The Journal of Laryngology & Otology (2007), ***121***, *763–767.*

doi:10.1017/S0022215107007177
Printed in the United Kingdom
First published online 1 May 2007

Main Article

Optimising outcome assessment of voice interventions, I: reliability and validity of three self-reported scales

A L Webb, P N Carding, I J Deary*, K MacKenzie†, I N Steen‡, J A Wilson**

Abstract

Background: There is an increasing choice of voice outcome research tools, but good comparative data are lacking.

Objective: To evaluate the reliability and validity of three voice-specific, self-reported scales.

Design: Longitudinal, cohort comparison study.

Setting: Two UK voice clinics: the Freeman Hospital, Newcastle upon Tyne, and the Glasgow Royal Infirmary.

Participants: One hundred and eighty-one patients presenting with dysphonia.

Main outcome measures: All patients completed the vocal performance questionnaire, the voice handicap index and the voice symptom scale. For comparison, each patient's voice was recorded and assessed perceptually using the grade–roughness–breathiness–aesthenia–strain scale. The reliability and validity of the three self-reported vocal performance measures were assessed in all subjects, while 50 completed the questionnaires again to assess repeatability.

Results: The results of the 170 participants with completed data sets showed that all three questionnaires had high levels of internal consistency (Cronbach's alpha = 0.81–0.95) and repeatability (voice handicap index = 0.83; vocal performance questionnaire = 0.75; voice symptom scale = 0.63). Concurrent and criterion validity were also good, although, of the grade–roughness–breathiness–aesthenia–strain subscales, roughness was the least well correlated with the self-reported measures.

Conclusion: The vocal performance questionnaire, the voice handicap index and the voice symptom scale are all reliable and valid instruments for measuring the patient-perceived impact of a voice disorder.

Key words: Voice; Voice Quality; Quality of Life; Outcome Assessment (Healthcare); Outcome Measures

Introduction

As in many other areas of evidence-based practice, there is an increasing demand in otorhinolaryngology for reliable outcome measures. Voice clinicians need valid, repeatable and change-sensitive outcome measures in order to evaluate speech and language therapy and/or phonosurgical interventions. To date, three main areas of laryngological outcome assessment have been developed: self-reporting, acoustic analysis and perceptual rating. Detailed investigation has shown minimal correlation between general acoustic measures and patient complaints.[1] Perceptual measurement has become the accepted 'gold standard' for voice assessment; however, the process is time intensive and requires the expertise of a trained observer, usually a speech and language therapist. Also, unlike self-reported measures, expert voice rating does not reflect quality of life impact. For these reasons, there has been increasing activity in the design of research tools which evaluate the quality of life impact of dysphonia from the patient's perspective.[2–6]

From the various available self-rating scales for the evaluation of voice-related quality of life, we selected three which we considered the most appropriate for potential use in the UK. These were the vocal performance questionnaire,[2] the voice handicap index[3] and the newly developed voice symptom scale.[7] The vocal performance questionnaire was the first scale developed for use within a British population. The sample size used in the evaluation of its reliability was small,[1] and there is little information in support of its validity.[8] The voice handicap index already has a body of evidence in support of its reliability and validity, but these studies were undertaken on small samples of North American

From the Department of Speech & Language Therapy, the **Institute of Health and Society, and the ‡Department of Otolaryngology Head and Neck Surgery, Newcastle University, Newcastle, the *Department of Psychology, University of Edinburgh, Scotland, and the †Department of Otorhinolaryngology–Head and Neck Surgery, Royal Infirmary, Glasgow, Scotland, UK.
Accepted for publication: 6 January 2007.

patients.[3,9,10] The voice symptom scale, which was designed to be applicable across the range of heterogeneous voice symptoms, was developed from British patient samples.[11] It underwent rigorous psychometric evaluation of its content validity, internal consistency and factorial structure, using a number of large samples of voice patients.[7–9]

The aim of this study was to evaluate the reliability (i.e. internal consistency and repeatability) and validity of these three self-rating scales.

Methods

Patient self-reported scales

Vocal performance questionnaire. This scale was designed for use in an evaluation study of voice therapy in cases of non-organic dysphonia.[12] It consists of 12 items which address the physical aspects of the voice problem and also its social and emotional impact. It is scored to give a total score only, with no subscales. The reliability of the questionnaire was originally assessed on a group of only 10 respondents. Validity was ascertained by discussing the questionnaire with the patients in the pilot study and by correlating the scores with a mean severity rating of voice quality, determined by external raters.[8]

Voice handicap index. This disability and handicap inventory was developed for use in a variety of voice disorders.[3] Its 30 items are grouped into three content domains representing functional, emotional and physical aspects of voice disorders. The items were selected from patients' case records. The reliability of the questionnaire was assessed on a sample of 65 consecutive patients. Construct validity was evaluated by correlating the voice handicap index with the domains of the SF36, quality of life measure, in 260 patients.[9] The sensitivity to change in voice was evaluated on a sample of 37 subjects with various vocal fold abnormalities.[10] This study concluded that the voice handicap index was a useful patient-based instrument for the measurement of change following intervention. The voice handicap index has been used in previous studies to assess patients' perception of the severity of their voice disorder due to a variety of aetiologies[13,14] and in efficacy studies of intervention for voice disorders.[6,15–17]

Voice symptom scale. This 30-item scale has three content domains – impairment, physical symptoms and emotional response – and a total score. The impairment domain has 15 items and reflects the impact of the voice problem and the patient's ability to use their voice. The physical symptoms domain has seven items and addresses the symptoms which regularly occur as concomitants of voice disorder (e.g. sore throat and throatclearing). These may result from and/or exacerbate dysphonia but are not synonymous with poor voice quality. The emotional domain reflects the impact of the voice disorder on the patient's psychological well-being.

In summary, these three different questionnaires attempt to reflect the breadth of patients' voice problems, but have different derivations and thus potentially different applicability to the general population of voice-disordered patients.

Perceptual analysis

The grade–roughness–breathiness–aesthenia–strain rating scale[18] scores each of these five parameters. Each parameter is scored using a four-point rating scale, from zero (normal) to three (extreme). There is a body of evidence in support of the reliability of this scale.[19–22]

Patients

One hundred and eighty-one patients complaining of hoarseness and attending otorhinolaryngology – head and neck surgery out-patient clinics in Newcastle and Glasgow gave consent to take part in the study at their initial out-patient consultation. Patient exclusion criteria were: laryngeal cancer; age less than 18 years; pregnancy; learning difficulties; stroke; aphasia; and English not being their first language. The 127 female and 54 male patients included had a mean age of 52 years (range 18 to 88 years). Forty-four (34 per cent) were smokers. Patients' voice disorder categories are shown in Table I. At the initial out-patient appointment, each participant completed the three voice questionnaires. A sub-group of 50 participants was asked to complete a second set of the same questionnaires, one week later. The gold standard with which each questionnaire was compared was perceptual analysis of the voice using the grade–roughness–breathiness–aesthenia–strain scale.[18] This analysis was determined for each participant following a standard protocol for recording and assessment. Each patient gave a speech sample, consisting of rote counting and the days of the week, a prolonged /a/ and /i/ vowel, and three sentences from the Rainbow Passage.[23] An independent, expert rater evaluated each of the voice recordings, blinded to all but the age and sex of the participant. Each of the ratings was recorded in a standardised, pre-designed proforma.

Statistical analysis

Reliability. The assessment of reliability was based on whether each scale gave consistent and reproducible results.[22]

Firstly, the vocal performance questionnaire, the voice handicap index and the voice symptom scale were evaluated for internal consistency. From the several assessment methods available, we selected

TABLE I

DIAGNOSTIC CATEGORIES OF PATIENTS' VOICE DISORDERS

Category[10]	Participants (*n*)
Non-organic	74
Organic	57
Movement disorder	25
Systemic disorder	24
No diagnosis	1
Total	181

https://doi.org/10.1017/S0022215107007177 Published online by Cambridge University Press

the most widely used – the Cronbach's alpha reliability coefficient.[24]

Secondly, the repeatability or stability of the measurements[25] was assessed, based on analysis of correlations between repeated measures. The measures were repeated over time (i.e. test–retest reliability) in 50 of the participants with dysphonia. Test–retest reliability was assessed by calculating the intra-class correlation coefficient based on a two-way analysis of variance (subjects by occasions), with both subjects and occasions being treated as random effects.[26]

Validity. Two aspects were assessed: concurrent validity and criterion validity.

Concurrent validity is the extent to which results obtained with one measure of a construct relate to results obtained with another measure of the same construct.[24] Concurrent validity was evaluated by Pearson correlations of the three different self-reported scales.

Criterion validity is a special case of construct validity in which a stronger hypothesis is made possible by reference to some outside validating criterion or gold standard.[25,27,28] There are no gold standards available for voice-specific, self-reported patient scales, therefore criterion validity was evaluated by comparing the ratings given to the participants' voice quality using the grade–roughness–breathiness–aesthenia–strain scale[18] with the scores on the three self-reported voice scales, using the Spearman rho correlation coefficient.

Results

One hundred and seventy participants with complete data sets for all three questionnaires, and 46 with a second complete set, were included in the analysis.

Internal consistency

Generally, Cronbach's alpha coefficients of at least 0.7–0.8 are regarded as necessary for adequate internal consistency.[24] Cronbach's alpha coefficient for the vocal performance questionnaire total score was 0.81. The alpha coefficients for the domains of the voice handicap index were: physical aspects 0.85, functional aspects 0.90 and emotional aspects 0.90, with a total score of 0.95. The alpha coefficients for the domains of the voice symptom scale were impairment 0.85, physical symptoms 0.73 and emotion 0.90, with a total score of 0.89.

Repeatability

Table II shows the test–retest coefficients and the 95 per cent confidence intervals for each of the domains within each scale and for their total scores. The voice handicap index demonstrated very good stability, with a total scale test–retest reliability coefficient of 0.83. The vocal performance questionnaire and the voice symptom scale both demonstrated adequate stability, with test–retest reliability coefficients of 0.75 and 0.63, respectively. It should also be noted that the test–retest reliability of the voice handicap index and the voice symptom scale domain scores were very good.

Concurrent validity

Table III presents a correlation matrix for the domains and total score of the voice symptom scale, the domains and total score of the voice handicap index, and the total score of the vocal performance questionnaire. Most components showed strong positive correlations, except the voice symptom scale physical symptoms domain, which included relevant but non-voice throat symptoms.

Criterion validity

Table IV presents a correlation matrix for the self-reported scales and the parameters of the grade–roughness–breathiness–aesthenia–strain auditory rating scale. Observer- and self-rated voice quality are two different things, and it is not surprising that the overall strength of correlations between the self-reported voice scales and the grade–roughness–breathiness–aesthenia–strain scale is less than that of correlations between the three self-reported scales. Results for the vocal performance questionnaire and the voice handicap index significantly correlated with all parameters of the grade–roughness–breathiness–aesthenia–strain scale except roughness. The highest vocal performance questionnaire correlation was with overall grade (0.32),

TABLE II

REPEATABILITY MEASURES FOR THE 3 SELF-REPORTED VOICE SCALES*

Scale	Domain	Test–retest reliability coefficient†	95% CI
VPQ	Total	0.75	0.60, 0.86
VHI	Physical aspects	0.73	0.57, 0.85
	Function	0.79	0.66, 0.88
	Emotional	0.90	0.83, 0.94
	Total	0.83	0.73, 0.91
VoiSS	Impairment	0.66	0.49, 0.81
	Physical symptoms	0.78	0.66, 0.88
	Emotion	0.79	0.65, 0.88
	Total	0.63	0.43, 0.79

*The vocal performance questionnaire (VPQ), the vocal handicap index (VHI) and the voice symptom scale (VoiSS). †Intra-class correlation coefficient. CI = confidence intervals

https://doi.org/10.1017/S0022215107007177 Published online by Cambridge University Press

TABLE III

PEARSON PRODUCT MOMENT CORRELATION MATRIX FOR THE 3 SELF-REPORTED VOICE SCALES*

	VoiSS impairment	VoiSS physical symptoms	VoiSS emotion	VoiSS total score	VPQ total score
VHI physical aspects	0.78††	0.25††	0.55††	0.78††	0.73††
VHI function	0.75††	0.14	0.73††	0.81††	0.73††
VHI emotion	0.59††	0.07	0.83††	0.71††	0.67††
VHI total score	0.77††	0.17†	0.80††	0.87††	0.76††
VPQ total score	0.79††	0.13	0.62††	0.78††	

*The voice symptom scale (VoiSS), the vocal performance questionnaire (VPQ) and the vocal handicap index (VHI). ††Correlation significant at the 0.01 level (two-tailed); †correlation significant at the 0.05 level (two-tailed).

TABLE IV

SPEARMAN CORRELATION COEFFICIENTS FOR THE 3 SELF-REPORTED VOICE SCALES* AND THE GRBAS SCALE

	Grade	Roughness	Breathiness	Asthenia	Strain
VPQ total score	0.32††	0.05	0.31††	0.28††	0.21††
VHI physical aspects	0.24††	−0.01	0.28††	0.20††	0.20†
VHI function	0.40††	−0.09	0.44††	0.41††	0.30††
VHI emotion	0.31††	−0.03	0.40††	0.30††	0.17†
VHI total score	0.38††	−0.05	0.42††	0.35††	0.24††
VoiSS impairment	0.36††	−0.03	0.43††	0.32††	0.30††
VoiSS physical symptoms	0.02	−0.10	−0.07	0.04	0.03
VoiSS emotion	0.27††	0.01	0.32††	0.30††	0.15
VoiSS total score	0.34††	−0.03	0.37††	0.32††	0.28††

*The vocal performance questionnaire (VPQ), the vocal handicap index (VHI) and the voice symptom scale (VoiSS). ††Correlation significant at the 0.01 level (two-tailed); †correlation significant at the 0.05 level (two-tailed). GRBAS = grade–roughness–breathiness–aesthenia–strain

while the voice handicap index and the voice symptom scale correlated most strongly with breathiness (0.44 and 0.43, respectively). The physical symptoms domain of the voice symptom scale was not related to the grade–roughness–breathiness–aesthenia–strain rating scale.

Discussion

This study demonstrated that all three self-reported patient questionnaires were reliable and valid instruments for measuring the patient-perceived impact of a voice disorder. We consider that the relatively minor differences between the scales, with regard to coefficient sizes, are of limited significance.

The vocal performance questionnaire, voice handicap index and voice symptom scale had good internal consistency and test–retest reliability (Table II).[7,2] Criterion validity entails comparing the scale under review with an outside validating criterion or gold standard. The adopted criterion in this study was the grade–roughness–breathiness–aesthenia–strain perceptual rating scale. Previously, the vocal performance questionnaire had been correlated with an overall rating of severity of voice quality, comparable to 'grade' on the grade–roughness–breathiness–aesthenia–strain scale, in 45 patients with non-organic dysphonia, giving a Spearman rho coefficient of 0.65.[2] In the present study, the total score of the vocal performance questionnaire again correlated significantly (0.32) with the grade parameter of the grade–roughness–breathiness–aesthenia–strain scale. All the self-reported scales, with the exception of the physical symptoms domain of the voice symptom scale, correlated significantly with all the parameters of the grade–roughness–breathiness–aesthenia–strain scale, except roughness. This supports the theory that the self-reported and perceptual assessments are in part measuring the same underlying concept.

- **There are several self-reported voice quality research tools available**
- **Most studies report on only one such tool**
- **The comparative reliability and validity of different tools is not known**
- **The voice performance questionnaire, the voice handicap inventory and the voice symptom scale have good internal consistency and test–retest reliability**
- **In comparison with observer rating of voice performance, all three scales emerged as valid; the vocal performance questionnaire appeared adequate for a synopsis of voice outcomes, whereas the vocal handicap index may be superior for emotional domains**
- **The voice symptom scale physical symptom domain score seemed independent of the other self- and observer-reported ratings**

The highest correlations were demonstrated between the function/impairment domains of the voice handicap index and voice symptom scale, respectively, and the 'breathiness' parameter of the grade–roughness–breathiness–aesthenia–strain scale. This may indicate that air wastage through the glottis has the largest subjective impact on the patient's ability to carry out their normal activities.

https://doi.org/10.1017/S0022215107007177 Published online by Cambridge University Press

However, although statistically significant, none of the correlations was high. In other words, a clinician's perception of voice quality, as recorded at one point in time, does not directly correspond to the patient's perception of voice quality and its impact on their daily activities.[6]

Conclusion

There were strong correlations between the vocal performance questionnaire, the voice handicap index and the voice symptom scale, and aspects which address impairment or alteration of function. The voice symptom scale has an additional domain not reflected in the other scales, that of associated physical symptoms. The vocal performance questionnaire gives little indication of emotional effects, but, like the shortened version of the voice handicap index (the voice handicap index 10), is a convenient, internally consistent, uni-dimensional voice outcome tool.[2]

A voice assessment tool that addresses voice problems in terms of physical, functional and emotional impacts may provide a more accurate indication of the outcomes of a particular treatment package. For example, a functional approach to therapy may more adequately be informed by assessment with the voice handicap index, whilst a medical approach to the treatment of symptoms may benefit from the results of the more symptom-based voice symptom scale. However, if the aim of assessment is to obtain a brief, simple indication of severity of impact, in order to determine intervention outcomes and to audit service provision, then a shorter, more general measure (such as the vocal performance questionnaire) would be more appropriate.

Acknowledgements

This research was supported by a grant from the Wellcome Trust.

References

1 Carding P, Steen N, Webb A, Mackenzie K, Deary IJ, Wilson JA. The reliability and sensitivity to change of the acoustic quality of voice. *Clin Otolaryngol* 2004;**29**: 538–44
2 Deary IJ, Webb A, Mackenzie K, Wilson JA, Carding P. Short self report voice symptom scales: psychometric characteristics of the Voice Handicap – 10 and the Vocal Performance Questionnaire. *Otolaryngol Head Neck Surg* 2004;**131**:232–5
3 Jacobson BH, Johnson A, Grywalski C, Silbergleit A, Jacobson G, Benninger MS. The Voice Handicap Index (VHI): development and validation. *Am J Speech Lang Pathol* 1997;**6**:66–70
4 Gliklich RE, Glovsky RW, Montgomery WW. Validation of a voice outcome survey for unilateral vocal cord paralysis. *Otolaryngol Head Neck Surg* 1999;**120**:153–8
5 Hogikyan N, Sethuraman G. Validation of an instrument to measure Voice-Related Quality of Life (V-RQOL). *J Voice* 1999;**13**:557–69
6 Ma E, Yiu E. Voice activity and participation profile: assessing the impact of voice disorders on daily activities. *J Speech Lang Hear Res* 2001;**44**:511–24
7 Wilson JA, Webb A, Carding P, Steen IN, Mackenzie K, Deary I. Comparing the Voice Symptom Scale (VoiSS) and the Voice Handicap Index (VHI) structure and content. *Clin Otolaryngol* 2004;**29**:169–74
8 Carding P, Docherty GJ. A study of the effectiveness of voice therapy in the treatment of 45 patients with nonorganic dysphonia. *J Voice* 1999;**13**:72–104
9 Benniger MS, Ahuja AS, Gardner G, Grywalski C. Assessing outcomes for dysphonic patients. *J Voice* 1998;**12**:540–50
10 Rosen C, Murray T. Nomenclature of voice disorders and vocal pathology. *Otolaryngol Clin North Am* 2000;**33**: 1035–45
11 Scott S, Robinson K, Wilson JA, MacKenzie K. Patient reported problems associated with dysphonia. *Clin Otolaryngol* 1997;**2**:37–40
12 Carding P, Horsley I. An evaluation study of voice therapy in nonorganic dysphonia. *Eur J Disord Commun* 1992;**27**: 137–57
13 Stewart M, Chen A, Stach C. Outcome analysis of voice and quality of life in patients with laryngeal cancer. *Arch Otolaryngol Head Neck Surg* 1998;**124**:143–8
14 Rosen CA, Murray T, Zinn A, Zullo T, Sonbolian M. Voice Handicap Index change following treatment of voice disorders. *J Voice* 2000;**14**:619–23
15 Courey MS, Garrett CG, Billante CR, Stove RE, Portell MD, Smith TL *et al*. Outcomes assessment following treatment of spasmodic dysphonia with botulinum toxin. *Ann Otol Rhinol Laryngol* 2000;**109**:819–22
16 Benninger MS, Gardner G, Grywalski C. Outcomes of botulinum toxin treatment for patients with spasmodic dysphonia. *Arch Otolaryngol Head Neck Surg* 2001;**127**:1083–5
17 Fung K, Yoo J. Vocal function following radiation for non laryngeal versus laryngeal tumours of the head and neck. *Laryngoscope* 2001;**111**:1920–3
18 Hirano M. *Clinical Examination of Voice*. Vienna: Springer-Verlag, 1981
19 Dejonckere PH, Obbens C, Leeper HA, Hawkins S, Heeneman H, Doyle PC *et al*. Perceptual evaluation of dysphonia: reliability and relevance. *Folia Phoniatrica* 1993; **45**:76–83
20 De Bodt M, Wuyts FL, Van de Heyning PH, Croux C. Test-retest of the GRBAS Scale: influence of experience and professional background on perceptual ratings of voice quality. *J Voice* 1997;**11**:74–80
21 Wuyts FL, De Bodt MS, Van de Heyning PH. Is the reliability of a visual analog scale higher than an ordinal scale? An experiment with the GRBAS Scale for the perceptual evaluation of dysphonia. *J Voice* 1999;**13**:508–17
22 Webb A, Carding P, Deary IJ, Mackenzie K, Steen IN, Wilson JA. A study of the reliability of three auditory perceptual scales for dysphonia. *Eur Arch Otorhinolaryngol* 2004;**261**:429–34
23 Fairbanks G. *Voice and Articulation Drillbook*. New York: Harper Row, 1960
24 Cronbach L. *Essentials of Psychological Testing*. London, Harper & Row, 1970
25 Hays R, Anderson R, Revicki D. Psychometric considerations in evaluating health-related quality of life measures. *Qual Life Res* 1993;**2**:441–9
26 Shrout P, Fleiss J. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 1979;**86**:420–8
27 Nunally J. *Introduction to Psychological Measurement*. New York: McGraw-Hill, 1970
28 Schuavetti N, Metz D. *Evaluating Research in Communicative Disorders*. Boston: Allyn and Bacon, 1997

Address for correspondence:
Mr Kenneth MacKenzie,
Department of ORL-HNS,
Glasgow Royal Infirmary,
Glasgow G31 2ER,
Scotland, UK.

E-mail: kenneth.mackenzie@northglasgow.scot.nhs.uk

Professor J Wilson takes responsibility for the integrity of the content of the paper.
Competing interests: None declared