

Optimising outcome assessment of voice interventions, II: sensitivity to change of self-reported and observer-rated measures

I N STEEN, K MACKENZIE*, P N CARDING†, A WEBB‡, I J DEARY‡, J A WILSON†

Abstract

Objectives: A wide range of well validated instruments is now available to assess voice quality and voice-related quality of life, but comparative studies of the responsiveness to change of these measures are lacking. The aim of this study was to assess the responsiveness to change of a range of different measures, following voice therapy and surgery.

Design: Longitudinal, cohort comparison study.

Setting: Two UK voice clinics.

Participants: One hundred and forty-four patients referred for treatment of benign voice disorders, 90 undergoing voice therapy and 54 undergoing laryngeal microsurgery.

Main outcome measures: Three measures of self-reported voice quality (the vocal performance questionnaire, the voice handicap index and the voice symptom scale), plus the short form 36 (SF 36) general health status measure and the hospital anxiety and depression score. Perceptual, observer-rated analysis of voice quality was performed using the grade–roughness–breathiness–asthenia–strain scale. We compared the effect sizes (i.e. responsiveness to change) of the principal subscales of all measures before and after voice therapy or phonosurgery.

Results: All three self-reported voice measures had large effect sizes following either voice therapy or surgery. Outcomes were similar in both treatment groups. The effect sizes for the observer-rated grade–roughness–breathiness–asthenia–strain scale scores were smaller, although still moderate. The roughness subscale in particular showed little change after therapy or surgery. Only small effects were observed in general health and mood measures.

Conclusion: The results suggest that the use of a voice-specific questionnaire is essential for assessing the effectiveness of voice interventions. All three self-reported measures tested were capable of detecting change, and scores were highly correlated. On the basis of this evaluation of different measures' sensitivities to change, there is no strong evidence to favour either the vocal performance questionnaire, the voice handicap index or the voice symptom scale.

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

Introduction

Dysphonia is a multifactorial disorder which affects expressive communication, mood and general health status.¹ As discussed in the preceding paper,² a number of well validated measures is now available to assess these different domains. The responsiveness to change of these very different tools is, however, much less clear. There are comparatively few studies of the outcomes of dysphonia treatment, and most use limited outcome data sets in a highly selected patient group (for example, vocal fold medialisation). Furthermore, the number

of studies of surgical as opposed to voice therapy outcomes is exceedingly small. Our objective was to compare the sensitivity to change of a number of different self-reported, perceptual and quality of life measures following conservative (i.e. speech and language therapy) and surgical intervention in a large, heterogeneous group of voice patients.

The specific aims of the study were: (1) to estimate the responsiveness to change of self-reported and perceptual ratings of voice quality and voice-related quality of life, in a large cohort of heterogeneous

From the Institute of Health and Society and the †Department of Otolaryngology Head and Neck Surgery, Newcastle University, the *Department of Otolaryngology Head Neck Surgery, Royal Infirmary, Glasgow, and the ‡Department of Psychology, University of Edinburgh, Scotland, UK.

Accepted for publication: 24 January 2007.

dysphonic patients; and (2) to estimate the range of effect sizes for voice therapy and surgical interventions, using self-reported and observer-rated measures, in order to inform future prospective, controlled trials.

Methods

The approach adopted was to identify patients with voice-related problems and then, using a range of generic and voice-specific measures, to assess their quality of life before and after medical intervention.

Three self-reported voice scales were completed by patients. The vocal performance questionnaire³ consists of 12 items which address the physical aspects of the voice problem and its social and emotional impact, scored to give a total score. The voice handicap index⁴ is a 30-item questionnaire with questions grouped into three content domains, representing the functional, emotional and physical aspects of voice disorders. Its sensitivity to change in voice was evaluated on a sample of 37 subjects with various vocal fold abnormalities.⁵ The voice symptom scale⁶ is a 30-item scale with three content domains and a total score, the reliability and validity of which have been assessed in a series of studies involving over 800 subjects.⁷

Perceptual, observer-rated analysis of voice quality was performed using the grade-roughness-breathiness-asthenia-strain scale.⁸⁻¹⁰ All voices were recorded on digital audiotape, following a standard procedure, both before and after treatment.¹¹ The recorded voice sample included rote counting and speaking the days of the week, prolonged /a/ and /i/ vowels, and three sentences from the Rainbow Passage. The five grade-roughness-breathiness-asthenia-strain parameters were scored using a four-point rating scale, from zero (normal) to three (extreme). Each participant was scored on a standard pro forma by an independent expert rater. Independent raters were blinded to treatment group and treatment status, but aware of each patient's age and sex. The short form 36 (SF 36)¹² is an extensively validated, self-administered, 36-item questionnaire assessing quality of life. It has eight subscales and two global domains (mental health and physical health), and a large body of normative data is available.¹³ The SF 36 is known to be abnormal in patients with voice disorders.¹⁴

All patients also completed the hospital anxiety and depression scale.¹⁵

Patients

One hundred and forty-four patients complaining of hoarseness and attending out-patient clinics in Newcastle and Glasgow were assessed by the above measures, before and after intervention. The patients included a subgroup of patients described in our companion paper.² The patient exclusion criteria were: no intervention undertaken; defaulting from follow-up; laryngeal cancer; age less than 18 years; and impaired language or receptive communication skills. At the initial out-patient appointment, each participant completed the three self-reported voice questionnaires, the SF 36 and the hospital anxiety

and depression scale, and also had their voice recorded. Ninety patients received a course of speech and language therapy, while 54 patients underwent laryngeal surgery.

Analysis

For each questionnaire or rating scale, the effect size was defined as the change in mean score divided by the standard deviation of change scores. The effect size is independent of scale and sample size and can be used to make comparisons between the different questionnaires and different groups of subjects. It is accepted^{16,17} that values around 0.2 represent small effect sizes, values around 0.5 represent medium effect sizes and values around 0.8 represent large effect sizes. If subjects experience an improvement in quality of life, the outcome measure with the largest effect size is clearly the most sensitive to change.

Results

Table I shows the mean baseline and follow-up scores for each patient group, along with the mean improvement in quality of life (with a 95 per cent confidence interval (CI)) and an estimate of effect size. A paired *t*-test indicates those improvements that are statistically significant.

Both groups of subjects reported medium to large improvements on all three voice questionnaires. The smallest changes were in the emotional subscale of the voice handicap index (effect sizes = 0.44 and 0.48 for speech and language therapy and surgery groups, respectively) and in the physical symptoms subscale of the voice symptom scale (effect sizes = 0.38 and 0.43 for speech and language therapy and surgery groups, respectively). The largest changes in the individual scales were in the physical aspects of voice subscale of the voice handicap index (effect sizes = 0.71 and 0.81 for speech and language therapy and surgery groups, respectively) and in the voice impairment subscale of the voice symptom scale (effect sizes = 0.78 and 1.00 for speech and language therapy and surgery groups, respectively). The effect sizes corresponding to the change in the total score, respective to speech and language therapy and surgery, for each of the three voice questionnaires, were: vocal performance questionnaire, 1.04 and 0.82; voice handicap index, 0.62 and 0.72; and voice symptom scale, 0.78 and 1.06. The two patient groups were very similar both at baseline and follow up (Figure 1), with no significant differences between them.

The changes in the three voice questionnaires were highly correlated, as follows: vocal performance questionnaire *vs* voice symptom scale, 0.74 (95 per cent CI: 0.65, 0.81); vocal performance questionnaire *vs* voice handicap index, 0.76 (95 per cent CI: 0.68, 0.83); and voice symptom scale *vs* voice handicap index, 0.83 (95 per cent CI: 0.76, 0.87). All differences were significant ($p < 0.0001$). The correlation between the voice symptom scale and the voice handicap index was greater, due in part to the small number of shared items between the two questionnaires. Changes in the subscale components of the voice

TABLE I
RESPONSIVENESS TO CHANGE OF VOICE-RELATED QUALITY OF LIFE MEASURES

Scale & subscale	Group	Baseline mean (SD)	Follow-up mean (SD)	QOL improvement				
				Mean*	95% CI	Effect size [†]	<i>t</i>	<i>p</i>
<i>VPQ</i>								
Total	SLT	32.3 (9.2)	21.8 (6.8)	10.5	8.3, 12.6	1.04	9.67	<0.001
	Surg	32.2 (9.2)	22.9 (8.3)	9.3	6.1, 12.4	0.82	5.92	<0.001
<i>VHI</i>								
Physical aspects	SLT	21.4 (6.7)	15.3 (8.2)	6.1	4.3, 7.9	0.71	6.72	<0.001
	Surg	19.4 (7.3)	13.1 (8.2)	6.3	4.2, 8.5	0.81	5.89	<0.001
Functional aspects	SLT	14.4 (9.0)	9.7 (8.1)	4.8	2.9, 6.6	0.54	5.16	<0.001
	Surg	14.6 (9.4)	10.4 (7.4)	4.2	2.3, 6.1	0.61	4.41	<0.001
Emotional aspects	SLT	11.8 (9.5)	7.9 (8.4)	3.9	2.05, 5.8	0.44	4.16	<0.001
	Surg	12.5 (10.4)	7.9 (7.4)	4.5	1.9, 7.1	0.48	3.51	<0.01
Total	SLT	47.8 (22.8)	32.9 (22.7)	14.9	9.9, 19.9	0.62	5.92	<0.001
	Surg	46.3 (24.6)	31.4 (21.2)	15.0	9.3, 20.8	0.72	5.24	<0.001
<i>VoiSS</i>								
Impairment	SLT	30.7 (11.8)	20.1 (11.7)	10.5	7.7, 13.4	0.78	7.27	<0.001
	Surg	32.2 (11.7)	20.4 (12.3)	11.8	8.5, 15.2	1.00	7.12	<0.001
Physical symptoms	SLT	9.7 (4.9)	8.1 (5.5)	1.6	0.7, 2.4	0.38	3.56	<0.01
	Surg	9.0 (5.0)	7.6 (5.2)	1.4	0.5, 2.3	0.43	3.15	<0.01
Emotion	SLT	8.2 (6.7)	4.7 (5.5)	3.4	2.04, 4.8	0.52	4.92	<0.001
	Surg	9.3 (8.0)	4.7 (5.5)	4.6	2.8, 6.5	0.69	4.99	<0.001
Total	SLT	48.4 (18.1)	32.5 (19.2)	16.0	2.2, 20.4	0.78	7.16	<0.001
	Surg	50.4 (19.9)	32.4 (19.5)	18.0	13.2, 22.8	1.06	7.57	<0.001
<i>GRBAS</i>								
Grade	SLT	1.9 (1.1)	1.3 (0.7)	0.7	0.4, 0.9	0.55	5.09	<0.01
	Surg	2.1 (1.1)	1.7 (0.8)	0.4	0.09, 0.8	0.35	2.53	0.01
Roughness	SLT	1.1 (0.9)	0.9 (0.8)	0.2	-0.05, 0.4	0.16	1.56	0.12
	Surg	1.2 (0.9)	0.9 (0.8)	0.3	0.01, 0.6	0.29	2.09	0.04
Breathiness	SLT	1.5 (0.9)	1.0 (0.6)	0.5	0.3, 0.7	0.57	5.21	<0.001
	Surg	1.8 (0.9)	1.5 (0.9)	0.3	0.03, 0.6	0.32	2.23	0.03
Asthenia	SLT	1.1 (1.0)	0.6 (0.6)	0.5	0.3, 0.7	0.49	4.58	<0.001
	Surg	1.4 (0.9)	1.2 (0.9)	0.2	-0.03, 0.5	0.25	1.77	0.08
Strain	SLT	0.9 (0.8)	0.4 (0.6)	0.5	0.3, 0.6	0.56	5.13	<0.001
	Surg	1.1 (0.8)	1.2 (0.9)	-0.1	-0.4, 0.2	-0.13	-0.90	0.37
<i>SF 36</i>								
Physical component	SLT	35.2 (14.2)	38.1 (17.1)	2.9	0.6, 5.3	0.26	2.50	0.01
	Surg	41.6 (14.3)	39.5 (15.6)	-2.1	-4.5, 0.4	-0.23	-1.67	0.10
Mental component	SLT	45.7 (12.5)	47.2 (12.1)	1.5	-0.7, 3.7	0.15	1.37	0.18
	Surg	45.7 (12.6)	48.4 (12.0)	2.7	-0.07, 5.4	0.27	1.96	0.06
<i>HAD</i>								
Depression	SLT	4.3 (3.4)	3.7 (3.7)	0.6	-0.02, 1.1	0.21	1.91	0.06
	Surg	4.5 (3.6)	4.1 (3.7)	0.4	-0.5, 1.3	0.12	0.88	0.39
Anxiety	SLT	7.2 (4.3)	6.6 (4.7)	0.6	-0.08, 1.3	0.19	1.77	0.08
	Surg	7.1 (4.0)	6.3 (4.6)	0.8	-0.1, 1.7	0.24	1.71	0.09

Data represent scores for the various measures. *Mean change from baseline in the direction consistent with an improvement in quality of life (QOL) (a negative score implies a deterioration in quality of life). [†]Effect sizes = improvement in quality of life divided by the standard deviation (SD) of change scores. CI = confidence intervals; VPQ = vocal performance questionnaire; SLT = speech and language therapy; surg = surgery; VHI = voice handicap index; VoiSS = voice symptom scale; GRBAS = grade-roughness-breathiness-asthenia-strain scale; SF 36 = short form 36; HAD = hospital anxiety and depression scale

handicap index and the voice symptom scale were also significantly correlated. The greatest correlation was between the voice handicap index physical aspects of health subscale and the voice symptom scale impairment subscale ($\rho = 0.76$; 95 per cent CI: 0.67, 0.82). Changes in the three voice handicap index subscales were correlated with each other (all correlations > 0.6). Changes in the three subscales of the voice symptom scale were less strongly correlated with each other (correlations between 0.2 and 0.6). The weakest correlations were between the voice symptom scale physical symptom subscale and all the other subscales (correlations between 0.2 and 0.4).

For both groups of subjects, there was some evidence of change on the grade-roughness-breathiness-asthenia-strain scale, but the effect sizes were much smaller than those observed with the self-reported

measures. For subjects undergoing speech and language therapy, there were small or medium effects in each grade-roughness-breathiness-asthenia-strain component. For subjects undergoing surgery, the effect sizes were smaller in all components except for roughness, the least sensitive component.

There was little evidence of substantive change in any of the generic health status instruments – all effect sizes were less than 0.3. Thus, although some of the changes were statistically significant, the small effect sizes suggest that they were not clinically important.

Discussion

The pattern of effect sizes observed in Table I indicates very small changes in the generic health

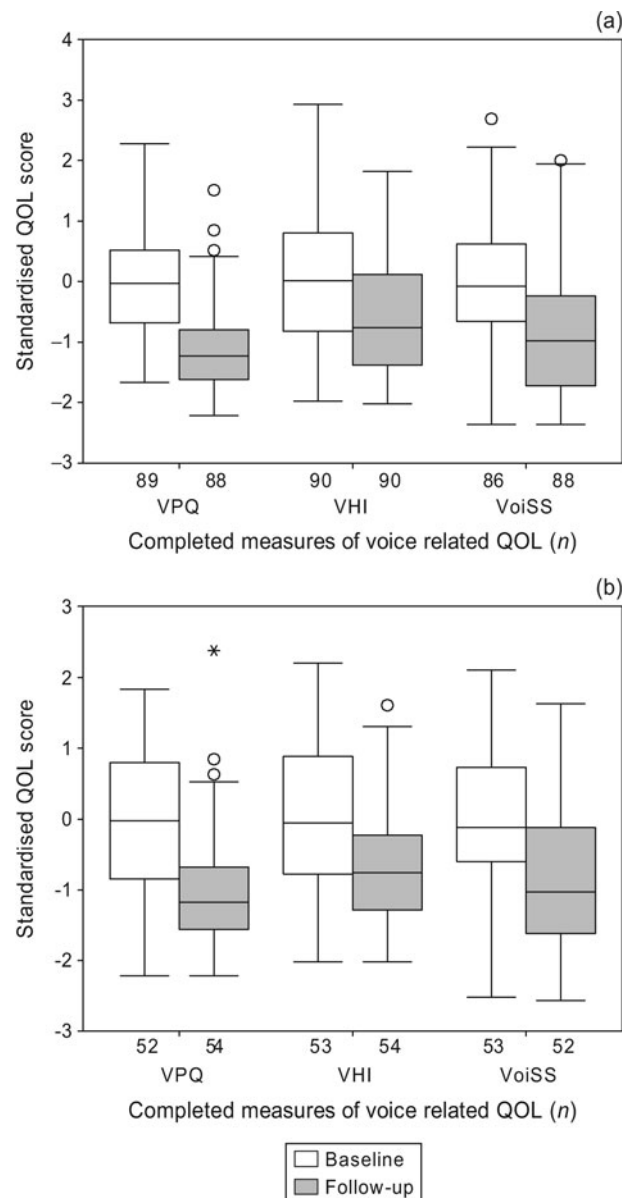


FIG. 1

Box and whisker plots of pre- and post-intervention scores for the three self-reported questionnaires completed by (a) the speech and language therapy group, and (b) the surgery group. Scores have been numerically standardised for comparison. The x axis shows the number of fully completed questionnaire sets, by questionnaire. The lower and upper edges of each box represent the 25th and 75th percentiles, respectively. Medians are indicated by the horizontal line within each box. The range is denoted by the whiskers; individual outliers are indicated as circles. QOL = quality of life; VPQ = vocal performance questionnaire; VHI = voice handicap index; VoiSS = voice symptom scale

status instruments in comparison with the voice-specific measures. This supports the conclusion of a previous study which assessed voice therapy alone.¹⁸ Following intervention, there were fairly large changes in self-reported, voice-related quality of life. Similar effect sizes were observed across the three self-reported voice questionnaires. The changes in total scores were highly correlated, suggesting that the three self-reported voice questionnaires were detecting changes in the same sets of patients.

When considering the voice handicap index and the voice symptom scale component scores, greatest change was observed in the voice symptom scale

impairment subscale and in the voice handicap index physical aspects of voice subscale. Inspection of the individual items that make up these two subscales suggests that they are more or less equivalent; both relate to voice quality. The very high correlation between the changes in these two subscales supports this suggestion. The remaining two voice handicap index components (functional and emotional aspects of health) broadly overlap with the voice symptom scale emotion component; this is reflected in the correlation in change scores between these components. The voice symptom scale physical symptom component has almost no overlap with any of the voice handicap index components; this is reflected in the

much weaker correlations of this subscale with the voice handicap index subscales, and indeed with the other voice symptom scale components.

The grade-roughness-breathiness-asthenia-strain components were less sensitive to change than the self-reported measures. Roughness did not alter much, and was also the only parameter which consistently failed to show any relationship with the total or subscales of any of the three self-reported measures.²

The participants reported comparatively little change in the generic health status questionnaires; a much larger change was observed in the self-reported voice scales. The SF 36 and other generic quality of life instruments have proven validity and reliability, and it is reasonable to suppose that any general, subjective increase in wellbeing would be reflected in these instruments. Similarly, there was little change in the mood variables (the hospital anxiety and depression scales and the SF 36 mental health component score). This suggests that the changes observed in the voice measures were independent of a general improvement in subjective wellbeing as a result of receiving a medical intervention, but rather reflected real improvement in voice-related quality of life. At the same time, the lack of change in the generic measures also highlights the need for a voice-specific questionnaire in assessing voice treatment effectiveness; the SF 36 does not include voice-sensitive components.

- **Several self-reported voice tools seem responsive to change**
- **The differing sensitivity of the various available tools is not known**
- **Identification of the effect sizes of different interventions according to different tools would be useful in the standardisation of interventions and in clinical voice outcome reporting**
- **The three self-reported voice tools studied (the vocal performance questionnaire, the voice handicap index and the voice symptom scale) all showed large effect sizes following either voice therapy or phonosurgery**
- **Smaller effect sizes were noted for the grade-roughness-breathiness-asthenia-strain perceptual rating scale, administered by an expert rater**
- **General health status measures were the least responsive to change**

Our work suggests that all three self-reported voice measures are capable of detecting change. There is no strong evidence, on the basis of sensitivity to change evaluation, to favour one measure over the other two. Whichever one is chosen, the means and standard deviations given in Table I can be used to help determine sample size when planning

an evaluation of an intervention. In the present, large sample of patients with benign disorders, the vocal performance questionnaire and the voice symptom scale showed the largest overall effect sizes, while the voice handicap index sensitivity was somewhat lower for both conservative and surgical interventions. The voice handicap index was, however, developed partly in laryngectomy patients; therefore, the responsiveness pattern may well change in subjects with laryngeal malignancy. Figure 1 shows that, especially for the surgery group, the vocal performance questionnaire and the voice handicap index inter-quartile ranges shrank post-treatment. This possibly supplies evidence for a floor effect in these questionnaires, which was not seen in the voice symptom scale.

The data shown here are among the first to record the approximate level of benefit from a heterogeneous group of phonosurgical patients. There was no prior expectation as to the relative effects of speech and language therapy and surgery. Both interventions produced very similar changes in self-reported voice quality, although the expert raters recorded larger improvements in perceived voice quality in the group undergoing speech and language therapy. The study was, however, designed principally to assess the behaviour of the clinical outcomes studied, rather than to compare different subgroups of intervention. The groups were not prospectively matched at baseline, either for disease severity or for diagnostic spread; no treatment comparison inference is therefore possible. Nonetheless, the results do provide useful evidence that the measures used can in future be applied across a range of interventions likely to affect voice-related quality of life.

Acknowledgements

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

References

- 1 Millar A, Deary IJ, Wilson JA, MacKenzie K. Is an organic/functional distinction psychologically meaningful in patients with dysphonia? *J Psychosom Res* 1999;**46**:497-505
- 2 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;1-5 [Epub ahead of print]
- 3 Deary IJ, Webb A, MacKenzie K, Wilson JA, Carding PN. Short self-report voice symptom scales: psychometric characteristics of the VHI-10 and the VPQ. *Otolaryngol Head Neck Surg* 2004;**131**:232-5
- 4 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
- 5 Benniger MS, Ahuja AS, Gardner G, Grywalski C. Assessing outcomes for dysphonic patients. *J Voice* 1998;**12**:540-50
- 6 Deary IJ, Wilson JA, Carding PN, MacKenzie K. VoiSS: a patient derived voice symptom scale. *J Psychosom Res* 2003;**54**:483-9
- 7 Wilson JA, Webb AL, Carding PN, Steen IN, MacKenzie K, Deary IJ. Comparing the Voice Symptom Scale (VoiSS) and the Voice Handicap Index: structure and content. *Clin Otolaryngol* 2004;**29**:169-74
- 8 Hirano M. *Clinical Examination of Voice*. New York: Springer-Verlag, 1981

- 9 Dejonckere PH, Obbens C, de Moor GM, Wieneke GH. Perceptual evaluation of dysphonia: reliability and relevance. *Folia Phoniatrica* 1993;**45**:76–83
- 10 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
- 11 Webb AL, Carding PN, Deary IJ, MacKenzie K, Steen N, Wilson JA. The reliability of three perceptual evaluation scales for dysphonia. *Eur Arch ORL* 2004;**261**: 429–34
- 12 Jenkinson C, Coulter A, Wright L. Short-form SF-36 health survey questionnaire: normative data for adults of working age. *Br Med J* 1993;**306**:1437–40
- 13 Brazier JE, Harper R, Jones NMB, O’Cathain A, Thomas KJ, Usherwood T *et al.* Validating the SF-36 health survey questionnaire: a new outcome measure for primary care. *BMJ* 1992;**305**:160–4
- 14 Wilson JA, Millar A, Deary IJ, MacKenzie K. The quality of life impact of dysphonia. *Clin Otolaryngol* 2002;**27**: 179–82
- 15 Zigmond AS, Snaith RP. The hospital anxiety and depression scale. *Acta Psychiatrica Scand* 1983;**67**:361–70
- 16 Cohen J. *Statistical Power Analysis for the Behavioral Sciences*, 2nd edn. Hillsdale, NJ: Erlbaum, 1988
- 17 Kazis LE, Anderson JJ, Meenen RF. Effect sizes for interpreting changes in health status. *Med Care* 1989; **27**(3 Suppl):S178–89
- 18 MacKenzie K, Millar A, Sellars C, Wilson JA, Deary IJ. Is voice therapy an effective treatment for dysphonia? A randomised controlled trial. *BMJ* 2001;**323**:658–61

Address for correspondence:
Prof. Janet A Wilson,
Dept of Otolaryngology Head Neck Surgery,
Freeman Hospital,
Newcastle upon Tyne NE7 7DN, UK.

Fax: (44) 191 223 1246
E-mail: j.a.wilson@ncl.ac.uk

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