individual differences in perception. *J Voice* 2006;20:211-222.
10. Wolfe VI, Martin DP, Palmer CI. Perception of dysphonic voice quality by naive listeners. *J Speech Lang Hear Res* 2000;43:697-705.
11. Webb M, Starr CD, Moller K. Effects of extended speaking on resonance of patients with cleft palate. *Cleft Palate Craniofac J* 1992;29:22-26.
12. Hirano M. *Clinical examination of voice*. Wien; New York: Springer, 1981.
13. Murry T, Medrado R, Hogikyan ND, Aviv JE. The relationship between ratings of voice quality and quality of life measures. *J Voice* 2004;18:183-192.
14. Martens JW, Versnel H, Dejonckere PH. The effect of visible speech in the perceptual rating of pathological voices. *Arch Otolaryngol Head Neck Surg* 2007;133:178-185.
15. Webb AL, Carding PN, Deary IJ, MacKenzie K, Steen N, Wilson JA. The reliability of three perceptual evaluation scales for dysphonia. *Eur Arch Otorhinolaryngol* 2004;261:429-434.
16. Lee M, Drinnan M, Carding P. The reliability and validity of patient self-rating of their own voice quality. *Clin Otolaryngol* 2005;30:357-361.

17. Karnell MP, Melton SD, Childes JM, Coleman TC, Dailey SA, Hoffman HT. Reliability of clinician-based (GRBAS and CAPE-V) and patient-based (V-RQOL and IPVI) documentation of voice disorders. *J Voice* 2007;21:576-590.

18. Feijoo S, Hernandez C. Short-term stability measures for the evaluation of vocal quality. *J Speech Hear Res* 1990;33:324-334.

19. De Bodt MS, Wuyts FL, Van de Heyning PH, Croux C. Test-retest study of the GRBAS scale: influence of experience and professional background on perceptual rating of voice quality. *J Voice* 1997;11:74-80.

20. de Krom G. Consistency and reliability of voice quality ratings for different types of speech fragments. *J Speech Hear Res* 1994;37:985-1000.

21. Eskenazi L, Childers DG, Hicks DM. Acoustic correlates of vocal quality. *J Speech Hear Res* 1990;33:298-306.

22. Kreiman J, Gerratt BR. Validity of rating scale measures of voice quality. *J Acoust Soc Am* 1998;104:1598-1608.

23. Kreiman J, Gerratt BR, Kempster GB, Erman A, Berke GS. Perceptual evaluation of voice quality: review, tutorial, and a framework for future research. *J Speech Hear Res* 1993;36:21-40.

24. Shrivastav R, Sapienza CM, Nandur V. Application of psychometric theory to the measurement of voice quality using rating scales. *J Speech Lang Hear Res* 2005;48:323-335.

25. Dejonckere PH, Remacle M, Fresnel-Elbaz E, Woisard V, Crevier-Buchman L, Millet B. Differentiated perceptual evaluation of pathological voice quality: reliability and correlations with acoustic measurements. *Rev Laryngol Otol Rhinol (Bord)* 1996;117:219-224.

26. Langeveld TP, Drost HA, Frijns JH, Zwinderman AH, Baatenburg de Jong RJ. Perceptual characteristics of adductor spasmodic dysphonia. *Ann Otol Rhinol Laryngol* 2000;109:741-748.

27. Ma EP, Yiu EM. Voice activity and participation profile: assessing the impact of voice disorders on daily activities. *J Speech Lang Hear Res* 2001;44:511-524.

28. Morsomme D, Jamart J, Wery C, Giovanni A, Remacle M. Comparison between the GIRBAS Scale and the Acoustic and Aerodynamic Measures Provided by EVA for the Assessment of Dysphonia following Unilateral Vocal Fold Paralysis. *Folia Phoniatr Logop* 2001;53:317-325.

29. Bele IV. Reliability in perceptual analysis of voice quality. *J Voice* 2005;19:555-573.

30. Bergan CC, Titze IR, Story B. The perception of two vocal qualities in a synthesized vocal utterance: ring and pressed voice. *J Voice* 2004;18:305-317.

31. Titze IR, Bergan CC, Hunter EJ, Story B. Source and filter adjustments affecting the perception of the vocal qualities twang and yawn. *Logoped Phoniatr Vocol* 2003;28:147-155.

32. Erickson ML. Effects of voicing and syntactic complexity on sign expression in adductor spasmodic dysphonia. *Am J Speech Lang Pathol* 2003;12:416-424.

33. Carding PN, Steen IN, Webb A, MacKenzie K, Deary IJ, Wilson JA. The reliability and sensitivity to change of acoustic measures of voice quality. *Clinical otolaryngology and allied sciences* 2004;29:538-544.

34. Dwire A, McCauley R. Repeated measures of vocal fundamental frequency perturbation obtained using the Visi-Pitch. *J Voice* 1995;9:156-162.

35. Gelfer MP. Stability in phonational frequency range. *J Commun Disord* 1989;22:181-192.

36. Gramming P, Sundberg J, Akerlund L. Variability of phonetograms. *Folia Phoniatr (Basel)* 1991;43:79-92.

37. 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.

38. Kania RE, Hans S, Hartl DM, Clement P, Crevier-Buchman L, Brasnu DF. Variability of electroglottographic glottal closed quotients: necessity of standardization to obtain normative values. *Arch Otolaryngol Head Neck Surg* 2004;130:349-352.

39. 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.

40. Mendoza E, Munoz J, Valencia Naranjo N. The long-term average spectrum as a measure of voice stability. *Folia Phoniatr Logop* 1996;48:57-64.

41. Stassen HH. Affective state and voice: the specific properties of overtone distributions. *Methods Inf Med* 1991;30:44-52.

42. Stone RE, Jr., Rainey CL. Intra- and Intersubject Variability in Acoustic Measures of Normal Voice. *J Voice* 1991;5:189-196.

43. Bough ID, Jr., Heuer RJ, Sataloff RT, Hills JR, Cater JR. Intrasubject variability of objective voice measures. *J Voice* 1996;10:166-174.

44. Higgins MB, Netsell R, Schulte L. Aerodynamic and electroglottographic measures of normal voice production: intrasubject variability within and across sessions. *J Speech Hear Res* 1994;37:38-45.

45. Jacobson BH, Johnson A, Grywalski Cet al. The Voice Handicap Index (VHI): Development and Validation. *American Journal of Speech-Language Pathology* 1997;6:66-70.

46. Van Gogh CD, Mahieu HF, Kuik DJ, Rinkel RN, Langendijk JA, Verdonck-de Leeuw IM. Voice in early glottic cancer compared to benign voice pathology. *Eur Arch Otorhinolaryngol* 2007;264:1033-1038.

47. Webb AL, Carding PN, Deary IJ, MacKenzie K, Steen IN, Wilson JA. Optimising outcome assessment of voice interventions, I: Reliability and validity of three self-reported scales. *J Laryngol Otol* 2007;121:763-767.

48. Lam PK, Chan KM, Ho WK, Kwong E, Yiu EM, Wei WI. Cross-cultural adaptation and validation of the Chinese Voice Handicap Index-10. *Laryngoscope* 2006;116:1192-1198.

49. Bach KK, Belafsky PC, Wasylik K, Postma GN, Koufman JA. Validity and reliability of the glottal function index. *Arch Otolaryngol Head Neck Surg* 2005;131:961-964.

50. Cohen SM, Jacobson BH, Garrett CGet al. Creation and validation of the Singing Voice Handicap Index. *Ann Otol Rhinol Laryngol* 2007;116:402-406.

51. Zur KB, Cotton S, Kelchner L, Baker S, Weinrich B, Lee L. Pediatric Voice Handicap Index (pVHI): a new tool for evaluating pediatric dysphonia. *Int J Pediatr Otorhinolaryngol* 2007;71:77-82.

52. Boseley ME, Cunningham MJ, Volk MS, Hartnick CJ. Validation of the Pediatric Voice-Related Quality-of-Life survey. *Arch Otolaryngol Head Neck Surg* 2006;132:717-720.

53. Hartnick CJ, Volk M, Cunningham M. Establishing normative voice-related quality of life scores within the pediatric otolaryngology population. *Arch Otolaryngol Head Neck Surg* 2003;129:1090-1093.

54. Bassich CJ, Ludlow CL. The use of perceptual methods by new clinicians for assessing voice quality. *J Speech Hear Disord* 1986;51:125-133.

55. Chhetri DK, Merati AL, Blumin JH, Sulica L, Damrose EJ, Tsai VW. Reliability of the perceptual evaluation of adductor spasmodic dysphonia. *Ann Otol Rhinol Laryngol* 2008;117:159-165.

56. Damrose JF, Goldman SN, Groessl EJ, Orloff LA. The impact of long-term botulinum toxin injections on symptom severity in patients with spasmodic dysphonia. *J Voice* 2004;18:415-422.

57. Halberstam B. Acoustic and perceptual parameters relating to connected speech are more reliable measures of hoarseness than parameters relating to sustained vowels. *ORL J Otorhinolaryngol Relat Spec* 2004;66:70-73.

58. Heuer R, Towne C, Hockstein NE, Andrade DF, Sataloff RT. The Towne-Heuer Reading Passage--a reliable aid to the evaluation of voice. *J Voice* 2000;14:236-239.

59. Munoz J, Mendoza E, Fresneda MD, Carballo G, Ramirez I. Perceptual analysis in different voice samples: agreement and reliability. *Percept Mot Skills* 2002;94:1187-1195.

60. van der Torn M, Verdonck-de Leeuw IM, Kuik DJ, Mahieu HF. Communicative suitability of voice following radiotherapy for T1 glottic carcinoma: testing the reliability of a rating instrument. *J Voice* 2002;16:398-407.

61. Moerman M, Martens JP, Crevier-Buchman Let al. The INFVo perceptual rating scale for substitution voicing: development and reliability. *Eur Arch Otorhinolaryngol* 2006;263:435-439.

62. Stewart CF, Allen EL, Tureen P, Diamond BE, Blitzer A, Brin MF. Adductor spasmodic dysphonia: standard evaluation of symptoms and severity. *J Voice* 1997;11:95-103.

63. Fang TJ, Li HY, Gliklich RE, Chen YH, Wang PC. Assessment of Chinese-version voice outcome survey in patients with unilateral vocal cord paralysis. *Otolaryngol Head Neck Surg* 2007;136:752-756.

64. Gliklich RE, Glovsky RM, Montgomery WW. Validation of a voice outcome survey for unilateral vocal cord paralysis. *Otolaryngol Head Neck Surg* 1999;120:153-158.

65. Hogikyan ND, Sethuraman G. Validation of an instrument to measure voice-related quality of life (V-RQOL). *J Voice* 1999;13:557-569.

66. Kazi R, Singh A, De Cordova Jet al. Validation of a voice prosthesis questionnaire to assess valved speech and its related issues in patients following total laryngectomy. *Clin Otolaryngol* 2006;31:404-410.

67. Ma EP, Yiu EM. Scaling voice activity limitation and participation restriction in dysphonic individuals. *Folia Phoniatr Logop* 2007;59:74-82.

68. van Gogh CD, Verdonck-de Leeuw IM, Boon-Kamma BA, Langendijk JA, Kuik DJ, Mahieu HF. A screening questionnaire for voice problems after treatment of early glottic cancer. *International journal of radiation oncology, biology, physics* 2005;62:700-705.

CHAPTER 3

The interobserver and test-retest variability of the Dysphonia Severity Index

Hakkesteegt MM, Wieringa MH, Brocaar MP, Mulder PG, Feenstra L.

Folia Phoniatr Logop 2008;60:86-90

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 three different days, with an interval of one week.
RESULTS:
The difference in DSI between two observers (interobserver difference) is 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 using 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), lowest intensity (I-low in dB SPL), maximum phonation time (MPT in s) and jitter (%). The DSI is constructed as DSI = 0.13 x MPT + 0.0053 x F_0-High - 0.26 x I-low - 1.18 x 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 Multi-Speech 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 maximum phonation time. 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, maximum phonation time 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 three times; the longest measured phonation time in seconds was used.

Jitter

The subjects phonated three times an /a/ at a comfortable pitch and loudness during approximately three seconds. The jitter was calculated on a sample of one second, 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 one 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 (1^{st}, 2^{nd}, 3^{rd}) 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 inter-subject 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 (inter-observer variance + residual variance). The SD of the difference between two duplicate measurements to the same subject equals $\sigma_{error} * \sqrt{2}$.

RESULTS

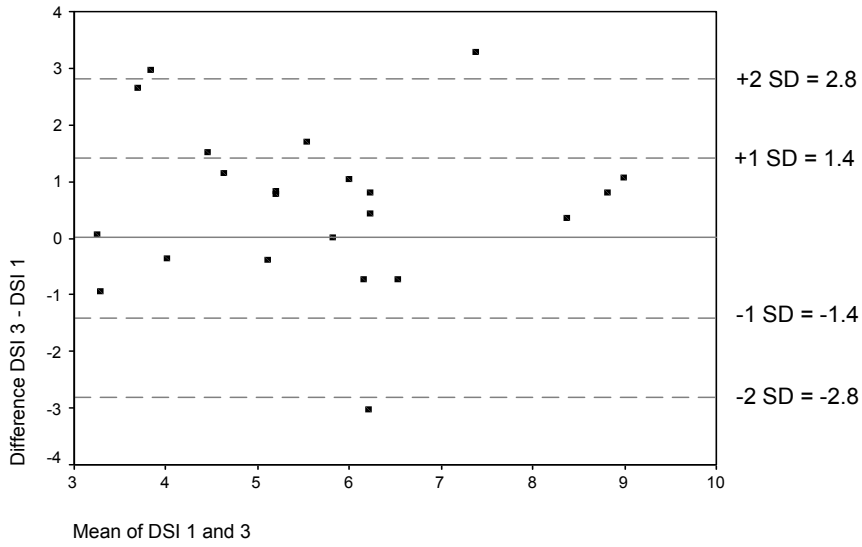
Of all 30 subjects, 22 completed 3 measurements (13 females, 9 males) and 8 subjects completed 2 measurements (6 females, 2 males). Five of those dropouts were measured twice by the same speech pathologist (schedule 1), three were measured by two speech pathologists (schedule 2).
The mean time interval 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).

Table 1. Mean values of DSI, F_0-High, I-low, MPT and jitter on measurement 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

The total variance of the DSI was 3.92. The intersubject variance was 3.11, the inter-observer 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 σ^2_{error} = 0.21 (interobserver variance) + 0.60 (residual variance) = 0.81.
Standard error of measurement σ_{error} = $\sqrt{0.81}$ = 0.90. The SD of the difference between two duplicate measurements by different observers is $\sigma_{error} * \sqrt{2} = 0.90 * \sqrt{2} = 1.27$.

Figure 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 standard deviation of the difference between DSI 1 and 3.



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 one-week interval, by two speech pathologists. Eight subjects did not complete the three measurements planned. Drop out 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. It is neither clear whether a similar effect will be present in patients. We chose a time interval of one 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 one week, and usually will be at least three 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 three 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 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 inter-observer 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 Carding et al. as well [13]. 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 ICC's 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 a 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 within one patient of 2.49 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 Pet 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 Get 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 FLet 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* 2006.
12. Hirano M. *Clinical examination of voice*. Wien; New York: 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. *Psychological Assessment* 1994;Vol 6:284-290.

25. Gonzalez J, Cervera T, Miralles JL. [Acoustic voice analysis: reliability of a set of multi-dimensional parameters]. *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 with Acustica* 2004;90:520-527.

CHAPTER 4

Reproducibility of the Dutch version of the Voice Handicap Index

Hakkesteegt MM, Wieringa MH, Gerritsma EJ, Feenstra L

Folia Phoniatr Logop 2006;58:132-138

ABSTRACT

OBJECTIVE:
To investigate the reproducibility of the Dutch translation of the Voice Handicap Index (VHI) by performing a test-retest study. To determine the relationship between the test-retest differences and the VHI scores.
PATIENTS AND METHODS:
A group of 104 patients (56 male, 48 female) with voice complaints completed the Voice Handicap Index twice with a mean interval of 13 days.
RESULTS:
There were no differences in scores between male and female patients. There was a good correlation between the first and the second measurement (r = 0.95). The difference between two VHI scores of any one patient and the total score were not related. We found that a 14-point difference in total score of the Dutch VHI for one patient at two points in time is significant.
CONCLUSION:
The Dutch version of the Voice Handicap Index has a good reproducibility. When used for clinical evaluation studies of patients with voice complaints, a difference of 14 points between two measurements is significant.

INTRODUCTION

During the past decades patient-based measuring instruments regarding quality of life and perceived handicap have become important in health care. Self-administered questionnaires are used to assess quality of life. Specifically for voice disorders several questionnaires are available: the Voice Related Quality of Life Instrument (VRQOL) [1-3], the Voice Symptoms Scale (VoiSS) [4] and the Voice Handicap Index (VHI) [4-8]. In 2001 the European Laryngological Society indicated the VHI as an important instrument for the assessment of voice disorders [9]. The VHI was introduced by Jacobson et al. [7] and consists of 30 items, divided into 3 subscales of 10 items each. These subscales concern 'Functional' (F), 'Emotional' (E) and 'Physical' (P) items. The score for each item ranges from 0-4, resulting in a maximum total score of 120. The higher the score, the larger the voice related problem. Jacobson et al.[7] tested the reproducibility and the internal consistency of the VHI, which appeared to be good. Various publications [10-14] reported that the VHI is clinically applicable and can be used to evaluate effectiveness of various therapies[15-23]. The VHI questionnaire has been translated into Flemish-Dutch [5].
The purpose of this study was to investigate the reproducibility of the translated version of the original VHI. The test-retest-variability was measured and the relationship of the variability to absolute score was determined.

PATIENTS AND METHODS:

All patients with voice related complaints who had an appointment at our department between November 2003 and April 2004 were sent the Dutch VHI questionnaire by post. They were asked to fill in this VHI-form approximately two weeks before their appointment. They noted the date of completion on the form. On the day of their appointment they immediately handed over this form to the receptionist. At that time they received a new VHI-form to complete in the waiting room (without having access to the form they had filled in previously) and subsequently returned it to the receptionist.
The Dutch translation of the VHI [5] was used, with some minor adjustments allowing for differences between Netherlands-Dutch and Flemish-Dutch. The sequence of the questions was not altered. As in the original VHI a five-point scale was used (0 = never and 4 = always). Patients had to check the box of the response that indicates how frequently they have this experience.
A time interval of approximately 14 days was chosen because this was estimated to be short enough so that not much was likely to have changed in the patients complaints, and long enough for the patient not to remember his or her previous responses.
Data was analysed anonymously.

STATISTICS

For the statistical analysis the SPSS statistical program release 10.1 for Windows was used. If the distribution of the VHI scores was normal, a T-test was used for comparison of the VHI-scores between males and females and a paired T-test for comparison of the sub-scores. In case of non-normality of the data, non-parametric tests were performed. For the total score and the sub-scores a Pearson Correlation coefficient was computed. For further interpretation of the reproducibility a Bland-Altman plot was made for the total VHI score [24]. Probability values of less than 0.05 were considered significant.

RESULTS

One hundred and four patients (56 male, 48 female) with a mean age of 51 years completed two VHI-forms. The median time interval between completion of the first and the second form was 13 days (90% between 5 and 24 days).
In table 1 the diagnoses of the patients are shown. Twenty-eight patients had non-organic dysphonia, 27 vocal fold paresis/paralysis, 20 chronic laryngitis or Reinke's oedema. Cysts/polyps and nodules were relatively rare in this study population, with 9 and 5 patients respectively.

Table 1 Diagnostics of the patients (N=104)

Diagnosis	N
Non-organic dysphonia	28
Unilateral vocal fold paralysis	23
Laryngitis / oedema	20
Cyst / polyp	9
Nodules	5
Bilateral vocal fold paralysis	4
Contact ulcer granuloma	2
Miscellaneous	13

Table 2 Mean and range of the VHI-scores in females, males and total group

	Females	Males	Total	P-values of difference females-males
VHI-total, mean (range)	41 (0-92)	43 (0-101)	42 (0-101)	0.65
VHI-F, mean (range)	10 (0-32)	13 (0-38)	12 (0-38)	0.08
VHI-E, mean (range)	13 (0-33)	13 (0-31)	13 (0-33)	0.96
VHI-P, mean (range)	18 (0-32)	17 (0-36)	18 (0-36)*	0.63

Number of participants: female participants, N=48 (46%); male participants, N=58 (54%); total, N=104.
Score significantly higher than scores on VHI-F and VHI-E: * $p < 0.01$.

Table 2 shows the mean and range of the total scores and of the sub-scores of the two VHI forms for females, males and the total group. The distribution of the total score and the sub-scores was normal. The differences between the mean total scores and sub-scores of males and females were not significant ($p>0.05$). Regarding the sub-scores the mean score on P (18) was significantly higher ($p<0.01$) than the scores on F (12) and E (13).

In figure 1 the first measurement (VHI 1) is plotted against the second measurement (VHI 2). Each point represents the first score of a patient (VHI 1) on the y-axis and the second score (VHI 2) on the x-axis. Of the total score the correlation coefficient (r) was 0.95. For the sub-scores on F, E, and P the correlation coefficient was 0.93, 0.92 and 0.88 respectively (table 3).

Figure 2 shows a Bland-Altman plot of the two VHI measurements. A Bland-Altman plot shows the difference between two measurements against their mean. In such a plot a relationship between this difference and the magnitude of the score is shown. In figure 2 the y-axis represents the difference between the first and the second measurement (VHI 1 – VHI 2), the x-axis shows the mean of VHI 1 and VHI 2 (VHI-mean). The difference between the first and the second measurement seemed to be independent of the mean score.

In table 3 the mean differences between VHI 1 and VHI 2 (with standard deviation) for the total score, and for the F-, E- and P-scores are shown. For the total score the mean difference was –0.42 and the standard deviation was 7.3. Therefore a significant difference between two VHI total scores within one patient is 1.96 x 7.3 = 14 points.

Figure 1 The first measurement (VHI 1) plotted against the second measurement (VHI 2)

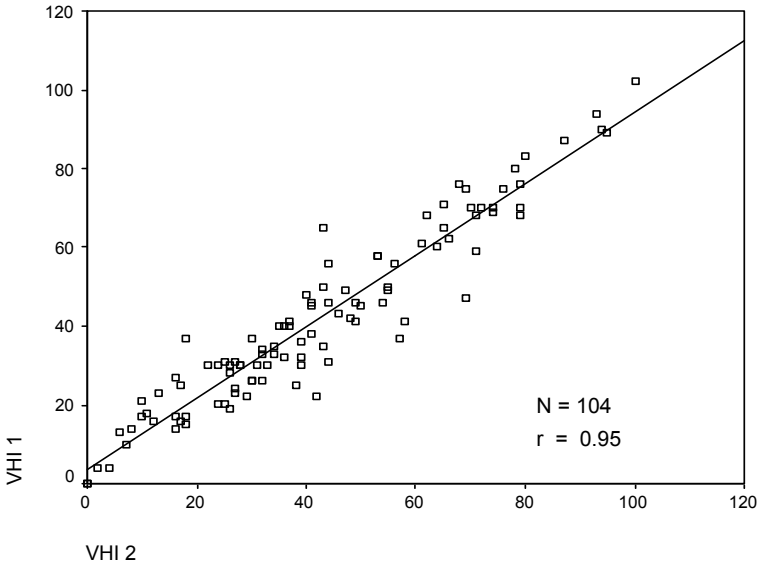


Figure 2 Bland-Altman plot: the difference between the first and the second measurement (VHI 1 – VHI 2) plotted against the mean of the first and the second measurement (VHI MEAN)

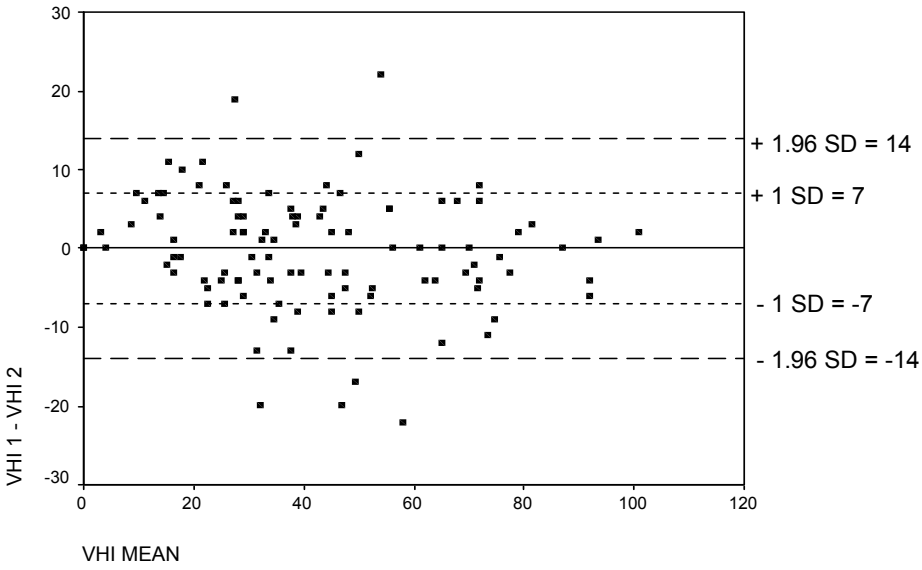


Table 3. Correlation coefficient between the first and second measurement; the mean difference between first and second measurement with accompanying Standard Deviation

	Correlation coefficient	VHI 1 – VHI 2 Mean (SD)	
VHI-total	0.95	-0.42	(7.3)
VHI-F	0.93	-0.27	(3.3)
VHI-E	0.92	0.18	(3.6)
VHI-P	0.88	-0.74	(3.9)

DISCUSSION

We investigated the reproducibility of the Dutch version of the VHI in 104 patients, which had not yet been studied.

To determine the test-retest variability a time interval between two measurements of approximately 14 days was used. With a longer time interval changes of the complaints, in laryngeal status, vocal use and environmental situation might have occurred whereas with a shorter interval the patient could more easily remember the first scoring. In the test-retest study of Jacobson et al. [7] the mean time interval was substantially longer: 29.3 days with a range from 6 till 71 days, with the disadvantages mentioned above.

In our study the mean total scores for males and females were almost equal. This corresponds to the results of Nawka et al. [8] in a group of German patients and Guimares and Abberton [6] in a group of Portuguese patients. We found that the highest mean score was measured on subscale P. The difference between the score on P and the scores on F and E was significant. A similar difference in the sub-scores is described by Speyer et al. [21] and Guimares and Abberton [6].

There was a wide range in sub-scores and total scores. This is probably due to the large heterogeneity in pathology seen in our department. Therefore the conclusions of this study can be generalized to a diverse group of patients.

Since our assumption was that the voice complaints of the patients had not changed between completing the two VHI's, both VHI scores should be the same. However small variances are still to be expected. In this study the mean differences within one patient between the two VHI scores were small for both the total score and the sub-scores (all < 1 point). The correlation coefficients of the total VHI, the VHI-F, VHI-E and VHI-P were 0.95, 0.93, 0.92 and 0.88, respectively. Jacobson et al. [7] did a test-retest study on 63 subjects (25 male, 38 female) and found correlation coefficients of 0.92, 0.84, 0.92 and 0.86. Hogikyan and Sehuraman [3] described comparable results for the Voice Related

Quality of Life instrument (VRQOL): in their test-retest study they found a correlation coefficient of 0.93. So in our study the test-retest stability appeared to be slightly better.
In this study we found a difference between the two measurements of 14 points or less (1.96*SD) in 95% of our patients (with unchanged voice complaints). This implies that a difference of more than 14 points represents a change in voice complaints. In clinical practice this means that if a patient completes the VHI twice, for example before and after therapy, a difference of at least 14 points can be interpreted as a significant change. For the subscales F, E and P differences of respectively 6, 7 and 8 points are significant. Jacobson et al. [7] found a difference of 18 points to be significant for the total score, and 8 points for the subscales F, E and P.
In this study we also investigated whether the difference between the two measurements becomes larger with a higher total score. This turned out not to be the case. This implies that the significant difference of 14 points is applicable to the entire range of total scores, without taking into account the magnitude of the total score. However, it seems that the difference is slightly more variable in the mid-range of the total scores (figure 2). This is probably due to the fact that people with minor complaints (low score) and people with major complaints (high score) often score in the extremes of the answer possibilities (0 or 4). This implies a very clear choice of answer, which might be more stable in time.
When assessing results of therapy, a clinically significant difference within one patient of 14 points can be used. This is an important finding because until now, the results of Jacobson et al. [7] (significant difference of 18 points) were used for clinical evaluation studies in The Netherlands.

CONCLUSION

The Dutch version of the Voice Handicap Index has a good reproducibility and can be used for clinical evaluation studies of patients with voice complaints. A difference of 14 points or more between two measurements within one patient indicates a change in voice complaints.

REFERENCES

1. Eggermont A, Buekers R. Kwaliteit van leven en stemkwaliteit. *Stem-, Spraak- en Taalpathologie* 2001;10:95-109.
2. Hogikyan ND, Rosen CA. A review of outcome measurements for voice disorders. *Otolaryngol Head Neck Surg* 2002;126:562-572.
3. Hogikyan ND, Sethuraman G. Validation of an instrument to measure voice-related quality of life (V-RQOL). *J Voice* 1999;13:557-569.
4. Wilson JA, Webb A, Carding PN, Steen IN, MacKenzie K, Deary IJ. The Voice Symptom Scale (VoiSS) and the Vocal Handicap Index (VHI): a comparison of structure and content. *Clin Otolaryngol* 2004;29:169-174.
5. De Bodt M, Jacobson B, Musschoot Set al. De Voice Handicap Index, een instrument voor het kwantificeren van de psychosociale consequenties van stemstoornissen. *Logopedie* 2000;13:29-33.
6. Guimaraes I, Abberton E. An investigation of the Voice Handicap Index with speakers of Portuguese: preliminary data. *J Voice* 2004;18:71-82.
7. Jacobson BH, Johnson A, Grywalski Cet al. The Voice Handicap Index (VHI): Development and Validation. *American Journal of Speech-Language Pathology* 1997;6:66-70.
8. Nawka T, Wiesmann U, Gonnermann U. [Validation of the German version of the Voice Handicap Index]. *Hno* 2003;51:921-930.
9. Dejonckere PH, Bradley P, Clemente Pet 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.
10. Benninger MS, Ahuja AS, Gardner G, Grywalski C. Assessing outcomes for dysphonic patients. *J Voice* 1998;12:540-550.
11. Moerman M, Martens JP, Dejonckere P. Application of the Voice Handicap Index in 45 patients with substitution voicing after total laryngectomy. *Eur Arch Otorhinolaryngol* 2003.
12. Murry T, Rosen CA. Outcome measurements and quality of life in voice disorders. *Otolaryngol Clin North Am* 2000;33:905-916.
13. Stewart MG, Chen AY, Stach CB. Outcomes analysis of voice and quality of life in patients with laryngeal cancer. *Arch Otolaryngol Head Neck Surg* 1998;124:143-148.
14. Verdonck-de Leeuw IM, Boon-Kamma B, Van Gogh CDL, Greven AJ, Peeters AJGE, Mahieu HF. Een multidimensioneel stemonderzoeksprotocol: het meten van stemkarakteristieken van patienten met een klein larynxcarcinoom. *Logopedie en Foniatrie* 2002;74:340-345.

15. Benninger MS, Gardner G, Grywalski C. Outcomes of botulinum toxin treatment for patients with spasmodic dysphonia. *Arch Otolaryngol Head Neck Surg* 2001;127:1083-1085.

16. Hogikyan ND, Wodchis WP, Terrell JE, Bradford CR, Esclamado RM. Voice-related quality of life (V-RQOL) following type I thyroplasty for unilateral vocal fold paralysis. *J Voice* 2000;14:378-386.

17. Peeters AJ, Van Gogh CD, Goor KM, Verdonck-De Leeuw IM, Langendijk JA, Mahieu HF. Health status and voice outcome after treatment for T1a glottic carcinoma. *Eur Arch Otorhinolaryngol* 2004.

18. Rosen CA, Murry T, Zinn A, Zullo T, Sonbolian M. Voice handicap index change following treatment of voice disorders. *J Voice* 2000;14:619-623.

19. Roy N, Weinrich B, Gray SD, Tanner K, Stemple JC, Sapienza CM. Three treatments for teachers with voice disorders: a randomized clinical trial. *J Speech Lang Hear Res* 2003;46:670-688.

20. Spector BC, Netterville JL, Billante C, Clary J, Reinisch L, Smith TL. Quality-of-life assessment in patients with unilateral vocal cord paralysis. *Otolaryngol Head Neck Surg* 2001;125:176-182.

21. Speyer R, Wieneke GH, Dejonckere PH. Self-assessment of voice therapy for chronic dysphonia. *Clin Otolaryngol* 2004;29:66-74.

22. 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.

23. 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.

24. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;1:307-310.

CHAPTER 5

The relationship between perceptual evaluation and objective multiparametric evaluation of dysphonia severity

Hakkesteegt MM, Brocaar MP, Wieringa MH, Feenstra L.

J Voice 2008;22:138-145

ABSTRACT

OBJECTIVE:
The purpose of this study was to investigate the usefulness of the Dysphonia Severity Index (DSI) as an objective multiparametric measurement in assessing dysphonia. The DSI was compared with the score on Grade of the GRBAS scale. Investigated was also whether the DSI is related to severity of dysphonia, which was represented by different diagnosis groups. Furthermore, it was investigated whether the DSI can differentiate between a group of patients and a control group.
METHOD:
A total of 294 patients with different voice pathologies were included. A control group consisted of 118 volunteers without any voice complaints. The voices of all participants were perceptually evaluated on Grade, and the DSI was measured.
RESULTS:
The groups of patients with voice complaints have a lower DSI and higher scores on Grade than the control group. The DSI was significantly lower when the score on Grade was higher. The DSI discriminates between patients with nonorganic voice disorders, vocal fold mass lesions and vocal fold paresis/paralysis. To determine whether the DSI discriminates between patients and controls, the sensitivity and specificity for different DSI cutoff points were calculated. With a DSI cut-off of 3.0 maximum sensitivity (0.72) and specificity (0.75) were found.
CONCLUSION:
We conclude that the DSI is a useful instrument to objectively measure the severity of dysphonia.

INTRODUCTION

Voice disorders have a variety of causes (organic and functional) and can lead to problems in work and social activities. According to the European Laryngeal Society, an assessment of voice disorders should consist of (video)laryngostroboscopy, perceptual voice assessment, acoustic analysis, aerodynamic measurements and subjective self-evaluation of voice [1]. Two of the advised assessment tools, the perceptual assessment and the acoustic analysis, address the voice quality. Both have specific features, and both have advantages and disadvantages.

The perceptual assessment in its most simple form is a description of the sound of the voice. This can be useful in clinical practice, but it lacks precision and is hardly useful to compare results of therapy in individuals or between groups of patients. Besides, communication between clinicians will be difficult, which is due to lack of agreement on definitions and terminology. On top of that, each clinician has his own internal standard to compare the perceived voice quality [2]. This internal standard is partly dependent on the range of severity of dysphonia a clinician is used to judge. This could very well result in different judgements between speech therapists working in private practice and in hospitals, which will complicate communication between them. To reduce these drawbacks, different scales have been introduced to score specific aspects of voice quality. The 'GRBAS scale' as introduced by Hirano [3] is widely used. The parameters of this scale are Grade, Roughness, Breathiness, Asthenia and Strain. For each parameter, a four-point scale is used to indicate the severity. The efficacy of this system has been evaluated. The interrater reliability is moderate, [2,4-6] and on the aspects Asthenia and Strain it is low. [7] The reliability of the overall score Grade is higher than of the parameters Roughness, Breathiness, Asthenia and Strain [2,6,7]. Therefore, it is suggested to use a simplified system of GRB [1]. On the other hand Hartl et al. [8] suggested that the GRBAS scale lacks detail and sensitivity and is therefore imperfect. Nevertheless in daily clinical practice, the GRBAS and similar scales are still most widely used and are the most practical for perceptual evaluation.

To improve and clarify the communication between clinicians and for standardization purposes, acoustic measurements could be used. Both acoustic measurements and perceptual assessment address voice quality. Perceptual assessment is still regarded as the 'gold standard' [9]. Therefore, acoustic measurements should be compared with the perceptual assessment. Several studies investigated the relationship between single acoustic measures and perceptual evaluation. The results of these studies are inconclusive [10-15]. However, a combination of several objective parameters seems to correlate better with perceptual analysis than single acoustic measures [8,16-22]. A disadvantage of some of these multiparametric methods is the need of specific equipment for some of the parameters used, like subglottic pressure [18] or the Lyapunov Coefficient [20]. The Dysphonia Severity Index (DSI) as proposed by Wuyts et al.[22] is also

an objective multiparametric measurement. The DSI was derived from a multivariate analysis of 387 subjects with the goal to describe the perceived voice quality, based on objective measures. The parameters used for the DSI are the highest frequency (F_0-high in Hz), lowest intensity (I-low in dBSPL), maximum phonation time (MPT in s) and jitter (%). The DSI is constructed as DSI = 0.13 x MPT + 0.0053 x F_0-High - 0.26 x I-low - 1.18 x Jitter (%) + 12.4. It is constructed so that a perceptually normal voice corresponds with a DSI of +5 and a severely dysphonic voice corresponds with a DSI of –5, but also scores beyond this range are possible (higher than +5 or lower than -5). An advantage of this DSI is that the parameters can be obtained relatively quickly and easily by speech pathologists in daily clinical practice.

To prove the usefulness of an objective measure in assessing dysphonia, the results should be compared with the perceptual assessment of patients with voice disorders. Furthermore, it would be interesting to know how well such a measure differentiates between people with and without dysphonia.

The purpose of this study was to investigate the usefulness of the DSI as an objective measurement in assessing dysphonia. The research questions were: Are the DSI and the score on Grade comparable? Is there a relationship between the DSI and the severity of dysphonia, represented by clinical diagnosis groups? Does the DSI differentiate between a patient group and a control group?

METHODS

PARTICIPANTS

Voice data of 294 patients with voice complaints who visited our department for the first time between January 2000 and February 2004 were analyzed. Patients with laryngeal cancer were excluded, because valid measurements could not be obtained. A control group consisted of 118 volunteers without any former or present voice complaints or voice disorders.

PROCEDURES

All patients visited the out-patient clinic of our department of otorhinolaryngology with voice complaints. They were examined by one of the two speech pathologists and one of two ENT specialists of the department. The speech pathologists perceptually evaluated the quality of the patients' voices on G (Grade) of the GRBAS (on a scale from 0 to 3). This was evaluated during spontaneous speech and reading of a short text. Subsequently, data recordings for the DSI-parameters were taken by the same speech pathologist. At that moment, the clinical diagnosis was still unknown. Thereafter, the clinical diagnosis was made by the ENT specialist with direct or indirect laryngoscopy.

All measurements were collected in a database. In this database, 10 different diagnoses were registered. These diagnoses were clustered into three groups, based on expected differences in severity of dysphonia: non-organic dysphonia, mass lesions (nodules, polyps, cysts, laryngitis and edema) and paresis / paralysis (unilateral and bilateral paresis and paralysis).
All volunteers of the control group visited our department for data recording. They did not have (nor have had) any serious voice complaints. They completed a questionnaire regarding former and present voice problems to confirm this. Two final-year speech-pathology students perceptually evaluated the voice quality on Grade and recorded the DSI parameters. These students were trained by the first author in performing the measurements.
All data recording took place in a room with "living room acoustics" [23].

Frequency and Intensity
The participants 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 dBSPL.

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

Jitter
The participants phonated three times an /a/ at a comfortable pitch and loudness during approximately three seconds. The percentage jitter was calculated on a sample of one second, starting half a second after the voice onset. To exclude technically invalid measurements, the lowest result of the three calculations was used.

EQUIPMENT

Recordings were made on a Sony Digital Audio Tape recorder (DTC-57ES, Sony, Tokyo, Japan) with a Sennheiser microphone (BG 2.0 dyn). The distance between mouth and microphone was 30 cm. Intensity and frequency measurements were obtained with an automatically recording phonetograph (Pabon/Laryngograph 1997). A stopwatch was used for measuring the MPT. The *Multi-Speech* program (Kay Elemetrics, Kay PENTAX, Lincoln Park NJ) was used for acoustic analysis of the sound files.

STATISTICAL ANALYSIS

For the statistical analysis, the *SPSS* statistical program release 10.1 for Windows (SPSS Inc., Chicago, IL) was used. The distributions of the DSI in the patient group and the control group were plotted. The distribution of the DSI was tested for normality with the one-sample Kolmogorov-Smirnov test. Because the DSI was not normally distributed median values and percentiles were used. To test the differences of the DSI between groups (control group versus patient group, diagnosis groups and Grade groups), the nonparametric Mann-Whitney test was used. Plots of the DSI against the G score and diagnosis groups were made. The sensitivity and specificity of the DSI were calculated.

RESULTS

The mean age of the 118 individuals of the control group (69 female, 49 male) was 44 years (range 20-79). The mean age of the 294 patients (196 female, 98 male) was 44 years (range 14-87). The number of patients in each diagnosis group is shown in Table 1.

All participants of the control group had a Grade score of 0. In the patient group, 49 (17%) patients had a score of 0, 162 (55%) had a score of 1, 51(17%) had a score of 2 and none had a score of 3. Of 32 patients (11%) the scores on Grade were missing, due to administrative causes. The missing scores were distributed over all diagnosis groups. The percentages of patients per G score within each diagnosis group are shown in Table 1.

The distributions of the DSI in the control group and the patient group are shown in Figure 1. The median DSI for the control group was 4.2, for the patient group 1.4. A Mann-Whitney test proved that the difference in DSI between the control and the patient group was significant ($P < 0.001$).

In Table 2, the median DSI with the 25^{th} and 75^{th} percentiles for the control group and the diagnosis groups are shown.

In Figure 2, the median DSI including the 25^{th} and 75^{th} percentile for each Grade score is shown, for the control group and the patient group separately. The differences of the DSI between the scores on grade were all significant ($P < 0.05$). The difference of the DSI between the control group and the patient group with Grade score of 0 is significant ($P < 0.05$).

In Figure 3, the median DSI including the 25^{th} and 75^{th} percentile for each diagnosis group is shown. For all groups the difference of DSI with the control group was significant ($P < 0.001$).

The *P*-values of the differences in DSI and G score between the diagnosis groups are shown in Table 3. All these differences were significant, except for the Grade score between the groups mass lesions and paresis/paralysis.

To determine whether the DSI discriminates between patients and controls, the sensitivity and specificity for different DSI cutoff points were calculated (Table 4). With a DSI cut-off of 3.0, maximum sensitivity (0.72) and specificity (0.75) were found. In Table 5, the sensitivity and specificity for a DSI cutoff of 3.0 between the control group and the diagnosis groups are shown.

Table 1. Grade scores within the diagnosis groups

		Grade			
		0	1	2	Missing
	N	%	%	%	%
Control group	118	100	-	-	-
Patients with:					
Non-organic dysphonia	122	25	53	7	15
Mass lesions	126	10	60	23	7
Paresis / paralysis	46	13	48	28	11
Total patient group	294	17	55	17	11

Figure 1. Distributions of DSI in the control group and the patient group

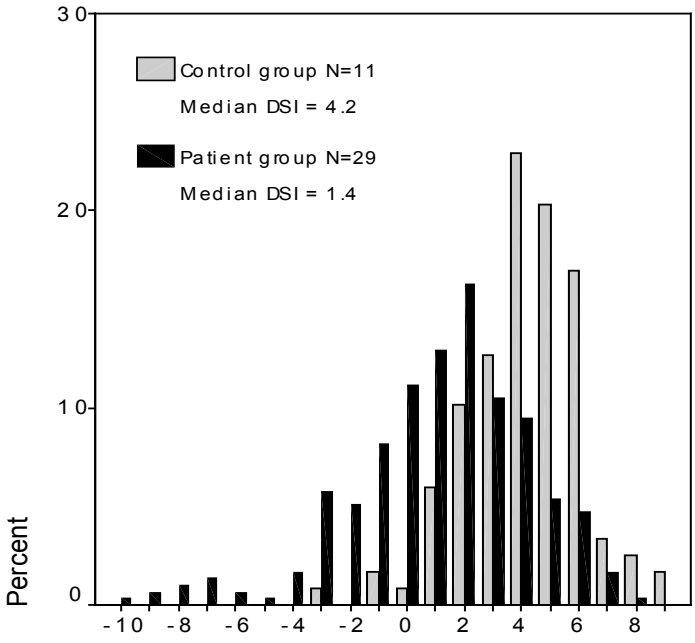
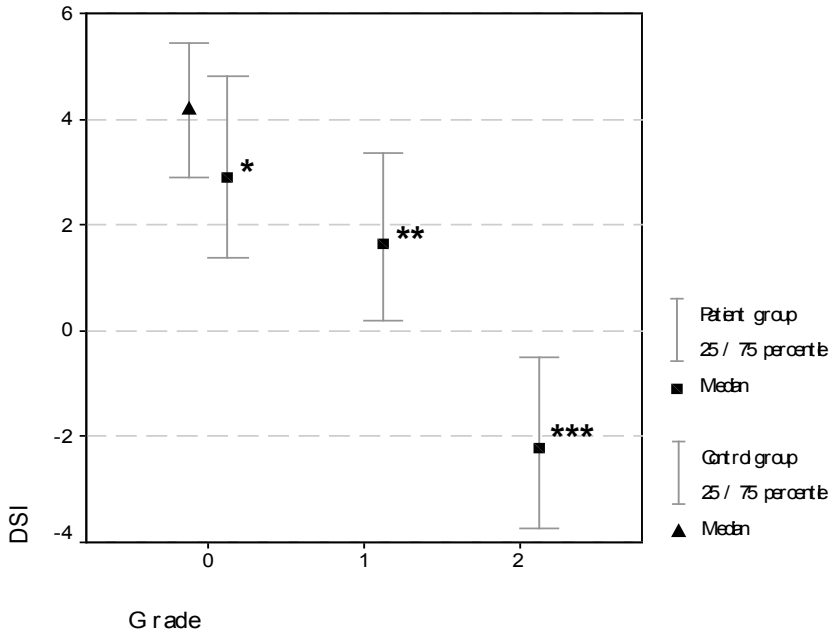


Table 2. Median and percentiles of DSI for the control group and the diagnosis groups

	DSI		
		Percentiles	
	Median	25th	75th
Control group	4.2	2.9	5.4
Patients with:			
Non-organic dysphonia	2.0*	0.1	4.0
Mass lesions	1.2*	-0.3	2.9
Paresis / paralysis	-0.7*	-2.8	1.5
Total patient group	1.4*	-0.8	3.3

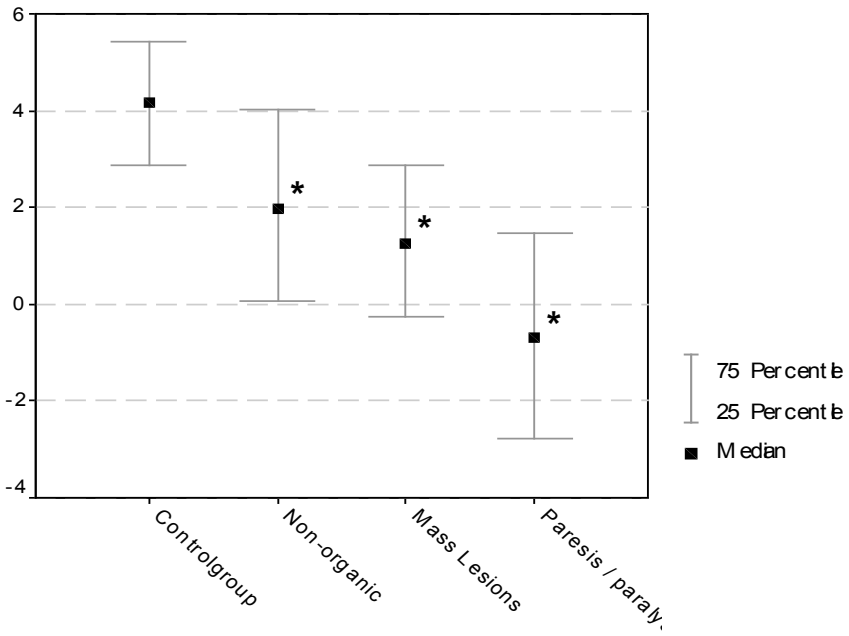
* = Difference with control group is significant (Mann-Whitney test, $p < 0.05$)

Figure 2. Median DSI with 25th and 75th percentile per G score for the patient group and the control group



* = Difference with control group is significant ($p < 0.05$)
** = Difference with patients with G0 is significant ($p < 0.05$)
*** = Difference with patients with G0 and G1 is significant ($p < 0.05$)

Figure 3. Median DSI with 25th and 75th percentile per diagnosis group



* = Difference with control group is significant ($p < 0.001$)

Table 3. P-values of differences in DSI and differences in G score between the diagnosis groups

	Non-organic dysphonia		Mass lesions		Paresis / paralysis	
	DSI	G	DSI	G	DSI	G
Control group	0.00*	0.00*	0.00*	0.00*	0.00*	0.00*
Non-organic dysphonia			0.03*	0.00*	0.00*	0.00*
Mass lesions					0.00*	0.75

* = Difference is significant (Mann-Whitney test, $p < 0.05$)

Table 4. Sensitivity and specificity of the DSI

Cut-off DSI	Sensitivity	Specificity
1.5	0.51	0.91
2.0	0.62	0.86
2.5	0.68	0.81
3.0	0.72	0.75
3.5	0.78	0.68
4.0	0.83	0.56
4.5	0.88	0.45
5.0	0.91	0.35

Table 5. Sensitivity and specificity (cut-off point 3.0) for the different patient groups by diagnosis

DSI cut-off 3.0	Sensitivity	Specificity
Control group vs. Patients with non-organic dysphonia	0.60	0.75
Control group vs. Patients with mass lesions	0.77	0.75
Control group vs. Patients with paresis/paralysis	0.93	0.75

DISCUSSION

We investigated the DSI in a group of patients with voice complaints and a group of controls without voice complaints. The DSI is a relatively simple and easy to obtain objective evaluation method for dysphonia, which can be used in daily clinical practice. For the perceptual assessment we used the Grade score of the GRBAS scale.

The analyses were not stratified for males and females. Although the highest frequency is used in the DSI, there is no difference in DSI between males and females [22,24]. The difference in highest frequency appears to be compensated by the difference in MPT between the sexes.

In the perceptual assessment, all controls in this study had a Grade score of 0. This is inherent to the fact that they were included only if they had no voice disorders. Some of the patients had also a Grade score of 0. These are possible patients who visited our

department because of, for example, work-related voice complaints, without dysphonia. The perceptual assessment of the patients showed that most of them had a Grade score of 1 or 2. In this study, a Grade score of 3 was not found, this could be due to the 'internal standard' of the investigators. The range of voice pathology presented to our phoniatrics department is very wide. Patients with good speaking voices (Grade score of 0) who have problems in singing are seen, as well as patients with a variety of postoperative sequelae after treatment for oncologic disorders who have severely dysphonic or aphonic voices. Therefore, in our department only the very severely dysphonic voices will be scored as Grade 3. Consequently, voices classified as Grade 1 or Grade 2 include a relatively wide range of dysphonic voice qualities. In this study, patients were excluded when a technically valid measurement could not be obtained. This resulted in exclusion of patients with a Grade score of 3. Especially valid frequency-related measurements, necessary for the DSI, cannot be obtained from severely dysphonic voices. The fact that the DSI cannot be used with severely dysphonic voices is a disadvantage of the DSI.

In our study, we found a significantly lower DSI with a higher perceptual score of overall severity of voice quality (Grade of the GRBAS scale). This is in concordance with the fact that the construction of the DSI is based on the G score [22]. The range of the DSI within each G score was quite large, possibly due to differences in severity of dysphonia that were not reflected in the G score. Therefore, the DSI possibly determines the severity of dysphonia in more detail than the G. It is remarkable that some patients (17%) have a G score of 0. These patients have a significantly lower DSI than the controls (a median of 2.9 versus 4.2). These patients often have good speaking voices, but do have other voice complaints, for example problems in their work situation or with their singing voice. This is apparently reflected in the DSI but not in the G score.

Several other studies showed a relationship as well between subjective and multiparametric objective evaluation of voice quality. Yu et al.[20] found a good relationship between perceptive evaluation and a combination of six acoustic and airflow measurements. Piccirillo et al.[19] found a good relationship between the GRBAS scale and their multi-parameter instrument: the Weighted Odds Ratio Index. However, these multiparametric objective evaluations contain parameters that are hard to measure in a clinical setting, like the Lyapunov coefficient and (estimated) subglottal pressure.

To identify the relationship between the DSI and the severity of dysphonia, beforehand we clustered the diagnoses as registered in our clinical database into groups of the same kind of organic disorder, although the clinical diagnosis is not a robust predictor of the severity of dysphonia. In all diagnosis groups, the median DSI was lower than in the control group. It appeared that the group of patients with non-organic dysphonia had a higher median DSI and a lower G score than the group with vocal fold mass lesions and the group of patients with paresis/paralysis. Subsequently, the group of patients with paresis/paralysis of the vocal fold(s) had a lower median DSI than the group with mass

lesions. In the latter two groups, there often is a combination of incomplete glottal closure and irregularity in vocal fold vibration. The difference in DSI could be due to the fact that the incomplete glottal closure is worse in the group with paresis/paralysis. It could be expected that the perceived severity of dysphonia would be different as well, however, there was no difference in G score. This difference between DSI and G score again suggests that the DSI might determine the severity of dysphonia in more detail than the G.

The median DSI of the control group was significantly higher than of the patient group. There was however an overlap. Some of the patients with nonorganic voice disorders have complaints of their voice related to their profession, e.g. teachers who have trouble talking loud enough, without having a dysphonic voice. On the other hand, people without voice complaints may have limitations in high frequencies or MPT, for example, while this is not causing them any problems in daily life.

To determine whether the DSI discriminates between controls and patients, the sensitivity and specificity were calculated. Depending on the aim of the clinical application, the choice for a higher sensitivity or specificity can be made. To classify people with dysphonia as dysphonic (true positives), a high sensitivity is needed. The higher the DSI cutoff is chosen, the higher the sensitivity. On the other hand, people without dysphonia should not be classified as dysphonic (false positive) and consequently treated as such. Therefore a high specificity is needed as well. With a DSI cut-off of 3.0, a maximum sensitivity (0.72) and specificity (0.75) are found. If a patient has a DSI of 3.0 or higher, this patient probably has other problems than dysphonia, like throat complaints or specific work related voice problems.

CONCLUSIONS

The DSI and the score on G are comparable. There is a relationship between the DSI and severity of dysphonia, represented by different DSI scores between groups of patients with non-organic voice disorders, vocal fold mass lesions, and paresis/paralysis. The DSI differentiates between groups of individuals with and without voice complaints. With a DSI cutoff of 3.0, the sensitivity is 0.72 and the specificity is 0.75. The DSI can be a part of the total examination of patients with voice complaints; it is useful to objectively classify the severity of dysphonia.

ACKNOWLEDGEMENTS

The authors would like to thank Hans Verschuure for his helpful comments on an earlier draft of this manuscript.

REFERENCES

1. Dejonckere PH, Bradley P, Clemente Pet 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. De Bodt MS, Wuyts FL, Van de Heyning PH, Croux C. Test-retest study of the GRBAS scale: influence of experience and professional background on perceptual rating of voice quality. *J Voice* 1997;11:74-80.
3. Hirano M. *Clinical examination of voice*. Wien; New York: Springer, 1981.
4. Dejonckere PH, Remacle M, Fresnel-Elbaz E, Woisard V, Crevier L, Millet B. Reliability and clinical relevance of perceptual evaluation of pathological voices. *Rev Laryngol Otol Rhinol (Bord)* 1998;119:247-248.
5. Webb AL, Carding PN, Deary IJ, MacKenzie K, Steen N, Wilson JA. The reliability of three perceptual evaluation scales for dysphonia. *Eur Arch Otorhinolaryngol* 2003.
6. Kreiman J, Gerratt BR. Validity of rating scale measures of voice quality. *J Acoust Soc Am* 1998;104:1598-1608.
7. Dejonckere PH, Obbens C, de Moor GM, Wieneke GH. Perceptual evaluation of dysphonia: reliability and relevance. *Folia Phoniatr (Basel)* 1993;45:76-83.
8. Hartl DA, Hans S, Vaissiere J, Brasnu DA. Objective acoustic and aerodynamic measures of breathiness in paralytic dysphonia. *Eur Arch Otorhinolaryngol* 2003;260:175-182.
9. Ma EP, Yiu EM. Multiparametric Evaluation of Dysphonic Severity. *J Voice* 2005.
10. Dejonckere PH, Lebacq J. Acoustic, perceptual, aerodynamic and anatomical correlations in voice pathology. *ORL J Otorhinolaryngol Relat Spec* 1996;58:326-332.
11. Rabinov CR, Kreiman J, Gerratt BR, Bielamowicz S. Comparing reliability of perceptual ratings of roughness and acoustic measure of jitter. *J Speech Hear Res* 1995;38:26-32.
12. Giovanni A, Revis J, Triglia JM. Objective aerodynamic and acoustic measurement of voice improvement after phonosurgery. *Laryngoscope* 1999;109:656-660.
13. Morsomme D, Jamart J, Wery C, Giovanni A, Remacle M. Comparison between the GIRBAS Scale and the Acoustic and Aerodynamic Measures Provided by EVA for the Assessment of Dysphonia following Unilateral Vocal Fold Paralysis. *Folia Phoniatr Logop* 2001;53:317-325.

14. Heman-Ackah YD, Michael DD, Goding GS, Jr. The relationship between cepstral peak prominence and selected parameters of dysphonia. *J Voice* 2002;16:20-27.

15. Heman-Ackah YD, Heuer RJ, Michael DDet al. Cepstral peak prominence: a more reliable measure of dysphonia. *Ann Otol Rhinol Laryngol* 2003;112:324-333.

16. Klein S, Piccirillo JF, Painter C. Student Research Award 1999: comparative contrast of voice measurements. *Otolaryngol Head Neck Surg* 2000;123:164-169.

17. Michaelis D, Frohlich M, Strube HW. Selection and combination of acoustic features for the description of pathologic voices. *J Acoust Soc Am* 1998;103:1628-1639.

18. Piccirillo JF, Painter C, Fuller D, Fredrickson JM. Multivariate analysis of objective vocal function. *Ann Otol Rhinol Laryngol* 1998;107:107-112.

19. Piccirillo JF, Painter C, Fuller D, Haiduk A, Fredrickson JM. Assessment of two objective voice function indices. *Ann Otol Rhinol Laryngol* 1998;107:396-400.

20. Yu P, Ouaknine M, Revis J, Giovanni A. Objective voice analysis for dysphonic patients: a multiparametric protocol including acoustic and aerodynamic measurements. *J Voice* 2001;15:529-542.

21. Yu P, Revis J, Wuyts FL, Zanaret M, Giovanni A. Correlation of instrumental voice evaluation with perceptual voice analysis using a modified visual analog scale. *Folia Phoniatr Logop* 2002;54:271-281.

22. Wuyts FL, De Bodt MS, Molenberghs Get 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.

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. Hakkesteegt MM, Brocaar MP, Wieringa MH, Feenstra L. Influence of age and gender on the dysphonia severity index. A study of normative values. *Folia Phoniatr Logop* 2006;58:264-273.

CHAPTER 6

Influence of age and gender on the Dysphonia Severity Index

A study of normative values

Hakkesteegt MM, Brocaar MP, Wieringa MH, Feenstra L.

Folia Phoniatr Logop 2006;58:264-273

ABSTRACT

OBJECTIVE:
Attempts have been made to find objective parameters to assess voice quality for many years. Objective measurements such as the Dysphonia Severity Index (DSI), using four parameters (highest frequency, lowest intensity, maximum phonation time and jitter), appear to correlate well with perceptual evaluation. The aim of this study was to investigate the influence, if any, of age and gender on the DSI.
METHOD:
The DSI of 118 non-smoking adults (69 females, 49 males, age-range 20-79 years) without voice complaints was measured.
RESULTS:
Age has a significant effect on the DSI and on its parameters highest frequency and lowest intenstiy (only in females). Gender has no effect on the DSI, although it has a significant effect on the parameters highest frequency and maximum phonation time.
CONCLUSION:
To be able to distinguish between the effects of (normal) ageing and a voice disorder, normative data of a wide age range are essential. As a result of this study normative DSI values for gender and age have been made available.

INTRODUCTION

Various parameters are used to assess voice quality objectively. Several acoustic parameters correlate with some aspects of voice quality but not with overall voice quality [1-7]. Therefore multiparametric methods have been developed, which appear to correlate better with overall voice quality [2,8-12]. One of those multiparametric methods is the Dysphonia Severity Index (DSI) developed by Wuyts et al. [10]. This index consists of a specific weighted combination of the highest fundamental frequency (F_0-high), lowest intensity (I-low), maximum phonation time (MPT) and jitter. These parameters can be obtained rather quickly and easily. The DSI is constructed as follows:

DSI = 0.13 x MPT + 0.0053 x F_0-High - 0.26 x I-low - 1.18 x Jitter (%) + 12.4

The resulting DSI-values vary between 5 (corresponding to no dysphonia) and -5 (corresponding to severe dysphonia). Since the range of possible scores on the separate parameters is wide, scores >5 (good voice quality) or <5 (poor voice quality) are possible as well [10]. The DSI has been used to compare the voice quality of different groups of speakers [13], and to assess outcome of voice therapy and voice training programs [14-17].

Before applying a new index in clinical practice, it is necessary to know the normal values. Wuyts et al. used a control group consisting of 68 subjects (43 female, 25 male) and concluded that DSI values for males and females differ not significantly. However, the mean age and age range of their control group was not mentioned.

It is well known that voices change with age. For example, listeners are able to estimate the age of someone just by listening to a person's voice [18-20]. Voices of elderly people are often described as breathy, weak and trembling [20-25], but also as warmer and more appreciated than younger voices [26]. Not all perceptive aspects occur in all elderly people. The age at which these aspects occur differs highly among subjects [27]. The degree of changes with age depends on several aspects, such as physical and psychological health, hereditary and social factors [22,28,29]. The effects of ageing on the vocal folds have been described widely [21,25,27,30-39]. The laryngeal change mostly described in ageing males is vocal fold atrophy and 'bowing' of the vocal folds [21,25,27,37]. The elastic fibers in the intermediate layer of the lamina propria of the vocal folds become less dense and atrophic; the collagenous fibers in the deep layer become denser and more fibrotic [27]. The laryngeal skeletal ossificates [39]. In ageing females vocal fold oedema is often described [21,27,37,39]. In those studies the voice quality of the subjects is not known, consequently the relation between the findings and the effect on voice quality remains unclear. Other investigators described the effects of ageing on different vocal quality parameters [21,22,24,28,30,37,39-42]. The effects of ageing on the mean speaking fundamental frequency are often described. The male voice shows a higher [21,28,37] and the female voice a lower [21,28,37,40,41] mean speaking fundamental frequency with advancing age. Teles-Magelhaes et al. [22] described a lowering of the highest frequency in females as well as a restriction of the minimum intensity limits. The MPT may be affected by a

decline of pulmonary function with age [24]. Also jitter values may increase with advancing age [28,42] although Ferrand [41] found no changes in jitter in ageing females.
Considering these effects of age and gender on voice quality, it could be expected that the parameters of the DSI might change with advancing age, as well as differ between the sexes with advancing age. To discriminate between voice disorders and normal effects of ageing it is important to have normative values of a group of subjects of a wide age range without voice complaints.
The aim of this study was to investigate the influence, if any, of age and gender on the DSI and its separate parameters and to obtain age- and gender-related normative data.

METHODS

SUBJECTS

Adult volunteers without voice complaints participated in this study, which was performed at our Department of Otorhinolaryngology. They were recruited from faculty members, students and residents of retirement homes in the vicinity. To avoid effects of vocal mutation and maturation, minimum age was 20 years. There was no maximum age for inclusion in this study. The subjects had no history of voice disorders or voice therapy. Their voices were judged as normal by the two investigators (G0 on the GRBAS-scale [43]). It is known that smoking has an influence on voice quality and is as such a potential confounder. Therefore smokers were excluded.

EQUIPMENT

Recordings were made on a Sony Digital Audio Tape recorder (DTC-57ES) with a Sennheiser microphone (BG 2.0 dyn). The distance between mouth and microphone was 30 cm. Intensity and frequency measurements were obtained with an automatically recording phonetograph (Pabon/Laryngograph 1997). Traditionally a stopwatch was used for measuring MPT.The Multi-Speech program (Kay Elemetrics) was used for calculating jitter.

MEASUREMENTS

Speech-pathology students in their final year performed the measurements. These students were trained by the first author in performing the measurements. All subjects completed a questionnaire regarding former and present voice problems, general health and smoking habits. Data recording took place in a room with ‘living room acoustics’ and low environmental noise (<40 dB(A)), as recommended by Schutte and Seidner [44], since the use of a strongly sound damped room (like audiometer cabines) might influence the auditive self-control of the subject’s own voice. The complete procedure lasted about half an hour.

Frequency and Intensity

The subjects were asked to phonate on /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. The clinician stimulated and modelled the subject to achieve the highest possible pitch. 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 three times; the longest phonation time was used.

Jitter

The subjects phonated three times on /a/ at a comfortable pitch and loudness during approximately 3 s. The percentage jitter was calculated on a sample of 1 s, starting half a second post voice onset. To rule out technically invalid measurements due to incorrect marking of the voiced periods, the lowest result of the three calculations was used.

STATISTICAL ANALYSIS

For the statistical analysis the SPSS statistical program release 10.1 for Windows was used. Since age was not normally distributed, medians and percentiles were calculated. Means with standard deviation (SD) and range of the 4 separate parameters and the DSI were calculated for males and females. Scatterplots with regression prediction lines and 95% confidence interval lines were made for females and males of all separate parameters and the DSI against age. A regression analysis was performed to determine the association of the separate parameters and the DSI with age and gender. When an association of the voice parameters with age was different for males and females (interaction), the interaction with gender was taken into account by including an interaction term in the model. We considered a p-value of 0.05 or less as significant.

RESULTS

SUBJECTS

The number of volunteers participated was 118 (69 female, 49 male). The median age was 43 years (range 20-79 years). The median age of the female group was 39 years (range 20-79 years), the median age of the male group was 49 years (range 20-79). In table 1 the numbers of subjects in age groups of 10 years are shown.

Table 1. Number of subjects per age group of 10 years

Age, years	Female	Male	Total
20-30	28	16	44
31-40	7	3	10
41-50	9	7	16
51-60	12	13	25
61-70	7	5	12
71-80	6	5	11

GENDER

Table 2 shows that significant differences between males and females were found for F_0-high and for MPT. The mean F_0-high was higher for females (943 Hz) than for males (650 Hz). The mean MPT was lower for females (19 s) than for males (25 s). The mean DSI was not significantly different for females (4.3) and males (3.8).

Table 2. Gender effect. Mean values for highest frequency (F_0-high), lowest intensity (I-low), maximum phonation time (MPT), jitter and the Dysphonia Severity Index (DSI)

Subjects	F_0-high Hz	I-low dB SPL	MPT s	Jitter %	DSI
Female					
Mean	**943***	57	**19***	0.73	4.3
S.D.	243	3.3	6.7	0.45	2.01
Range	415 - 1397	51 - 66	8 - 39	0.30 - 2.89	-1.2 - 9.3
Male					
Mean	**650***	56	**25***	0.75	3.8
S.D.	161	2.9	9.3	0.70	1.94
Range	294 - 988	51- 62	11 - 48	0.19 - 4.31	-2.8 - 7.8
Total					
Mean	821	56	21	0.77	4.1
S.D.	298	3.2	8.4	0.57	2.00
Range	294 - 1397	51- 66	8 - 48	0.19 - 4.31	-2.8 - 9.3

* = Significant at 0.01 level

AGE

Table 3 shows the influence of age on all parameters. Only if the influence of age was different for males and females, the interaction with gender was taken into account and indcluded in table 3.
F_0-high decreased significantly with advancing age, as well in females as in males (figure 1a). I-low became significantly higher with advancing age in the female group (figure 1b). There were no significant changes with advancing age in MPT (figure 1c) and jitter (figure 1d). The DSI decreased significantly with advancing age in the female as well as the male group (figure 1e).

Table 3. Age effect: regression coefficient (B) and 95% confidence intervals for B, for all parameters on gender, age and if applicable the interaction between gender and age

	B (95% confidence interval)				
	F_0-high	I-low	MPT	Jitter	DSI
Gender (M vs. F)	**-272.2*** (-441.7 to -202.7)	2.91 (-0.17 – 5.99)	**5.6*** (2.67 – 8.58)	-0.01 (-0.20 – 0.22)	-0.37 (-1.04 – 0.30)
Age, years	**-5.8*** (-7.7 to -3.8)	**0.07*** (0.02 – 0.11)	-0.02 (-0.10 – 0.06)	0.00 (0.00 – 0.01)	**-0.05*** (-0.07 to -0.03)
Interaction gender-age	-	**-0.08*** (-0.14 to -0.02)	-	-	-

* = Significant at 0.01 level

DISCUSSION

In this study the influence of gender and age on the DSI and its separate parameters was investigated. We found that age had a significant influence on the DSI, while the influence of gender was not significant (p=0.28).

GENDER

In this study a (significantly) higher F_0-high for females was found; this is in concordance with the results of other studies [10,45,46]. The mean values in our study (943 Hz for females and 650 Hz for males) were somewhat higher than in the studies of Wuyts et al. (905 and 602 Hz) and Van de Heyning et al. (867 and 586 Hz) and somewhat lower than those reported by Hollien et al. (1108 and 698 Hz). In our study the MPT was

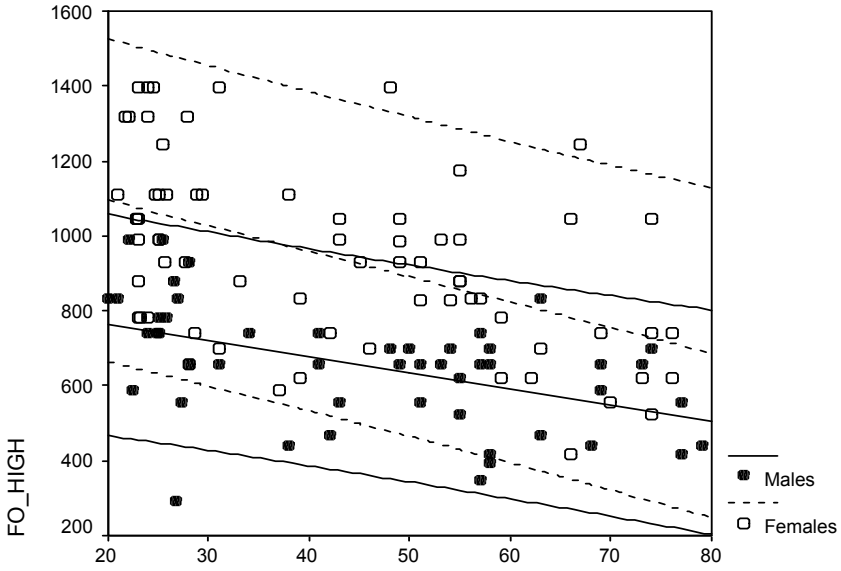
(significantly) higher for males. The mean values reported in various studies are rather diverse. For females mean values of 17-26 s [46,47] have been reported (19 s in our study). For males the mean values range from 22 to 35 s [46,47] (25 in our study). In all studies the mean values for males are higher than for females [10,43,46,47]. I-low did not differ between the sexes. This was also found in other studies [10,46]. Sulter et al. [48], however, found that males were able to phonate softer than females. We found higher values than reported by Wuyts et al. [10] (57 vs. 51 dB for females and 56 vs. 50 dB for males). Wuyts et al. [10] measured the lowest intensity in dB(A). This filter affects the frequencies below 1000 Hz. All subjects achieved their lowest intensity phonating at their lower frequencies, well below 1000 Hz. We did not use a filter and measured in dB SPL, which probably explains the difference in values. Jitter did not differ between the sexes, the values of females (0.73%) and males (0.75%) were almost equal. Wuyts et al. [10] reported a higher jitter for females (0.79 %) than for males (0.63%). Also in other studies the jitter values for males were lower than for females [46,49]. This difference might be due to the age of the subjects in this study because especially in males the jitter values become higher with advancing age.

Wuyts et al.[10] claim that DSI does not show gender differences, because the differences in F_0-high (higher in females) and MPT (higher in males) are opposite and counteracting. In our study the mean DSI for females was not significantly higher (4.3) than for males (3.8). Apparently also in our study the differences in F_0-high and MPT neutralized in the DSI. Our values are lower than in the study of Wuyts et al. [10] (5.2 for females and 4.7 from males). This seems mainly due to the fact we did not use a filter for measuring intensity.

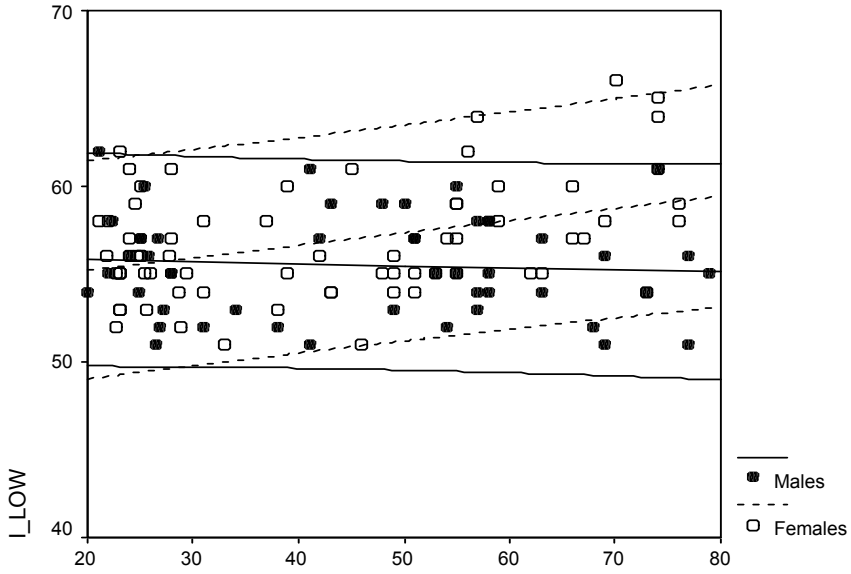
AGE

In this study, F_0-high was measured. The results showed that F_0-high decreased significantly in both females and males with advancing age. Hollien et al. [45] found no effect on the highest frequency with advancing age in theri group of subjects (male and female). This is probably due to the fact that their subjects were younger (between 18 and 36 years) than our subjects. For females our results correspond to the results of Teles-Magalhaes et al. [22]. The lowering of the highest frequency is in line with the described increase of vocal fold edema. The higher mean speaking F_0 in elderly males seems to be in contradiction with our findings of a decrease in F_0-high in males. However, this is probably the result of two different mechanisms. The lowering of F_0-high (in males and females) is possibly the result of the decreased elasticity of the vocal folds [39]. The higher mean speaking F_0 in males could be the result of the greater effort needed to close atrophic vocal folds.

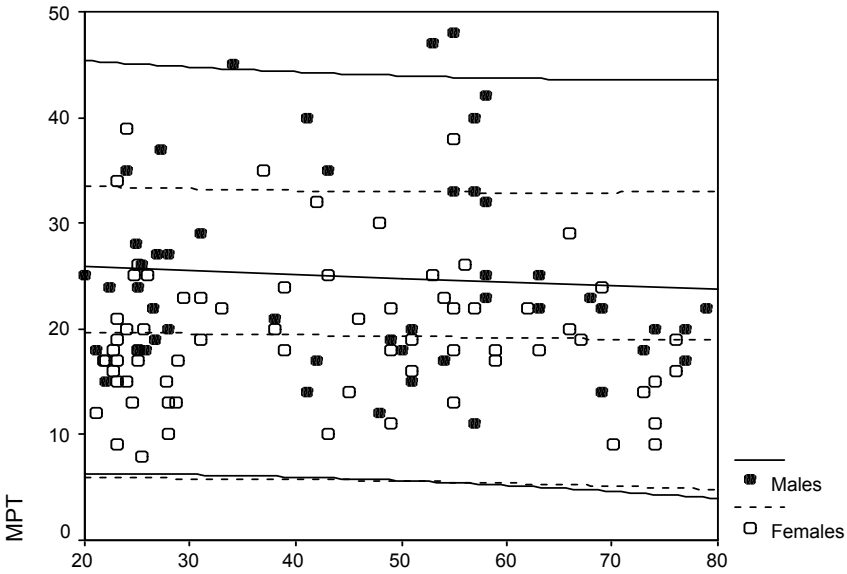
Figure 1 Scatterplots of F_0-high (1a), I-Low (1b), MPT (1c), jitter (1d) and DSI (1e) with age for females and males with regression lines and 95% confidence intervals



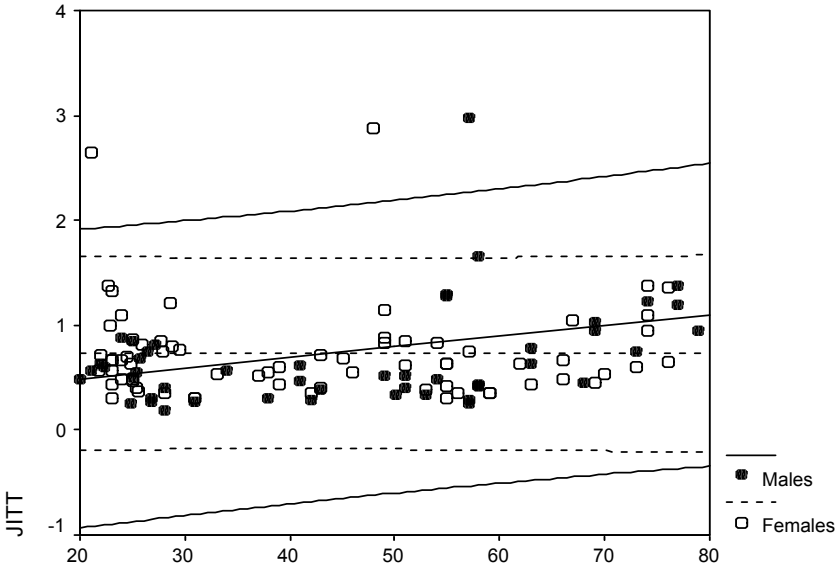
1a AGE



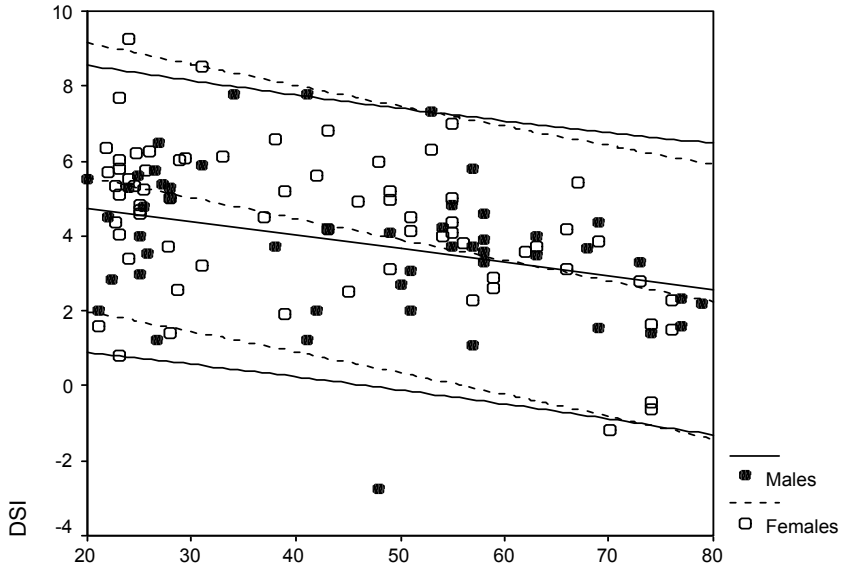
1b AGE



1c



1d



1e

The lowest intensity became significantly higher with advancing age only in females; this is in concordance with the findings of Teles-Magalhaes et al. [22]. Possibly a diminished control on breath support and phonation plays a role. It is not clear why there is a difference between males and females. Apparently the 'bowing' (and incomplete closure) of the vocal folds as described in males [21,25,27,37] did not lead to an increase in I-low.

MPT has a rather large range of 'normal values'. The mean values in healthy people described by several investigators vary from 16 s [46] to 35 s [47]. The mean value found in the present study was 21 s, with a very large range of values (from 8-48 seconds). It would not have been surprising if young persons had a longer MPT than older persons, caused by a decline of pulmonary function with age [24]. However, no such effect was found. This is possibly due to the fact that there is no relationship between MPT and vital capacity, as Solomon et al.[50] investigated in healthy young subjects.

In the present study jitter became higher with advancing age in males; this was however not significant (p=0.07). For females no changes were found. There seemed to be a tendency for jitter to become higher in all subjects over sixty (figure 1d). This tendency was confirmed by a regression analysis performed for all subjects over sixty, and was significant (p=0.00).

As could be expected from the values of the different parameters, age had an effect on the DSI, i.e., DSI decreased significantly with advancing age in both females and males. The age effect was somewhat, but not significantly, stronger in females than in males.

This effect on the DSI seemed to be mostly the result of the lowering F_0-high with advancing age and probably of a higher jitter, especially in subjects over sixty. In daily clinical practice this means that the DSI of a patient with voice complaints should be compared to the score that would be expected given the patients age and gender.

PREDICTION OF DSI

It is possible to predict the normative DSI, given the age and gender of a subject. The figures of the regression analysis (in table 3) and the corresponding constant term are used to construct a formula for this prediction. The formula for females is as follows: DSI = 6.33 - 0.05 x age (in years). For males it is: DSI = 5.96 – 0.05 x age (in years).

CONCLUSIONS

It is known that age and gender affect voice quality. In this study we found that age has a significant effect on the DSI, and on its parameters F_0-high and I-low (only in females). Gender has no effect on the DSI, although it has a significant effect on the parameters F_0-high and MPT. To be able in clinical practice to distinguish between the effects of normal ageing and a voice disorder, normative DSI-values for age and gender are now provided.

ACKNOWLEDGEMENTS:
The authors wish to thank the participating volunteers and the students who recruited and examined them. We also thank Elien Gerritsma, Ineke Hartgring and Hans Verschuure for their time and help.

REFERENCES

1. Munoz J, Mendoza E, Fresneda MD, Carballo G, Lopez P. Acoustic and perceptual indicators of normal and pathological voice. *Folia Phoniatr Logop* 2003;55:102-114.
2. Michaelis D, Frohlich M, Strube HW. Selection and combination of acoustic features for the description of pathologic voices. *J Acoust Soc Am* 1998;103:1628-1639.
3. Dejonckere PH, Lebacq J. Acoustic, perceptual, aerodynamic and anatomical correlations in voice pathology. *ORL J Otorhinolaryngol Relat Spec* 1996;58:326-332.
4. Giovanni A, Revis J, Triglia JM. Objective aerodynamic and acoustic measurement of voice improvement after phonosurgery. *Laryngoscope* 1999;109:656-660.
5. Morsomme D, Jamart J, Wery C, Giovanni A, Remacle M. Comparison between the GIRBAS Scale and the Acoustic and Aerodynamic Measures Provided by EVA for the Assessment of Dysphonia following Unilateral Vocal Fold Paralysis. *Folia Phoniatr Logop* 2001;53:317-325.
6. Heman-Ackah YD, Michael DD, Goding GS, Jr. The relationship between cepstral peak prominence and selected parameters of dysphonia. *J Voice* 2002;16:20-27.
7. Heman-Ackah YD, Heuer RJ, Michael DDet al. Cepstral peak prominence: a more reliable measure of dysphonia. *Ann Otol Rhinol Laryngol* 2003;112:324-333.
8. Piccirillo JF, Painter C, Fuller D, Fredrickson JM. Multivariate analysis of objective vocal function. *Ann Otol Rhinol Laryngol* 1998;107:107-112.
9. Yu P, Ouaknine M, Revis J, Giovanni A. Objective voice analysis for dysphonic patients: a multiparametric protocol including acoustic and aerodynamic measurements. *J Voice* 2001;15:529-542.
10. Wuyts FL, De Bodt MS, Molenberghs Get 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.
11. Piccirillo JF, Painter C, Fuller D, Haiduk A, Fredrickson JM. Assessment of two objective voice function indices. *Ann Otol Rhinol Laryngol* 1998;107:396-400.
12. Yu P, Revis J, Wuyts FL, Zanaret M, Giovanni A. Correlation of instrumental voice evaluation with perceptual voice analysis using a modified visual analog scale. *Folia Phoniatr Logop* 2002;54:271-281.
13. Timmermans B, De Bodt MS, Wuyts FLet al. Poor voice quality in future elite vocal performers and professional voice users. *J Voice* 2002;16:372-382.

14. 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.

15. 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.

16. 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.

17. 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.

18. Ramig LO, Gray S, Baker Ket al. The aging voice: a review, treatment data and familial and genetic perspectives. *Folia Phoniatr Logop* 2001;53:252-265.

19. Horii Y, Ryan WJ. Fundamental frequency characteristics and perceived age of adult male speakers. *Folia Phoniatr (Basel)* 1981;33:227-233.

20. Biondi S, Zappala M, Amato G, Consoli F. [The senile voice]. *Acta Otorhinolaryngol Ital* 1992;12:69-79.

21. Honjo I, Isshiki N. Laryngoscopic and voice characteristics of aged persons. *Arch Otolaryngol* 1980;106:149-150.

22. Teles-Magalhaes LC, Pegoraro-Krook MI, Pegoraro R. Study of the elderly females' voice by phonetography. *J Voice* 2000;14:310-321.

23. Ryan WJ, Burk KW. Perceptual and acoustic correlates of aging in the speech of males. *J Commun Disord* 1974;7:181-192.

24. Sataloff RT, Rosen DC, Hawkshaw M, Spiegel JR. The aging adult voice. *J Voice* 1997;11:156-160.

25. Ximenes Filho JA, Tsuji DH, do Nascimento PH, Sennes LU. Histologic changes in human vocal folds correlated with aging: a histomorphometric study. *Ann Otol Rhinol Laryngol* 2003;112:894-898.

26. Sulter AM, Schutte HK, Miller DG. Standardized laryngeal videostroboscopic rating: differences between untrained and trained male and female subjects, and effects of varying sound intensity, fundamental frequency, and age. *J Voice* 1996;10:175-189.

27. Hirano M, Kurita S, Sakaguchi S. Ageing of the vibratory tissue of human vocal folds. *Acta Otolaryngol* 1989;107:428-433.

28. Decoster W, Debruyne F. The ageing voice: changes in fundamental frequency, waveform stability and spectrum. *Acta Otorhinolaryngol Belg* 1997;51:105-112.

29. Ramig LA, Ringel RL. Effects of physiological aging on selected acoustic characteristics of voice. *J Speech Hear Res* 1983;26:22-30.

30. Linville SE. Source characteristics of aged voice assessed from long-term average spectra. *J Voice* 2002;16:472-479.

31. Ramig LA. Effects of physiological aging on speaking and reading rates. *J Commun Disord* 1983;16:217-226.

32. Sato K, Hirano M. Age-related changes in the human laryngeal glands. *Ann Otol Rhinol Laryngol* 1998;107:525-529.

33. Sato K, Hirano M, Nakashima T. Age-related changes of collagenous fibers in the human vocal fold mucosa. *Ann Otol Rhinol Laryngol* 2002;111:15-20.

34. Takeda N, Thomas GR, Ludlow CL. Aging effects on motor units in the human thyroarytenoid muscle. *Laryngoscope* 2000;110:1018-1025.

35. Woo P, Casper J, Colton R, Brewer D. Dysphonia in the aging: physiology versus disease. *Laryngoscope* 1992;102:139-144.

36. Kersing W, Jennekens FG. Age-related changes in human thyroarytenoid muscles: a histological and histochemical study. *Eur Arch Otorhinolaryngol* 2003.

37. Higgins MB, Saxman JH. A comparison of selected phonatory behaviors of healthy aged and young adults. *J Speech Hear Res* 1991;34:1000-1010.

38. Bloch I, Behrman A. Quantitative analysis of videostroboscopic images in presbylarynges. *Laryngoscope* 2001;111:2022-2027.

39. Paulsen F, Kimpel M, Lockemann U, Tillmann B. Effects of ageing on the insertion zones of the human vocal fold. *J Anat* 2000;196 (Pt 1):41-54.

40. Decoster W, Debruyne F. Longitudinal voice changes: facts and interpretation. *J Voice* 2000;14:184-193.

41. Ferrand CT. Harmonics-to-noise ratio: an index of vocal aging. *J Voice* 2002;16:480-487.

42. Linville SE, Fisher HB. Acoustic characteristics of women's voices with advancing age. *J Gerontol* 1985;40:324-330.

43. Hirano M. *Clinical examination of voice*. Wien; New York: Springer, 1981.

44. Schutte HK, Seidner W. Recommendation by the Union of European Phoniatricians (UEP): standardizing voice area measurement/phonetography. *Folia Phoniatr (Basel)* 1983;35:286-288.

45. Hollien H, Dew D, Philips P. Phonational frequency ranges of adults. *J Speech Hear Res* 1971;14:755-760.

46. Van de Heyning PH, Remacle M, Van Cauwenberge Pet al. Voice assessment: what and why - results of a multicenter study concerning 1100 Belgian subjects. In: Dejonckere PH, Peters HFM, eds. *24th IALP congress*. Amsterdam: Nijmegen University Press, 1998:263-266.

47. Hirano M, Koike Y, Von Leden H. Maximum phonation time and air usage during phonation. Clinical study. *Folia Phoniatr (Basel)* 1968;20:185-201.

48. Sulter AM, Schutte HK, Miller DG. Differences in phonetogram features between male and female subjects with and without vocal training. *J Voice* 1995;9:363-377.

49. Fitch JL. Consistency of fundamental frequency and perturbation in repeated phonations of sustained vowels, reading, and connected speech. *J Speech Hear Disord* 1990;55:360-363.

50. Solomon NP, Garlitz SJ, Milbrath RL. Respiratory and laryngeal contributions to maximum phonation duration. *J Voice* 2000;14:331-340.

CHAPTER 7

The applicability of the Dysphonia Severity Index and Voice Handicap Index in evaluating effects of voice therapy and phonosurgery

Hakkesteegt MM, Brocaar MP, Wieringa MH

J Voice 2009; [Epub ahead of print]

ABSTRACT

OBJECTIVES

To investigate the applicability of the Dysphonia Severity Index (DSI) and the Voice Handicap Index (VHI) in evaluating effects of intervention between groups of patients and for intrasubject differences and whether DSI and VHI are complementing measurements.

METHODS

Analyses of measurement data before and after intervention of 171 patients with voice disorders. The voice quality was measured objectively with the DSI. The perceived voice handicap was measured with the VHI. Three groups of patients were used: patients who had voice therapy, phonosurgery, or no intervention.

RESULTS

DSI and VHI improved significantly after intervention in the voice therapy and the surgery group (median difference DSI 1.19 and 3.03, VHI –8 and –26, respectively). The intrasubject results were analyzed based on the test-retest variability of DSI and VHI. Significant better DSI and VHI scores after intervention were found in, respectively, 22% and 38% of the patients with voice therapy, and 56% and 78% of the patients with surgery. In the no intervention group, this was 11% and 12%. In 37% of the patients, the differences before and after intervention in DSI and VHI were in discordance.

CONCLUSION

The DSI and VHI are able to show significant differences after intervention for voice disorders between groups of patients. The DSI and VHI can be used to determine a significant intrasubject result of intervention. The DSI and VHI measure each different aspects of the voice and are complementing measurements. The DSI is therefore applicable in clinical practice for objective evaluation of voice quality and the VHI for subjective evaluation of the perceived handicap by the patient self.

INTRODUCTION

Evaluating effects of intervention is of growing importance in today's health care, because of the need for evidence-based intervention. For voice disorders also, research on the effects of intervention is needed. There are however not yet well-accepted standardized instruments that can be used to assess the effects of intervention for voice disorders. When evaluating effects of intervention, there are two different aspects to take into account: the differences between groups of patients (intersubject differences) and the difference within one patient before and after intervention (intrasubject differences). The differences in outcome between groups are needed for research purposes: to compare a new type of intervention with a commonly used type of intervention, or to determine what the best type of intervention is for a certain diagnosis. Therefore, the intersubject variance of the used measurements has to be known. In daily clinical practice, it is important to be able to interpret differences between measurements of one patient made on different points in time (e.g. before and after intervention). To know whether differences are significant, the intrasubject variance of the used measurements has to be known.

Because voice disorders consist of different aspects (voice quality, voice handicap), several measurements should be used. Clinical assessment of voice disorders should consist of (video)laryngostroboscopy, perceptual voice assessment, objective measurements (acoustic analysis and aerodynamic measurements) and subjective self-evaluation of voice [1]. However, not all these aspects appear equally suitable for evaluating effects of intervention.

Although (video)laryngostroboscopy is a very important clinical tool for diagnosing and evaluating patients with voice disorders, it has not been widely used as a research tool because the interpretation is subjective and reliable and quantifiable tools for research purposes are not yet available [2]. For the perceptual voice assessment, the 'GRBAS scale' as introduced by Hirano [3] is widely used. The reliability of Grade has been investigated [4,5]. However, these investigations are expressed as levels of agreement (kappa values) for inter- and intra-rater and test-retest reliabilities, and are not expressed as the intrasubject variance. The reason for this is probably that Grade is scored on a categorical scale and consequently calculations cannot be made. Therefore, Grade appears not to be suitable for evaluation of intervention effects, neither between groups of patients, nor for intrasubject differences [6,7]. There is no consensus on what objective measurements for voice quality are best suitable to measure effects of intervention. In studies describing effects of intervention for voice disorders, a variety of measurements is used. The choice for the used measurements in evaluation studies can be based on expected changes in specific aspects of voice quality [8-11]. However, in daily clinical practice it is most practical to use the same objective measurement for all voice disorders. It is already known that multiparametric measures are more suitable for

evaluation of voice quality than single measures [12-15]. The Dysphonia Severity Index (DSI) [14] is such a multiparametric measure. The DSI has a good relationship with the perceptual evaluation on Grade of the GRBAS scale [16]. An advantage of the DSI is that the parameters can be obtained relatively quick and easy by speech pathologists in daily clinical practice.

For the self-evaluation of voice, the Voice Handicap Index (VHI) is a widely used tool [17]. The VHI is a subjective self-administered questionnaire addressing the patients perceived disability. The VHI and the DSI measure each different aspects of voice and the outcomes on both measurements are therefore not necessarily related. The patient's perception of the voice disorder is not only related to voice quality (as measured with the DSI) but is also related to, for example, professional and social vocal demands and personal aspects.

For both the DSI and VHI the clinical significance (standard deviation [SD]) is known, obtained from test-retest variability investigations [18,19]. This clinical significance is necessary to interpret the differences between measurements before and after intervention. Therefore, these measurements could be suitable to evaluate effects of intervention. The VHI is already used for evaluation, the DSI however is until now only used on a very limited scale.

We hypothesized that it is possible to evaluate results of intervention for voice disorders with a widely applicable objective measurement. We also hypothesized that measurements of different aspects of voice disorders will complement each other. The purposes of this study were as follows:

To investigate whether the DSI and the VHI can be used to evaluate effects of different types of intervention for voice disorders between groups of patients.

To investigate whether the DSI and VHI can be used in daily clinical practice to determine a significant intrasubject effect of intervention for voice disorders.

To investigate whether the DSI and VHI are complementing measurements, by investigating the relationship between the differences before and after intervention on the DSI and the VHI.

MATERIALS AND METHODS

SUBJECTS

The measurements of the patients with voice disorders visiting the outpatient clinic of our department of Otorhinolaryngology are collected in a database. The measurements are done at the first visit and at follow-up visits. When patients have voice therapy, the follow-up measurements are done at least 3 months after the first therapy session. When patients have phonosurgery, the follow-up measurements are done at least 2 months

after surgery. When there is no intervention, the minimum time interval had to be 6 weeks. From this database, the data of the patients of whom DSI measurements were available of their first visit (preintervention) and at least one follow-up visit (postintervention) were used. In case of several follow-up measurements, the last measurement was used as postintervention measurement. This resulted in measurement data of 171 patients (74 male, 97 female) with a mean age of 43 years (range, 15-82, SD 15 years). From 122 of those patients, also VHI measurements of both visits were available. The VHI was not for all patients available, partly because we introduced the VHI later than we started with DSI measurements, partly because some of the patients were not able to complete the questionnaire due to language problems. The median time interval between the measurement of the first and the last visit was 29 weeks (minimum 6, maximum 171 weeks). Patients were classified in three diagnosis groups: nonorganic dysphonia, mass lesions (nodules, polyps, cysts, laryngitis and edema), and paresis/paralysis (unilateral and bilateral paresis and paralysis). This classification was already used in a previous study [16]. There were two intervention groups: surgery combined with voice therapy (further called 'surgery group') and voice therapy only. Patients who visited the department only for follow-up, while there was no (further) intervention were assigned to a 'no intervention' group. These were patients for whom no intervention was available or who chose not to be treated.

PROCEDURES

The patients were asked to fill in a VHI form while they were in the waiting room. All patients were examined by one of the two speech pathologists of the department, who measured the DSI parameters. Thereafter, the clinical diagnosis was made with laryngostroboscopy by one of the two ear, nose, and throat ENT/voice-specialists of the department.

MEASUREMENTS

The parameters used for DSI measurements are the highest fundamental frequency (F_0-high in Hz), lowest intensity (I-low in dB sound pressure level (SPL)), maximum phonation time (MPT in s) and jitter (%). The DSI is constructed as DSI = 0.13 x MPT + 0.0053 x F_0-High - 0.26 x I-low - 1.18 x 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. Scores beyond this range (higher than +5 or lower than -5) are also possible. To obtain I-low, the subjects were asked to phonate an /a/ as softly as possible at a comfortable pitch. To obtain F_0-high, 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. To measure MPT, the subjects were asked to inhale deeply and sustain an

/a/ for as long as possible at a comfortable pitch and loudness. The MPT was recorded three times; the longest measured phonation time in seconds was used. To calculate jitter, the subjects phonated three times an /a/ at a comfortable pitch and loudness during approximately 3 seconds. The jitter was calculated on a sample of one second, starting half a second after the voice onset. The lowest result of the three calculations was used.

EQUIPMENT

Intensity and frequency measurements were obtained with an automatically recording phonetograph (Pabon/Laryngograph 1997). The *Multi-Speech* program (Kay Elemetrics, Lincoln Park, NJ) was used for calculating jitter. Audio recordings were made with a sampling rate of 11,025 Hz and 16 bits quantization. A Sennheiser microphone (BG 2.0 dyn) was used. The distance between mouth and microphone was 30 cm. Data recording took place in a room with 'living room acoustics' [20].

STATISTICS

For the statistical analysis, the SPSS statistical program release 10.1 for Windows was used. The distribution of the measurement data was tested for normality with the Kolmogorov-Smirnov one sample test. When the distribution of the measurements was normal, a Student's t-test was used for comparison of the effects of intervention in the patient groups. In case of nonnormality of the data, nonparametric paired tests were performed. Probability values of less than 0.05 were considered significant. For the intrasubject differences, a 'better' score on the DSI was defined as a difference between the pre- and post-intervention score of ≥2.5 (1.96 X SD) and 'worse' ≤-2.5. A 'better' score on the VHI was defined as a difference of ≤-14 (1.96 X SD) and 'worse' ≥14. These definitions of significant intrasubject differences are based on test-retest studies [18,19]. For the associated results of the DSI and the VHI, a plot was made of the difference before and after intervention in VHI against the difference before and after intervention in DSI.

RESULTS

The diagnoses and types of intervention of the 171 patients with voice disorders are shown in Table 1.

Table 1. Diagnosis and type of intervention for all patients

	Nonorganic Dysphonia	Mass Lesions	Paresis / Paralysis	Total
No Intervention	14	22	9	45
Voice Therapy	20	58	14	92
Surgery	0	20	14	34
Total	35	100	37	171

DIFFERENCES BETWEEN INTERVENTION GROUPS

The DSI and VHI results for the intervention groups are shown in Figure 1. In this figure, the medians and 25th and 75th percentiles of the measurements before and after the different types of intervention are shown for the DSI (Figure 1a) and the VHI (Figure 1b). Both the median baseline DSI and the median baseline VHI of the no intervention group were better than the median baseline DSI and VHI of the patients in the surgery and voice therapy group. The DSI improved significantly ($P<0.001$) after intervention in the voice therapy group and the surgery group (median of the difference 1.19 and 3.03, respectively). The VHI improved significantly ($P<0.001$) as well in the voice therapy group and the surgery group (median of the difference –8 and –26, respectively).

INTRA-SUBJECT DIFFERENCES

For clinical practice purposes, the intrasubject effects of intervention measured with the DSI and the VHI are expressed as percentages of patients who scored worse, the same or better as defined in the statistics section. These results are shown in Table 2 for each type of intervention.

In Figure 2a, the relationship between the DSI before and after each type of intervention is shown. Points left of the diagonal represent improvement. Points left of the dotted reference lines of 2.5 (1.96 X SD) and right of the dotted reference line of –2.5 indicate a significant intrasubject difference. This figure shows that patients who had a low DSI before intervention showed the largest improvements. Patients who had a good DSI before intervention, close to the mean DSI of people without voice disorders, showed only small differences after intervention. Figure 2 also shows that the largest part of the patients had after intervention a better DSI than before intervention.

In Figure 2b, the relationship between the VHI before and after each type of intervention is shown. Points right of the diagonal represent improvement. Points left of the dotted reference lines of 14 (1.96 X SD) and right of the dotted reference line of –14 indicate a significant intrasubject difference. Most patients (74%) had a better VHI after intervention.

Table 2. Percentages of patients who scored significantly worse, the same and significantly better on DSI and VHI for each type of intervention

Intervention	Difference DSI			Difference VHI		
	Worse (< -2.5) (%)	Same (-2.5–2.5) (%)	Better (>2.5) (%)	Worse (>14) (%)	Same (-14–14) (%)	Better (<-14) (%)
No Intervention	11	78	11	16	72	12
Voice Therapy	1	77	22	8	53	38
Surgery	6	38	56	0	22	78

For the DSI and the VHI the definitions for worse, the same and better are based on the significant intra-subject difference (1.96 X SD).

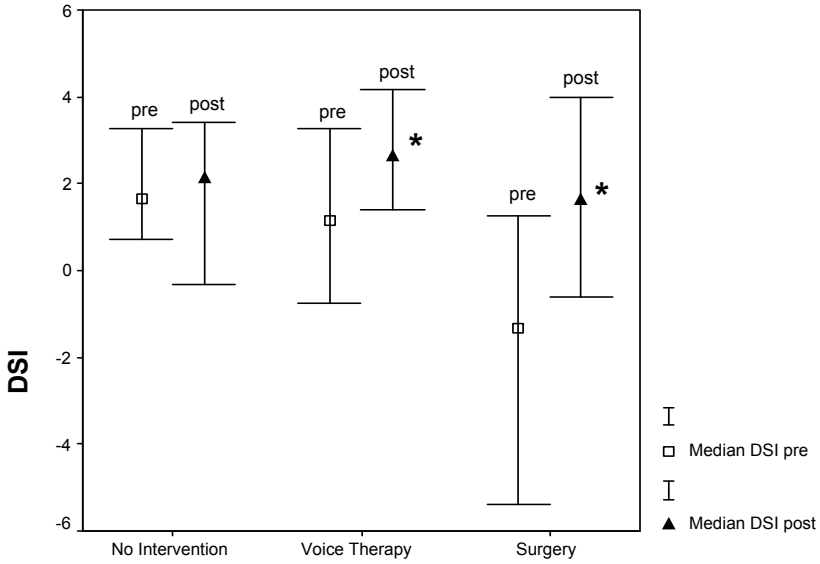
RELATIONSHIP BETWEEN THE DSI AND THE VHI

Figure 3 shows the relationship between the difference before and after intervention in DSI and VHI. This relationship is shown for the three types of intervention. The points within the grey area represent the nonsignificant differences (±1.96 X standard deviation) for both the DSI and the VHI. In 53% of the patients both DSI and VHI were improved, in 10% both were deteriorated. In 16% the DSI was improved, but the VHI was deteriorated. In 21% of the patients the DSI was deteriorated while the VHI was improved.

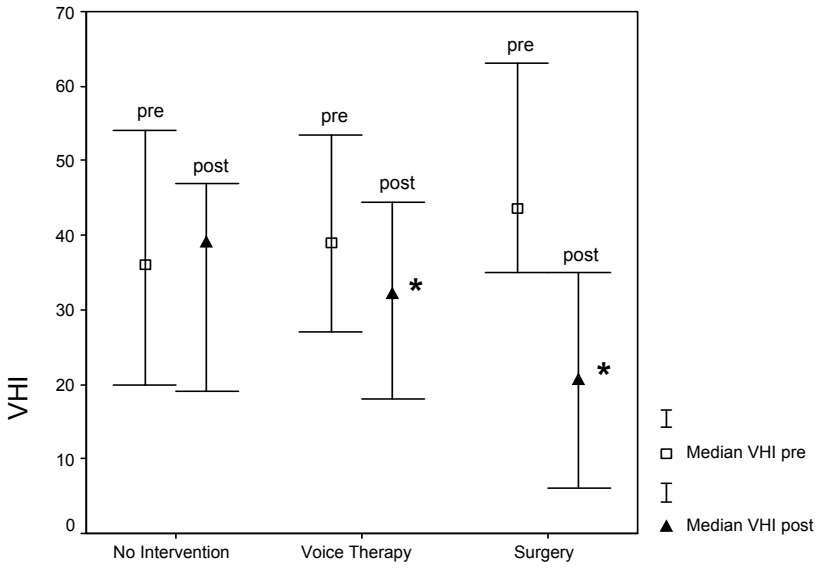
DISCUSSION

In this clinical follow-up study, data of 171 patients were analyzed. The measurements took place at two different visits to our department, with a minimum time-interval of 6 weeks. Between the two measurements, patients underwent surgery or had voice therapy. Part of the patients had no intervention. In most studies evaluating effects of intervention, only the results of a particular therapy for groups of patients are investigated, whereas the results for individual patients (intrasubject results) are not investigated [8,21-23], even though these intra-subject results are relevant in daily clinical practice for the evaluation of effects of intervention (e.g., before and after voice therapy). A possible explanation for this lack of data is that perceptual measurements scored on categorical scales are often used and consequently calculations cannot be made. Therefore, these measurements are not suitable for evaluation of intrasubject results. In this study, the applicability of the DSI and VHI for both evaluating the results for groups of patients and intrasubject differences were investigated.

Figure 1 Results for the groups of patients before and after intervention with the median scores and the 25th and 75th percentiles for DSI (a) and VHI (b). A higher DSI corresponds with a better voice quality, a lower VHI corresponds with less complaints



1a * Difference between pre and post is significant (p<0.05)



1b * Difference between pre and post is significant (p<0.05)

Figure 2 The relationship between the measurements before and after intervention for DSI (a) and VHI (b). The dashed lines represent the significant differences (± 1.96 X standard deviation = +2.5 and –2.5 for DSI and +14 and –14 for VHI). In Figure a, points left of the diagonal represent improvement, points right of the diagonal represent deterioration. In Figure b, points left of the diagonal represent deterioration, points right of the diagonal represent improvement.

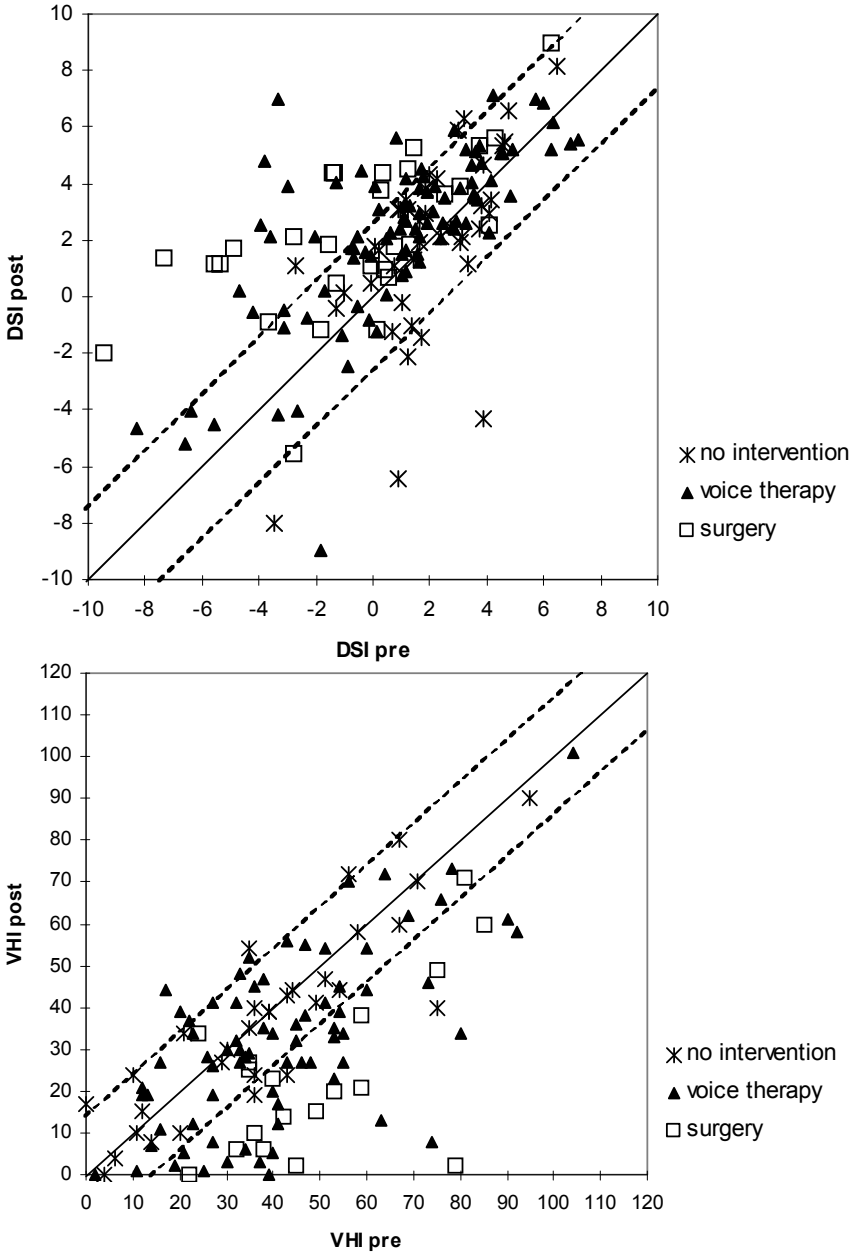
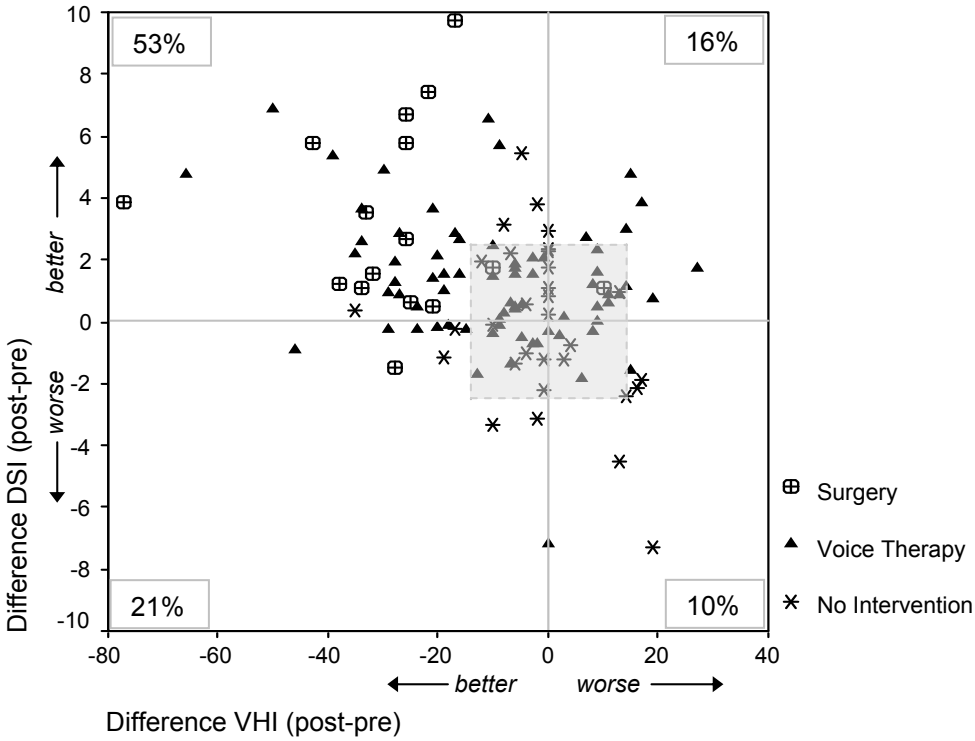


Figure 3 The relationship between the difference before and after intervention in DSI and VHI, for the different types of intervention. The points within the grey area represent the nonsignificant differences (± 1.96 X standard deviation) for both the DSI and the VHI. The percentages of the total group are shown for each quadrant. A higher DSI corresponds with a better voice ('better'), a lower DSI corresponds with a worse voice ('worse'). A lower VHI corresponds with less complaints ('better'), a higher VHI corresponds with more complaints ('worse').



DIFFERENCES BETWEEN INTERVENTION GROUPS

Investigation of the effect of different types of intervention is needed for research purposes: to compare a new type of intervention with a commonly used type of intervention, or to determine what the best type of intervention is for a certain diagnosis. To investigate whether the DSI and the VHI can be used for this purpose, three different groups of patients were used. When comparing these groups of patients, the results of our study show that surgery and voice therapy lead to a significant improvement in DSI (voice quality) and in VHI (self-evaluation). The largest differences were seen in the surgery group. The patients who were measured during follow-up without intervention

showed no significant differences in DSI and VHI within time. This demonstrates that the DSI and VHI are able to compare different types of intervention.
The preintervention VHI scores of the surgery group were worse than the preintervention scores of the voice therapy group, although, the scores were not as much worse as expected based on the differences between both groups in the DSI scores (Figure 1). This could indicate that the patients in the surgery group were used to their voice disorder before intervention, and that not all of them were really aware of the limits their voice caused in every day life. Therefore, they tended to value their perceived handicap (VHI) before surgery better than expected based on the objectively measured voice quality (DSI). Only after improvement of the voice quality after surgery, they became conscious of what they missed and evaluated their voice with the VHI as much better than before the surgery. It is remarkable that the postintervention VHI scores of the surgery group were better than the postintervention scores of the voice therapy group (Figure 1b), whereas the postintervention DSI scores of the surgery group were worse than of the voice therapy group (Figure 1a). Apparently, also after surgery, the voice was evaluated better (VHI) by the surgery group than expected based on the measured voice quality (DSI). Besides the before-mentioned underestimation of the preintervention voice handicap, patients who had surgery possibly assume their voice must have improved, because of the invasive character of the intervention.

INTRA-SUBJECT DIFFERENCES

The intrasubject differences are relevant in daily clinical practice for the evaluation of effects of intervention for an individual patient. Therefore, the intrasubject effects of intervention measured with the DSI and VHI were evaluated as well. In the voice therapy group, 22% of the patients showed a significantly better DSI after intervention, whereas the VHI was significantly improved in 38% of the voice therapy group. The larger percentage of improvement on the VHI than on the DSI could be due to the fact that, although the voice quality (DSI) was not improved, the patients learned in voice therapy to cope better with their (disordered) voice. From the patients in the voice therapy group with no significant changes in DSI (77%) and VHI (53%), the majority shows an improvement, as shown in Figures 2a,b. Only 1% had a significant worse DSI, and 8% had a worse VHI. For the patients who had surgery, the group with a significantly improved VHI was also larger (78%) than the group with a significantly better DSI (56%). This could be due to a combination of better coping with their voice, and the assumption that the voice must have improved after the invasive intervention (surgery). From the patients in the surgery group with no significant changes in DSI (38%), the majority shows an improvement, as shown in Figure 2a. In the surgery group 6% had a significant worse DSI, and none of the patients had a worse VHI. Most patients in the no intervention group showed no significant changes in DSI (78%) and the VHI (72%) at the follow-up measurement. From the patients with significant changes in the no intervention

group 11% had a worse DSI and 11% had a better DSI. For the VHI, this was 16% worse and 12% better. For both the DSI and the VHI, it is obvious in all patient groups that a preintervention score close to scores of subjects without voice disorders limits the possible improvement (Figures 2a,b).

Relationship between the DSI and the VHI

The VHI and the DSI both measure different aspects of voice disorders and are not necessarily related. The patient's perception of the disorder is not only related to voice quality (as measured, for example, with the DSI), but is probably even more related to professional and social vocal demands and personality aspects. Professional voice users will suffer more from a relatively little deviant voice quality than someone who, on average, is not very talkative. Furthermore, some patients with voice disorders do not visit the clinic because of the complaints they have, but to exclude a possibly malignant voice disorder. We investigated the relationship between the differences before and after intervention on both measurements. We expected that the measurements for DSI and VHI would often both show an improvement or deterioration. However, because the DSI and the VHI measure different aspects of voice disorders, it is conceivable that the differences in DSI and VHI before and after intervention will be in discordance. For example, perhaps patients become used to their voice quality, or have learned in voice therapy to cope better with their vocal limitations. This could result in a better VHI and an unchanged or worse DSI. The opposite is also possible: patients who had successfully voice therapy can become even more conscious of the limitations their voice causes them, because they are more aware of their voice due to the therapy. This could result in a better DSI and a worse VHI. In the study population, for most of the patients (63%) the results for DSI and VHI were in concordance. In 37% of the patients, the results were in discordance: in 16% the DSI was better whereas the VHI was worse, in 21% the DSI was worse, whereas the VHI was better.

The two groups with discordant results were further analyzed, including the non-significant differences.

The group of patients with a better DSI and a worse VHI consists of 17% of the voice therapy group (16 patients), 3% of the surgery group (one patient) and 2% of the total no intervention group (one patient). Most of the patients who had voice therapy had mass lesions (nodules, laryngitis). Perhaps, they became more conscious of their voice due to the voice therapy (worse VHI), but it is also possible that they learned how to use their voice better (better DSI), but that they were not (yet) able to apply this in daily voice use (worse VHI).

The group with a worse DSI and a better VHI consists of 16% of the voice therapy group (15 patients), 3% of the surgery group (one patient) and 20% of the no intervention group (nine patients). In this group also, most of the patients who had voice therapy had mass lesions (most nodules). Apparently, these patients indeed learned to cope with the

limitations of their voice quality, due to the voice therapy. The nine patients who had no intervention had various diagnoses. Possibly, these patients were getting used to the limitations of their voice quality. Furthermore, it was remarkable that four of those nine patients had a relatively good first DSI (>3), their second DSI was not significantly worse.
In 33% of all patients who had voice therapy, the results on DSI and VHI were in discordance: 17% with better DSI and worse VHI, and 16% with worse DSI and better VHI. These patients were comparable regarding the diagnosis and as far as could be retrieved, regarding their vocal demands as well. Apparently, the VHI and DSI describe each different aspects of voice. Behrman et al.[24] also concluded that patient's perception of severity of voice disorders was independent of factors like auditory perceptive evaluation, phonatory glottal closure and lesion type. Evaluation of patient's perception of the voice handicap appears an important element in the assessment of voice disorders and in therapeutic decision making.
Results on both the DSI and the VHI should be taken into account when decisions about intervention are made. If a patient with a benign vocal fold mass lesion has a low score on the VHI (little complaints), there is possibly no indication for surgery, regardless of the DSI. On the other hand, for a comparable patient with a high score on the VHI, but a good DSI, surgery may also not be the first choice. Because the voice quality cannot improve much, and therefore the complaints (VHI) might not improve much as well. For these patients, it could be a better choice to start with voice therapy.
In the evaluation of voice disorders, the DSI and the VHI are complementing measurements. It is useful to evaluate the voice quality (DSI) and the perceived handicap (VHI) combined.

CONCLUSION

In the evaluation of effects of intervention for voice disorders, two aspects are important: the intersubject and intrasubject differences. The intersubject differences are needed in research for evaluation of different types of intervention. The DSI and VHI are both able to show differences between groups of patients, and can therefore be used to evaluate effects of different types of intervention. The intrasubject differences are needed in clinical practice to determine whether measurements made on different points in time of one patient (e.g., before and after intervention) are significantly different. The DSI and VHI can both be used for this purpose. The DSI and the VHI measure each different aspects of voice and can be seen as complementing measurements.

REFERENCES

1. Dejonckere PH, Bradley P, Clemente Pet 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. Rosen CA. Stroboscopy as a research instrument: development of a perceptual evaluation tool. *Laryngoscope* 2005;115:423-428.
3. Hirano M. *Clinical examination of voice*. Wien; New York: Springer, 1981.
4. De Bodt MS, Wuyts FL, Van de Heyning PH, Croux C. Test-retest study of the GRBAS scale: influence of experience and professional background on perceptual rating of voice quality. *J Voice* 1997;11:74-80.
5. Webb AL, Carding PN, Deary IJ, MacKenzie K, Steen N, Wilson JA. The reliability of three perceptual evaluation scales for dysphonia. *Eur Arch Otorhinolaryngol* 2004;261:429-434.
6. Shrivastav R, Sapienza CM, Nandur V. Application of psychometric theory to the measurement of voice quality using rating scales. *J Speech Lang Hear Res* 2005;48:323-335.
7. 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.
8. Cantarella G, Mazzola RF, Domenichini E, Arnone F, Maraschi B. Vocal fold augmentation by autologous fat injection with lipostructure procedure. *Otolaryngol Head Neck Surg* 2005;132:239-243.
9. Krengli M, Policarpo M, Manfredda Iet al. Voice quality after treatment for T1a glottic carcinoma--radiotherapy versus laser cordectomy. *Acta Oncol* 2004;43:284-289.
10. Roy N, Weinrich B, Gray SD, Tanner K, Stemple JC, Sapienza CM. Three treatments for teachers with voice disorders: a randomized clinical trial. *J Speech Lang Hear Res* 2003;46:670-688.
11. Selby JC, Gilbert HR, Lerman JW. Perceptual and acoustic evaluation of individuals with laryngopharyngeal reflux pre- and post-treatment. *J Voice* 2003;17:557-570.
12. Piccirillo JF, Painter C, Fuller D, Haiduk A, Fredrickson JM. Assessment of two objective voice function indices. *Ann Otol Rhinol Laryngol* 1998;107:396-400.
13. Piccirillo JF, Painter C, Fuller D, Fredrickson JM. Multivariate analysis of objective vocal function. *Ann Otol Rhinol Laryngol* 1998;107:107-112.

14. Wuyts FL, De Bodt MS, Molenberghs Get 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.

15. Yu P, Ouaknine M, Revis J, Giovanni A. Objective voice analysis for dysphonic patients: a multiparametric protocol including acoustic and aerodynamic measurements. *J Voice* 2001;15:529-542.

16. Hakkesteegt MM, Brocaar MP, Wieringa MH, Feenstra L. The Relationship Between Perceptual Evaluation and Objective Multiparametric Evaluation of Dysphonia Severity. *J Voice* 2007.

17. Jacobson BH, Johnson A, Grywalski Cet al. The Voice Handicap Index (VHI): Development and Validation. *American Journal of Speech-Language Pathology* 1997;6:66-70.

18. Hakkesteegt MM, Wieringa MH, Brocaar MP, Mulder PG, Feenstra L. The Interobserver and Test-Retest Variability of the Dysphonia Severity Index. *Folia Phoniatr Logop* 2008;60:86-90.

19. Hakkesteegt MM, Wieringa MH, Gerritsma EJ, Feenstra L. Reproducibility of the Dutch version of the Voice Handicap Index. *Folia Phoniatr Logop* 2006;58:132-138.

20. Schutte HK, Seidner W. Recommendation by the Union of European Phoniatricians (UEP): standardizing voice area measurement/phonetography. *Folia Phoniatr (Basel)* 1983;35:286-288.

21. Schindler A, Palonta F, Preti G, Ottaviani F, Schindler O, Cavalot AL. Voice quality after carbon dioxide laser and conventional surgery for T1A glottic carcinoma. *J Voice* 2004;18:545-550.

22. Uloza V, Pribuisiene R, Saferis V. Multidimensional assessment of functional outcomes of medialization thyroplasty. *Eur Arch Otorhinolaryngol* 2005;262:616-621.

23. Schneider B, Denk DM, Bigenzahn W. Acoustic assessment of the voice quality before and after medialization thyroplasty using the titanium vocal fold medialization implant (TVFMI). *Otolaryngol Head Neck Surg* 2003;128:815-822.

24. Behrman A, Sulica L, He T. Factors predicting patient perception of dysphonia caused by benign vocal fold lesions. *Laryngoscope* 2004;114:1693-1700.

CHAPTER 8

General discussion

A ‘voice disorder’ is a quite broad concept: it contains everything that patients might experience as changes in their voice quality or limitations in their vocal functioning, i.e. limited frequency range, limited intensity, limited duration of voice use or physical complaints like a sore throat after speaking. For most people such a disorder causes problems in their social and/or professional functioning. Different aspects need to be assessed in a description of any voice disorder: the complaints of the patient, the sound of the voice and the aetiology. The aetiology is determined by history taking and physical examination. The laryngeal anatomy and the function of the vocal folds are examined through laryngo-stroboscopy. All aspects together usually will lead to a diagnosis.
After the diagnosis a therapeutic scheme is drafted. This may consist of a surgical procedure, some medication, voice therapy or any combination of these.
To establish the effect of any intervention some kind of an evaluation is mandatory. So far studies on evaluation are scarce. A possible reason for this lack of research data might be the missing ‘gold standard’ for voice measurements. This hampers seriously the comparison between published studies. Another reason might be that up till now the concept of ‘evidence based practice’ as in medicine is rather new or even non-existent in voice therapy, possibly because most voice therapists have had only little or no scientific training. Another problem consists in the heterogeneity in aetiology and therapy of patient groups. Moreover it takes a long time to collect homogeneous groups of significant size. All this makes comparison between different types of therapy hard or even impossible.
Consensus exists that voice disorders have many facets and therefore measurements also need to be multidimensional. This thesis emphasizes the evaluation of both the voice quality and the complaints of the patient. Evaluation consists in both the voice quality and the complaints of the patient. Evaluation of voice quality usually is done by listening to the voice and describing its qualities (subjective perceptual evaluation) and/or by instrumental measuring different aspects of the sound of the voice (objective evaluation). For the complaints of the patient different validated quality of life type questionnaires can be used.
These three methods of evaluation will be discussed.

PERCEPTUAL EVALUATION

Since the most ‘ear-catching’ of a voice disorder is its sound it is merely logical to use hearing to evaluate voice quality. Normal hearing is perfectly capable to detect even minimal changes in a voice like emotions, especially in the voices of people well known to the listener. We are even capable up to a certain point to estimate age and gender of a voice just by sheer listening. We are also able to hear abnormality within a voice. This of course is a relative judgment as we compare any voice with our internal standard of normality. This standard probably varies a little on a daily basis of any listener and also

between different listeners. It goes without saying that most listeners will agree on the differences between a normal voice and a severely disordered one, but between small variations this tends to become more problematic.

One may compare it with judging someone's length. We have an inner feeling about the length of humans with which we more or less judge someone as tall or short taking into account – with children that is – their age and gender. To use such a subjective judgment is fine on a daily basis but not acceptable for scientific aims. For the latter, large groups of children of different ages are assembled and a mean length with a certain bandwidth is accepted for any specific age, taking into account systematic and chance variations.

For voice quality, unfortunately, so far no such objective measurements exist. Traditionally, the perceptual evaluation is the only measurement used and therefore often considered to be the 'gold standard'. A description of a voice in 'free prose' obviously is not really useful. Therefore definitions of terminology and scoring systems of severity have been introduced, to make comparison between investigators and between different institutes possible. The system mostly used is probably the GRBAS-system of Hirano. In this system Grade (G), Roughness (R), Breathiness (B), Asthenia (A) and Strain (S) are scored on a four-point scale. The reliability of this system was investigated by several investigators (see chapter 2, table 3). Especially the intra- and inter-observer reliability is investigated and appeared to be moderate. Some investigators have studied scoring systems that differed from the original version, like ten-point scales or Visual Analogue Scales, but the results are ambiguous.

However, these 'scores' are still subjective judgments put into categories, which are not mathematical at all. Categories cannot easily be used to calculate for statistical analysis. Mean values and a standard deviation cannot be computed. Therefore computing a bandwidth including 95% of the subjects is not even possible. This distribution is needed to determine a significant deviation, needed to evaluate effects of intervention. Besides, the perceptual evaluation seems to have only limited applicability in evaluating intervention. As in daily practice, it frequently happens that both the patient and the therapist are of the opinion that the voice quality has changed, while this is not reflected in changes in the perceptual evaluation. We conclude that perceptual evaluation has only limited value and is definitely not a robust 'gold standard'.

There are no articles on intra-subject variability in perceptual judgments of voice qualities probably because the intra- and inter-observer reliability is very modest. Moreover it should be taken into account that any voice fluctuates over time. These fluctuations may be small but it seems relevant to know how large these fluctuations are in order to interpret correctly the results of measurements.

A possibility to evade some difficulties with the perceptual evaluation is to avoid the problem of the 'personal-individual' standard through using the patient's voice as its own standard, i.e. compare two samples of the same patient by employing sound files. In this

way factors like articulation, intonation or an accent don't disturb judgment. The two samples of one subject could be judged as 'the same', 'better' or 'worse' for several perceptual parameters of voice quality, for example G, R, B, A and S. Overall 'quantification' of any voice change will be reflected in the compound number of changed parameters: the more parameters are changed in the same direction (i.e. better or worse) the larger the difference is. A suggestion for future research is to investigate the reliability of these comparative judgments.
Technical progress made comparison easier. Voices used to be recorded on tapes. A lot of work was needed to make a tape with sets of samples of any one patient. In the clinical setting moreover comparison of the voice was difficult as older records were most of the time on different tapes. Nowadays due to the digital era these problems are solved. Sound files can be stored and retrieved easily. Also medical records of patients are increasingly digital which makes it possible to attach sound files to the patient's record. Those can be compared immediately, even during a patient's visit, making comparative perceptual judgments workable in daily practice.
It is anyhow important that, when using subjective perceptual judgment of voice quality, one is aware of its modest reliability.

OBJECTIVE VOICE QUALITY MEASUREMENTS

The human voice is a complex sound signal and as a consequence measuring its quality is difficult. In the first place we should be clear which qualities of the voice we are interested in. Thereafter, a cut-off point should be chosen to determine what should be considered normal and abnormal.
The 'ideal voice quality measurement' has yet to be discovered. Many objective parameters are somehow related with the perceptual judgment, because it is considered the best available 'gold standard'. Measurements combining a variety of parameters are considered to relate better with our perceptive evaluations than any single one. The Dysphonia Severity Index is such a multi-parameter measurement. The parameters used in the DSI are the highest fundamental frequency, the lowest intensity, the maximum phonation time and jitter. An advantage of the DSI is that the parameters can be obtained relatively quickly and easily by speech pathologists in daily practice. Since it is a relatively new measurement, only limited data about the reliability and applicability were available. Therefore we studied the test-retest reliability, normative values of the DSI, the relationship between the perceptual evaluation and the DSI and the clinical applicability of the DSI in evaluating intervention for voice disorders.
The test-retest variability of the DSI appeared to be rather large. We found that a minimal significant difference within one patient on two occasions on the DSI has to be 2.5 points. A possible weakness of our test-retest study was that we tested subjects without voice complaints or voice disorders. It is possible that in patients the test-retest variability

differs from the control group. Besides, we tested these healthy persons three times within three weeks and the variability appeared to be partly due to a 'learning effect'. In patients, the time interval between measurements before and after intervention will usually be at least three months and then a 'learning effect' will probably not occur. In our analysis for computing significant differences learning effects were taken into account.

It is well known that voices change with age and it is conceivable that the parameters of the DSI may change with advancing age. Therefore we investigated the influence of aging on the DSI and we established age related normal values. These normal values are important for use in clinical practice as they make it possible to distinguish between natural aging effects on the voice and pathology of the voice. We found no need for gender-related normal values, probably since the substantial difference between males and females on the highest frequency is compensated by the results on the maximum phonation time.

To evaluate the use of the DSI for follow up of therapies three groups of patients were studied. One group without intervention; one group received voice therapy and one group had a surgical intervention. As stated above a 'gold standard' would have been ideal, but is lacking as yet. Therefore we had to rely on our clinical experience. Although our three groups of patients were heterogeneous in respect to their diagnosis, we expected clear differences in the results of their respective interventions. The results of the DSI indeed did show the expected results: no or only small differences were encountered within the group who had no intervention, modest differences were seen in the voice therapy group and clear differences were met within the surgery group. These results suggest that the DSI is useful for evaluating results of intervention. To study this further, it would be worthwhile to compare the results of the DSI with comparative perceptual judgment as described before.

PATIENT BASED QUESTIONNAIRES

The complaints of a patient with a voice disorder are often not exclusively related to the sound of their voice. Often the limitations in daily functioning are also troublesome for the patient. To record the complaints of the patient with a validated questionnaire is valuable, both at the first contact and as a system to compare variations over time.

In our study we used the Dutch version of the Voice Handicap Index (VHI). This questionnaire is widely used and cited in literature. We studied the test-retest variability to establish the intra-subject variability. Our results show that a difference of fourteen points between two different measurements as for instance before and after therapy is significant. This cut-off of fourteen points is easy to use in daily practice and very informative for both the therapist and the patient.

Since a questionnaire is relatively reliable and easy to evaluate the voice, it is an essential part of voice assessment. The VHI is applicable in research and in daily practice. Most patients can complete the form easily.

APPLICABILITY

To determine whether the DSI can be used to discriminate between patients and controls, the sensitivity and specificity for different cut-off points were calculated. With a DSI cut-off of 3.0, the sensitivity is 0.72 and the specificity is 0.75. Nevertheless, some patients visiting our clinic with voice problems have a DSI above 3.0. These are mostly patients without voice quality problems, but who are experiencing limitations in the length of time they can use their voice without strain or sometimes having a sore throat after speaking.
For the applicability in evaluating results of intervention, the significant difference of a measure has to be known. For the DSI the cut-off point for a significant difference is 2.5, for the VHI this is fourteen points. We consider these cut-off points reasonable and applicable in daily practice. As the differences in DSI between different observers were not significant it is concluded that the findings of different observers and different institutes are comparable.
The applicability of the DSI and VHI in measuring results of intervention was also tested. As could be expected, the results of the VHI and the DSI turned out not always to be in concordance. For example, in some patients the voice quality improves (better DSI), but possibly they become more aware of the limitations of the voice, resulting in a worse VHI. Apparently the complaints of the patient and the voice quality are two different facets of a voice disorder and therefore need to be evaluated separately. The DSI and VHI are measuring different aspects of a voice disorder and therefore need to be used together.

CONCLUSIONS AND RECOMMENDATIONS

Perceptual evaluation up till now is generally considered to be a gold standard for voice evaluation, but more than one such perceptual system is used. The reliability of these systems is moderate at best. It is important to be aware of its limitations. We recommend using sound files in addition to perceptual judgment. These sound files can be attached to digital patient's records, making a comparative perceptual judgment, for example before and after treatment, possible.
We conclude that the DSI is a valuable objective and quantitative addition to the subjective perceptual evaluation of voice quality, applicable in evaluation of results of

intervention and making comparison of results of intervention between different institutes possible because the measurements can be standardised.
To investigate the complaints of a patient the VHI is valuable and clinical applicable.
Evaluation of voice problems has to be multidimensional. The combination of both DSI and VHI is recommended.

CHAPTER 9

Summary / Samenvatting

SUMMARY

The human voice is important in daily communication in spite of the increasing number of alternative electronic tools of communication like text messaging and e-mail. Approximately one third of the working population nowadays needs their voice to earn their money and it goes without saying that most people use their voice for daily social activities. Only when the voice causes trouble, the importance of a good voice dawns upon people. Diagnosis and treatment of voice disorders is by that of great importance.
This thesis focuses on the evaluation of intervention for treatment of benign voice disorders by speech therapists through available measurement tools.
Not only the voice is complex, its disorders are as well. For the evaluation of voice disorders there is as yet no consensus about the exact outcome parameters to be used, apart of the consensus that voice measurements should be multidimensional.
To evaluate the complaints of the patient standardised self-evaluation questionnaires (to be filled in by the patient personally) may be used. Traditionally a perceptual (subjective) evaluation is used, which is often viewed upon as the 'gold standard'. Beside this subjective evaluation, also more objective measurements can be used. An important advantage of objective measurements is that - provided that the measurements are standardised - the results of different speech centres and measurements over time are interchangeable and comparable.
In our study we made use of the Dutch version of the Voice Handicap Index (VHI). This VHI is frequently used in research, can be completed easily by most patients, and is known the whole world through.
For the perceptual evaluation we used an overall rating of voice quality by means of the score for 'Grade' on a four-point scale.
For the objective evaluation of the voice quality we used the Dysphonia Severity Index (DSI). This is a multiparametric index. The used parameters are the highest fundamental frequency, the lowest intensity, the maximum phonation time and the percentage of the jitter. Speech therapists are able to measure these parameters easily and quickly in daily practice.

Chapter 2 presents an overview of the literature of the intra-subject reproducibility of voice measurements. The intra-subject reproducibility is the degree of consistency of the voice measurements of one patient (subject) at different times. This is relevant to discriminate between normal fluctuations and real changes in voice quality, for example after therapy. Review of the literature shows that the reliability of the subjective perceptual evaluation has been studied mostly. Such investigations are mainly focused on the intra- and inter-observer variability. The intra-observer variability is the degree of consistency of the different judgments of one patient by one observer at different times. The inter-observer variability is the degree of consistency of the different judgments of one patient by different observers. No research so far has been described on the intra-

subject variability of the perceptual evaluation. In general the intra- and inter-observer variability of perceptual evaluation is moderate at best. A limited number of studies describes the test-retest variability of objective voice measurements. In most studies the correlation between the measurements at different times are investigated and general conclusions about the reliability are drawn. Only rarely a differentiation between intra-subject and intra-observer variability has been made. The inter-observer variability has been addressed rarely.
In most studies about questionnaires the correlation between the results of one subject at two times (without intervention) are investigated. Only in a few studies the intra-subject variability has been addressed.
We conclude that there exists only a limited amount of available literature on reproducibility of voice measurements and even less on intra-subject variability.

Chapter 3 describes the inter-observer variability and the intra-subject reproducibility of the DSI. The DSI was measured at three different times by two different observers in thirty subjects without any voice complaints. The differences between the different observers were not significant. This is relevant in daily practice since two measurements in one patient at two different times (for example before and after therapy), will often be performed by different observers. The standard deviation of the difference between two measurements in one subject is 1.27. For daily practice this implicates that the difference between two measurements in one subject (for example before and after therapy) has to be larger than 2.49 to be accepted as significant.

Chapter 4 describes the reproducibility of the Dutch version of the VHI. In total 104 patients with voice complaints completed the questionnaire at two different times. The time interval was approximately two weeks. The standard deviation of the difference between two measurements of one patient was seven points. For daily practice this implicates that the difference between two measurements in one patient (for example before and after therapy) has to be larger than fourteen points to be accepted as significant.

In chapter 5 we investigate the relationship between the perceptual evaluation of voice quality and the DSI in 294 patients and 118 controls. Patients with a high score on Grade (bad voice quality) have a low DSI (also indicating bad voice quality). Moreover there are significant differences between the scores on DSI between patients with different causes of voice disorders. Furthermore we investigated the cross point where the DSI discriminates between patients and controls. With a DSI cutoff of 3.0, maximum sensitivity (0.72) and specificity (0.75) were found.

Chapter 6 investigates the possible effects of gender and age on the DSI. The DSI was measured in 118 controls without voice complaints with an age range of 20 – 79 years. Despite significant differences between males and females on the highest fundamental

frequency and the maximum phonation time, the differences on the DSI were not significantly different between males and females. Age has a significant effect on the DSI in both males and females: the DSI decreases with advancing age.

In chapter 7 we compare the DSI and the VHI of 171 patients with voice disorders measured on two times. Some patients were measured twice without intervention in between the two measurements, some patients received voice therapy and some patients had phonosurgery for voice improvement. The DSI and VHI improved significantly in the groups who had intervention (voice therapy or surgery). At the individual level (intra-subject) the DSI improved significantly after intervention in 22% of the patients with voice therapy, the VHI improved significantly in 38% of the patients in this group. For the patients with surgery, this was 56% and 78% respectively. In 37% of the patients, the differences in DSI and VHI before and after intervention were in discordance. The reason for this discordance is probably due to the different aspect of voice disorders that are measured with DSI and VHI. We conclude that both measurements are applicable for evaluating the results of interventions for voice disorders. Both measurements have to be used, since they are complementary

We conclude that the perceptual evaluation is of limited use in both daily practice and research. In daily practice we recommend to use sound files that can be attached to digital patient's records. This allows for direct comparative perceptual judgment of two samples of one patient (for example before and after treatment).
The DSI is an objective and quantitative evaluation of voice quality, applicable in evaluation of interventions. As the measurements can be standardised, comparison of the results of intervention between different institutes is therefore possible.
The VHI is a clinical applicable standardised questionnaire for evaluation of the complaints of the patient.
Evaluation of voice problems has to be multidimensional. We recommend the combination of both DSI and VHI.

SAMENVATTING

De menselijke stem is belangrijk in de dagelijkse communicatie, ondanks het toenemend gebruik van schriftelijke communicatiemiddelen, SMS en e-mail. Voor ongeveer eenderde van de beroepsbevolking is de stem onmisbaar voor hun werkzaamheden. Bovendien gebruiken de meeste mensen hun stem voor sociale activiteiten. Pas als er sprake is van een stemstoornis, merken veel mensen het belang van een goede stem. Diagnose en behandeling van stemstoornissen zijn daarom van groot belang.
Het onderzoek beschreven in dit proefschrift richt zich op de evaluatie door logopedisten van de behandeling van goedaardige stemstoornissen, met gebruikmaking van reeds bestaande meetinstrumenten.
Niet alleen de stem is complex, stemstoornissen zijn dat ook. Voor evaluatie van stemstoornissen bestaat geen consensus over de uitkomstmaten die daarvoor gebruikt moeten worden; alleen is men het over eens dat meer aspecten naast elkaar beoordeeld moeten worden.
Voor evaluatie van de klachten van de patiënt kan gebruik worden gemaakt van gestandaardiseerde vragenlijsten. Voor evaluatie van de stemkwaliteit wordt van oudsher gebruik gemaakt van de - subjectieve - perceptuele beoordeling die vaak wordt gezien als de 'gouden standaard' voor stemevaluatie. Daarnaast kan voor het onderzoeken van de stemkwaliteit gebruik worden gemaakt van meer objectieve maten. Deze hebben het voordeel dat ze - mits gestandaardiseerd toegepast - de mogelijkheid bieden meetresultaten uit verschillende instituten en op verschillende tijden bepaald, uitwisselbaar en vergelijkbaar maken.
Voor het onderzoek beschreven in dit proefschrift is gebruik gemaakt van de Nederlandstalige versie van de Voice Handicap Index (VHI). Dit is een wereldwijd gebruikte vragenlijst, die door de meeste patiënten eenvoudig kan worden ingevuld.
Voor de perceptuele beoordeling wordt gebruik gemaakt van de beoordeling van de totale stemkwaliteit ('Grade'), beoordeeld op een vier-puntschaal.
Voor objectieve stemmeting wordt gebruik gemaakt van de Dysphonia Severity Index (DSI). Dit is een multiparametrische index. De gebruikte parameters zijn de hoogst haalbare frequentie, de minimale intensiteit, de maximale fonatieduur, en het percentage van de jitter. Deze parameters zijn in de dagelijkse praktijk door logopedisten relatief snel en eenvoudig te meten.

Hoofdstuk 2 is een overzicht van de literatuur over de intra-subject reproduceerbaarheid van stemmetingen. De intra-subject reproduceerbaarheid is de mate van consistentie van de stemmetingen van één bepaalde patiënt op verschillende tijdstippen. Dit is van belang om onderscheid te kunnen maken tussen normale fluctuaties en daadwerkelijke veranderingen in de stemkwaliteit, bijvoorbeeld na therapie. Uit dit literatuuroverzicht blijkt dat er vooral onderzoek is gedaan naar de betrouwbaarheid van de subjectieve perceptuele beoordeling. Dat onderzoek richt zich vooral op de intra- en inter-

beoordelaarsvariabiliteit. De intra-beoordelaarsvariabiliteit is de mate van consistentie tussen verschillende beoordelingen van één bepaalde patiënt door één en dezelfde beoordelaar op verschillende tijdstippen. De inter-beoordelaarsvariabiliteit is de mate van consistentie tussen verschillende beoordelingen van één bepaalde patiënt door verschillende beoordelaars. Er is geen onderzoek beschreven naar de intra-subject variabiliteit van de perceptuele beoordeling. In het algemeen kan worden gesteld dat intra- en inter-beoordelaarsbetrouwbaarheid van de perceptuele beoordeling hooguit matig is. Er is slechts een beperkt aantal studies beschreven over de test-hertest betrouwbaarheid van objectieve stemmetingen. Bij deze onderzoeken is veelal de correlatie tussen de metingen op twee tijdstippen onderzocht en worden algemene conclusies getrokken over de betrouwbaarheid van de onderzochte maten. Zelden wordt onderscheid gemaakt tussen de intra-subject en intra-beoordeleaarsvariabiliteit. Aan inter-beoordelaarsvariabiliteit wordt eveneens zelden aandacht besteed.
Bij vragenlijsten wordt vaak onderzocht hoe de correlatie is als dezelfde lijst door een persoon op twee momenten wordt ingevuld (zonder tussentijdse interventie). De intra-subject variabiliteit (standaard deviatie) daarvan wordt zelden bepaald.
Er is dus slechts zeer beperkt onderzoek gedaan naar de reproduceerbaarheid van stemmetingen en nog minder naar de intra-subject variabiliteit.

Hoofdstuk 3 beschrijft het onderzoek naar de inter-beoordelaarsvariabiliteit en test-hertest betrouwbaarheid van de DSI. Hiervoor werd bij dertig proefpersonen zonder stemklachten op drie momenten de DSI gemeten, door twee verschillende onderzoekers. Uit het onderzoek blijkt dat het verschil tussen twee onderzoekers niet significant is. Dit is relevant voor de dagelijkse praktijk, omdat twee metingen bij een patiënt (bijvoorbeeld voor en na behandeling) niet altijd door dezelfde onderzoeker zullen worden gedaan. De standaard deviatie van het verschil tussen twee metingen bij een proefpersoon (intra-subject variabiliteit) is 1,27. Dit betekent voor de praktijk dat het verschil tussen twee metingen bij een patiënt (voor en na behandeling) significant is als het groter is dan 2,49.

Hoofdstuk 4 beschrijft het onderzoek naar de reproduceerbaarheid van de Nederlandstalige VHI. In totaal vulden 104 patiënten met stemklachten de vragenlijst op twee momenten in, met een tijdsinterval van ongeveer twee weken. De standaarddeviatie van het verschil tussen de twee meetmomenten was zeven punten. Het verschil was onafhankelijk van de totaalscore. In de praktijk betekent dit dat het verschil tussen twee metingen bij een patiënt (voor en na behandeling) significant is als het minimaal veertien punten is.

In hoofdstuk 5 onderzoeken we de relatie tussen de subjectieve perceptuele beoordeling en de DSI bij 294 patiënten en een controlegroep van 118 personen. Bij een hoge score op Grade (slechte stemkwaliteit) is de score op de DSI inderdaad lager (dus ook slechter). Bovendien blijkt dat er verschillen in uitkomsten op de DSI zijn tussen

patiënten met verschillende oorzaken van stemstoornissen. Ook is onderzocht of de DSI onderscheid maakt tussen patiënten en gezonde proefpersonen. Bij een afkappunt van 3,0 is de sensitiviteit (correct positief) 0,72 en de specificiteit (correct negatief) 0,75.

Hoofdstuk 6 beschrijft het onderzoek naar de mogelijke effecten van geslacht en leeftijd op de DSI. Hiervoor werd de DSI gemeten bij 118 personen zonder stemklachten in de leeftijd van 20 – 79 jaar. Ondanks dat de verschillen tussen mannen en vrouwen voor de hoogst haalbare frequentie en de maximale fonatieduur significant zijn, zijn de verschillen voor de DSI dat niet. Leeftijd heeft wel een significant effect op de DSI, bij zowel mannen als vrouwen: met het hoger worden van de leeftijd, wordt de DSI lager.

In hoofdstuk 7 vergeleken we de DSI en de VHI van 171 patiënten met stemstoornissen gemeten op twee momenten. Een deel van de patiënten werd gevolgd zonder interventie, een deel kreeg alleen stemtherapie en een deel onderging een stemverbeterende operatie. De DSI en de VHI verbeterden significant in de groepen die behandeling ondergingen (stemtherapie of een operatie). Op individueel niveau (intra-subject) had in de stemtherapiegroep 22% een significant verbeterde DSI en 38% een significant verbeterde VHI. In de operatiegroep was dit respectievelijk 56% en 78%. Bij 37% van alle patiënten waren de resultaten op de DSI en VHI niet met elkaar in overeenstemming. Dit is te verklaren doordat de DSI en de VHI beiden een apart aspect van de stemstoornis meten. We concluderen dat beide maten geschikt zijn om interventie bij stemstoornissen te meten, en dat beide maten naast elkaar gebruikt dienen te worden, aangezien de maten complementair zijn.

We trokken de conclusie dat de perceptuele beoordeling beperkingen heeft bij toepassing in zowel de dagelijkse praktijk als voor wetenschappelijk onderzoek. Voor de praktijk verdient het aanbeveling gebruik te maken van geluidsbestanden die aan een elektronisch patiëntendossier gekoppeld kunnen worden. Een directe vergelijking van twee fragmenten van een patiënt wordt hiermee mogelijk.
De DSI is een objectieve en kwantitatieve evaluatie van stemkwaliteit, geschikt voor het evalueren van de resultaten van interventie (zoals stemtherapie). Omdat de metingen gestandaardiseerd kunnen worden, maakt dit vergelijking van de resultaten tussen verschillende instellingen mogelijk.
De VHI is een geschikte gestandaardiseerde vragenlijst voor het evalueren van de klachten van de patiënt.
Evaluatie van stemstoornissen dient multidimensioneel te zijn, wij bevelen het gebruik van de DSI en de VHI naast elkaar aan.

DANKWOORD

De exercitie die uiteindelijk leidde tot dit proefschrift is ooit begonnen als een onschuldig ‘plannetje’. Dit plannetje ontstond tijdens het napraten over een bezocht symposium, waarbij we zeer geïnspireerd waren door de onderzoeken van anderen. Van het een kwam het ander:
Wat begon als een plannetje werd groter, en werd uiteindelijk een Plan.
Het bijbehorende onderzoekje werd serieuzer en werd een Onderzoek.
Mijn doel was ‘eerst eens een artikel’ en dat werd een heus Proefschrift.

Dit alles ging niet vanzelf, en natuurlijk was mij dit nooit in mijn eentje gelukt, ik ben velen dank verschuldigd, waarvan ik enkelen met nadruk wil bedanken:

Michael Brocaar: zonder jou was het plannetje nooit ontstaan, maar ook tijdens de uitvoer van het Plan was je onmisbaar, mede door jouw audiologische blik op de stem.
Marjan van den Brink - Wieringa: naar mate het werk vorderde, werden onze afspraken voor mij steeds belangrijker, altijd een goede, maar gezellige, stok achter de deur. Zonder jou was ik waarschijnlijk nu nog steeds bezig.
Professor Feenstra: van u kreeg ik de kans naast mijn werk als klinisch logopedist onderzoek te doen, zonder u had het plannetje nooit geleid tot een heus proefschrift.
Elien Gerritsma: voor alle steun in met name de beginfase van het onderzoek.
De studenten logopedie die de gezonde proefpersonen wisten te ‘strikken’ en te meten.
Alle proefpersonen die geheel belangeloos hebben meegewerkt.
Hans Verschuure: voor alle opbouwende kritische opmerkingen, op zowel taalkundig als statistisch vlak.
Mijn directe collega’s, in het bijzonder Ineke Hartgring en Jet de Gier: voor de steun, de ruimte en de hulp die jullie hebben geboden.
Hanneke Kalf en Hans Bogaardt, mijn paranimfen, maar vooral ook ‘partners in crime’: onze DDA, tijdens het betere wetenschappelijk veldwerk opgericht, leidde tot een mooie mix van pret en wetenschap, met nu dus de eerste promotie. Als het goed is volgen er nog twee!
Tot slot mijn familie: misschien hebben jullie stiekum wel eens gedacht dat het nooit af zou komen. Maar zie hier, het is gelukt!

CURRICULUM VITAE

Maria Margreet (Marieke) Hakkesteegt werd op 19 november 1965 te Gouda geboren. In 1984 behaalde zij het eindexamen Gymnasium β aan het Christelijk Lyceum-Havo te Gouda. In hetzelfde jaar begon zij met de Hogere BeroepsOpleiding Logopedie aan de Hogeschool Midden Nederland te Utrecht. In 1988 behaalde zij aldaar het diploma Logopedie.

Na haar afstuderen had zij in de periode van 1988 – 1990 diverse part-time aanstellingen, meestal meer tegelijk. Zij werkte in die periode in twee vrijgevestigde praktijken (in Harderwijk en Utrecht), op een MLK-school in Harderwijk, in een verpleeghuis in Nijkerk en begeleidde stagiairs logopedie bij diverse onderwijsinstellingen voor anderstaligen.

Vanaf 1990 werkt zij als klinisch logopedist op de afdeling KNO-Foniatrie van het Academisch Ziekenhuis Rotterdam Dijkzigt, inmiddels het ErasmusMC.

Daarnaast bekleedde zij van 1993 tot 2002 bij de Nederlandse Vereniging voor Logopedie en Foniatrie (NVLF) verschillende (bestuurs)functies (eerst lid en later voorzitter van het seciebestuur NVLF/G, interim-hoofdredacteur voor het Tijdschrift Logopedie en Foniatrie, lid van het hoofdbestuur). Zij is sinds 2004 lid van de Werkgroep Toegepaste Wetenschap in de Logopedie. Deze werkgroep organiseert studiedagen voor logopedisten. Van 2005 tot 2007 was zij lid van de commissie Richtlijnen van de NVLF. Tevens is zij medeoprichter en bestuurslid van de landelijke Paramedische Werkgroep Hoofd-Hals Tumoren, de PWHHT (2002 – heden).

PUBLICATIONS

Gerritsma, EJ; Brocaar, MP; Hakkesteegt, MM; Birkenhager, JC
Virilization of the voice in post-menopausal women due to the anabolic steroid nandrolone decanoate (Decadurabolin). The effects of medication for one year
Clin Otolaryngol 1994;19:79-84.

Hakkesteegt, MM; Wieringa, MH; Gerritsma, EJ; Feenstra, L
Reproducibility of the Dutch version of the Voice Handicap Index
Folia Phoniatr Logop 2006;58:132-138.

Hakkesteegt, MM; Brocaar, MP; Wieringa, MH; Feenstra, L
Influence of age and gender on the dysphonia severity index. A study of normative values
Folia Phoniatr Logop 2006;58:264-273.

Sewnaik, A; Hakkesteegt, MM; Meeuwis, CA; de Gier, HH; Kerrebijn, JD
Supracricoid partial laryngectomy with cricohyoidoepiglottopexy for recurrent laryngeal cancer
Ann Otol Rhinol Laryngol 2006;115:419-424.

Bogaardt HC, Hakkesteegt MM, Grolman W, Lindeboom R
Validation of the voice handicap index using Rasch analysis.
J Voice 2007;21:337-344.

Hakkesteegt, MM; Wieringa, MH; Brocaar, MP; Mulder, PG; Feenstra, L
The interobserver and test-retest variability of the dysphonia severity index
Folia Phoniatr Logop 2008;60:86-90.

Hakkesteegt, MM; Brocaar, MP; Wieringa, MH; Feenstra, L
The relationship between perceptual evaluation and objective multiparametric evaluation of dysphonia severity
J Voice 2008;22:138-145.

Teguh, DN; Levendag, PC; Sewnaik, A; Hakkesteegt, MM; Noever, I; Voet, P; van der Est, H; Sipkema, D; van Rooij, P; Baatenburg de Jong, RJ; Schmitz, PI
Results of fiberoptic endoscopic evaluation of swallowing vs. radiation dose in the swallowing muscles after radiotherapy of cancer in the oropharynx
Radiother Oncol 2008;89:57-63.

Hakkesteegt, MM; Brocaar, MP; Wieringa, MH
The Applicability of the Dysphonia Severity Index and the Voice Handicap Index in Evaluating Effects of Voice Therapy and Phonosurgery
J Voice 2009; [Epub ahead of print]

LIST OF ABBREVIATIONS

ADSD	Adductor Spasmodic Dysphonia
CAPE-V	Consensus Auditory Perceptual Evaluation-Voice
dB	Decibel
dBA	Decibel A-weighting
dBSPL	Decibel Sound Pressure Level
DME	Direct Magnitude Estimation
DPF	Directional Perturbation Factor
DSI	Dysphonia Severity Index
EGG	Electroglottograph
ENT	Ear Nose Throat
F_0	Fundamental frequency
GCQ	Glottal Closed Quotient
GFI	Glottal Function Index
GIRBAS	Grade Instability Roughness Breathiness Asthenia Strain
GRBAS	Grade Roughness Breathiness Asthenia Strain
HNR	Harmonics-to-Noise Ratio
Hz	Hertz
ICC	Intraclass Correlation Coefficient
INFVo	Impression Intelligibility Noise Fluency Voicing
JF	Jitter Factor
LTAS	Long-Term Average Spectrum
MPT	Maximum Phonation Time
NHR	Noise-to-Harmonics Ratio
PPQ	Pitch Perturbation Quotient
P-VHI	Pediatric Voice Handicap Index
P-VOS	Pediatric Voice Outcome Survey
P-VRQOL	Pediatric Voice-Related Quality-of-Life survey
RAP	Relative Average Perturbation
S-VHI	Singing Voice Handicap Index
USDRS	Unified Spasmodic Dysphonia Rating Scale
VAPP	Voice Activity and Participation Profile
VAS	Visual Analogue Scale
VHI	Voice Handicap Index
VHI-F	Voice Handicap Index Functional subscale
VHI-E	Voice Handicap Index Emotional subscale
VHI-P	Voice Handicap Index Physical subscale
Voice AP	Voice Activity Limitation
Voice PR	Voice Participation Restriction
VoiSS	Voice Symptom Scale
VOS	Voice Outcome Survey
VPA	Vocal Profile Analysis scheme
VPQ	Vocal Performance Questionnaire
VPQ	Voice Prosthesis Questionnaire
V-RQOL	Voice Related Quality-of-Life survey
VSR	Visual Sort and Rate