The Journal of Laryngology & Otology (2009), ***123***, *755–763.*

doi:10.1017/S0022215109004605

Main Article

Reliability of perceptions of voice quality: evidence from a problem asthma clinic population

C Sellars, A E Stanton*†, A McConnachie‡, C P Dunnet, L M Chapman, C E Bucknall*, K MacKenzie**

Abstract

Introduction: Methods of perceptual voice evaluation have yet to achieve satisfactory consistency; complete acceptance of a recognised clinical protocol is still some way off.

Materials and methods: Three speech and language therapists rated the voices of 43 patients attending the problem asthma clinic of a teaching hospital, according to the grade-roughness-breathiness-asthenicity-strain (GRBAS) scale and other perceptual categories.

Results and analysis: Use of the GRBAS scale achieved only a 64.7 per cent inter-rater reliability and a 69.6 per cent intra-rater reliability for the grade component. One rater achieved a higher degree of consistency. Improved concordance on the GRBAS scale was observed for subjects with laryngeal abnormalities. Raters failed to reach any useful level of agreement in the other categories employed, except for perceived gender.

Discussion: These results should sound a note of caution regarding routine adoption of the GRBAS scale for characterising voice quality for clinical purposes. The importance of training and the use of perceptual anchors for reliable perceptual rating need to be further investigated.

Key words: Asthma; Reproducibility of Results; Voice Quality

Introduction

Formal and informal perceptual evaluation of speakers' voices is a well established field of study. Early (and later) studies demonstrated negative listener reaction to dysphonic voices.[1–5] Davis and Harris showed that elementary school teachers could identify children with disordered voices, a skill that might be useful for screening purposes.[6] Bonet and Casan proposed perceptual screening for dysphonia as a means of identifying a population (child choristers) at risk of vocal pathology.[7] Perceptual analysis remains an important feature of vocal assessment, both in research and clinical practice, and is endorsed by some as a necessary counterpart to other measures of voice, such as acoustic analysis.[8–13] However, there continues to be an issue regarding variability in listener perception in relation to normal and pathological voices.[14,15] Indeed, the collective work of Kreiman's group has led them to conclude that, '…pathologic voice quality assessment using traditional perceptual labels [viz 'breathy', 'rough', etc] is not generally useful'.[16]

The GRBAS scale is a perceptual rating scale widely reported in voice research.[17] Using a four-point scale (where zero = normal and three = severe), this system characterises the voice according to grade (i.e. overall severity), roughness, breathiness, asthenicity and strain. Some authors have proposed the adoption of this scale for routine clinical use.[10,13] Various studies have considered the clinical utility of the scale, including its rater reliability and method of presentation (i.e. ordinal versus visual analogue scale).[12,18–22] Other authors have employed the scale, fully or partially, as a ratings tool, and have incidentally reported on rater reliability.[23,24]

The raters employed in such studies have varied in number from two[13,18,22] to 28.[25] Some raters have been specially 'trained' in use of the scale,[26] while others have been more or less experienced in the field of clinical voice management, being variously patients,[25] students,[24] speech and language therapists and pathologists,[12,23,25,27] and ENT surgeons and phoniatricians.[19–22]

The number of voices rated in these studies has ranged from nine[19] to 943.[21] Some studies have included normal voices.[13,23,24,27] Where dysphonic voices have been rated, there has generally been limited detail on background pathology.[23,24] Only

From the Departments of Speech and Language Therapy, *Respiratory Medicine and **Otolaryngology, Glasgow Royal Infirmary, the ‡Robertson Centre for Biostatistics, University of Glasgow, Scotland, and the †Osler Chest Unit, Churchill Hospital, Oxford, UK.
Accepted for publication: 27 October 2008. First published online 2 March 2009.

Dejonckere *et al.* have reported the severity of ratings according to the broad category of vocal pathology involved; they noted an incomplete pattern of increasing severity of GRBAS scores for patients with functional problems, vocal fold nodules, benign tumours, vocal fold paralysis and malignant tumours.[20]

Specific rating trends have been reported, including: highest agreement for overall severity (i.e. grade);[13,19–22] greater test–retest reliability among more experienced raters and (for some sub-scales) among speech and language pathologists compared with ENT surgeons;[19] more severe rating by patients themselves compared with speech and language therapists;[25] and improved inter-rater agreement with sustained use of the GRBAS scale over time.[21] However, it is of concern that while some groups report gratifyingly high rates of inter- and intra-rater agreement (e.g. near-perfect agreement among experienced speech and language pathologists in Murry and colleagues' study,[27] and $\geq$0.92 per cent agreement for all GRBAS scoring among speech and language pathology students in Piccirillo and colleagues' study),[24] rater reliability continues to be a significant issue (with some groups reporting the highest κ value for inter-rater reliability as no better than 'moderate' (for overall grade)).[19,21]

Inhaled asthma medication has the potential to directly affect the larynx, and therefore asthma, or its treatment, may have a direct effect on patients' voices. Up to 50 per cent of patients taking inhaled corticosteroids may suffer from dysphonia[28] which is usually reversible.[29] This has been attributed to fungal infection or steroid-induced adductor myasthenia of the larynx,[30] although laryngoscopy or voice laboratory assessment may reveal more complicated abnormalities such as apposition abnormalities and cycle to cycle irregularity.[31–34]

The present study formed a component of a larger study characterising the vocal quality, self-perception of vocal morbidity, and laryngeal and nasal appearances in patients attending a problem asthma clinic, the results of which have been reported in summary form elsewhere.[35–38] The current study aimed to add to the general sum of knowledge in the field of voice perception, while highlighting issues of rater reliability in relation to the GRBAS scale and potentially outlining vocal features of a more specific patient population.

Materials and methods

Patients were recruited to the study from a problem asthma clinic based in a central teaching hospital. All patients attending the problem asthma clinic were eligible for inclusion in the study. Initially, 121 letters of invitation to take part in the study were sent to patients attending the clinic. If no response was obtained, attempts were made (by telephone or during clinic visits) to reiterate our invitation. Additional patients from the clinic were invited to participate. Sixty patients agreed to take part in the study (17 of whom subsequently withdrew) and 27 declined. Further attempts to contact remaining patients for recruitment were unsuccessful. Forty-three patients were ultimately included in the protocol, which involved attendance on a single afternoon. This study was conducted in accordance with the recommendations of the Helsinki Declaration of 1975 and was approved by the North Glasgow University Hospitals NHS Trust local research and ethics committee (research and ethics committee reference number 03RE002). All patients gave written, informed consent for their participation in the study.

Patient investigation included separate nasal and laryngological examinations. The latter were undertaken by a single otolaryngologist observer using a fibre-optic laryngoscope. Following topical application of co-phenylcaine to the nose, each patients' larynx and laryngopharynx were examined. Laryngeal assessment was based on structure and function. Laryngeal appearance was noted, along with the mobility of the vocal folds on phonation, inspiration and expiration. Findings were documented as normal or abnormal, with the latter category being further subdivided into organic or functional abnormalities. Patients also completed the voice symptom scale, a 30 item questionnaire which has been thoroughly evaluated in the self-assessment of voice quality.[39]

Voice recordings were then undertaken in a soundproof booth housed within the otolaryngology department. Recordings were made using digital audio tape in a digital tape recorder. Patients were asked to speak approximately 10 to 15 cm away from the microphone. They were asked to state their name and to engage in a few seconds of simple, spontaneous speech (topics suggested by the researcher included how they had got to the hospital that day, what they had watched on television the previous evening, etc) before reading the standard 'rainbow passage'.[40] These recordings were made by one of two independent observers who were not involved in any further data analysis.

After all patients' recordings had been made, the recordings were transferred onto two compact discs (CDs) by the medical illustration department, for review by the raters. All stimuli were randomised and then further assigned in a different order on each CD. Each patient's recording therefore corresponded to an individual track on each CD. A master list was kept in which the track numbers were linked to patient names; this list was not seen by the raters.

The raters (A, B and C) were three experienced speech and language therapists who were already very familiar with the GRBAS scale, both in their daily work and from earlier rating exercises undertaken in response to the lack of formalised training for this scale.[10] In an attempt to ensure optimal inter-rater agreement for the study, the raters engaged in pre-rating discussion around their individual understanding of the GRBAS scale and additional categories, and undertook listening exercises with recorded (non-study) pathological voices until consensus was reached.

The raters then graded the study patients' voices according to the GRBAS scale, with a further

assessment of fluctuations in voice quality (i.e. instability).[18,21,41] In addition, the speech and language therapists rated: audible respiration (reflecting any auditorily detectable respiration whether of laryngeal or lung origin) on an ad hoc, zero to three scale (interpreted as per the GRBAS rating scores); pitch range (as reduced or normal); overall pitch height (as low, medium or high); perceived gender (as male or female); and perceived age (by decade). Each CD was listened to and independently rated on two occasions at least seven days apart.

Patients' total GRBAS scores were calculated without using the instability component, as this parameter was not in widespread use. Mean values and standard deviations of the total score, the five component items and the two additional items were reported for all recordings rated, and separately for each rater and for each of the two rating occasions. Analysis of variance (ANOVA) methods were used to test for differences between the average ratings of the three raters (controlling for between-patient and between-occasion differences) and between the two rating occasions (controlling for between-patient and between-rater differences).

Inter- and intra-rater reliability coefficients were calculated using the methods of the generalisability theory,[42] using random effects ANOVA models to estimate the components of variance; the patient × occasion variance terms were assumed to be zero in all models, since the same recording was rated on each rating occasion. Inter-rater reliability estimates were also estimated separately for the two rating occasions, and intra-rater reliability estimates were estimated separately for each rater. Bootstrap methods were used to construct 95 per cent confidence intervals (CIs) for all reliability estimates, and to test for differences in reliability between raters or between rating occasions, based on 10 000 bootstrap samples from the 43 patients.[43]

In order to compare subgroups of patients, mean values for the total GRBAS scale, for each scale item and for the two additional items were calculated for each patient using the six scores (i.e. from three raters on two occasions). SPlus® for Windows (version 6.1) and Minitab™ (version 14) software was used to perform all calculations.

Results and analysis

Mean ratings

Table I shows means and standard deviations (SDs) for the total GRBAS scale scores, the five component items and the two additional items. There were significant differences between the mean ratings allocated by the three raters for all scale items ($p = 0.014$ for grade, and $p < 0.0001$ for all other items). Rater A tended to give higher ratings, except for breathiness and asthenicity. There was a trend among all three raters towards reduction in mean rating on the second occasion. For the asthenicity scale item, this reduction was significant ($p = 0.0026$), with weaker evidence for ratings of strain ($p = 0.046$) and total GRBAS scores ($p = 0.033$).

Reliability

Table II shows the inter- and intra-rater reliability estimates, with 95 per cent CIs, for the total GRBAS score, the five component items and the two additional items. Figure 1 shows the estimates of inter-rater reliability separately for each rating occasion, and the estimates of intra-rater reliability separately for each rater.

The total GRBAS scores showed good overall performance, with inter-rater reliability estimated as 78.1 per cent (95 per cent CI: 66.6, 88.2 per cent) and intra-rater reliability as 81.8 per cent (95 per cent CI: 68.5, 90.6 per cent). However, this level of reliability was not sustained across the individual scale items. The reliability of the grade item appeared best, with an overall inter-rater reliability of 64.7 per cent and an intra-rater reliability of 69.6 per cent. Asthenicity achieved the lowest inter- and intra-rater reliability estimates, at 43.4 and 49.6 per cent, respectively. The optional item instability had an overall inter-rater reliability of only 50.5 per cent; for this item, there was a marked improvement in inter-rater reliability, comparing the first assessment (35.4 per cent) and the second (65.9 per cent), an increase of 30.5 per cent (95 per cent CI: 13.9, 52.7 per cent). None of the other items showed a statistically significant difference in inter-rater reliability between the two measurement occasions. The instability item had rather better

TABLE I

SCORES FOR GRBAS SCALE TOTAL, INDIVIDUAL GRBAS ITEMS AND ADDITIONAL ITEMS, BY RATER AND RATING OCCASION

Parameter	Overall	Rater			p^*	Occasion		p^*
		A	B	C		1	2	
GRBAS total score	3.60 (3.34)	4.63 (3.26)	2.94 (2.78)	3.24 (3.69)	<0.0001	3.81 (3.29)	3.40 (3.38)	0.033
GRBAS item scores								
Grade	0.59 (0.67)	0.69 (0.66)	0.57 (0.64)	0.51 (0.72)	0.014	0.61 (0.67)	0.57 (0.68)	0.35
Roughness	0.88 (0.67)	1.15 (0.58)	0.78 (0.71)	0.72 (0.64)	<0.0001	0.88 (0.66)	0.88 (0.69)	1.00
Breathiness	0.43 (0.69)	0.41 (0.66)	0.24 (0.51)	0.63 (0.81)	<0.0001	0.47 (0.70)	0.39 (0.68)	0.17
Asthenicity	0.27 (0.52)	0.21 (0.46)	0.13 (0.34)	0.48 (0.65)	<0.0001	0.34 (0.58)	0.20 (0.44)	0.0026
Strain	0.62 (0.75)	0.97 (0.74)	0.51 (0.72)	0.40 (0.67)	<0.0001	0.68 (0.75)	0.57 (0.75)	0.046
Additional item scores								
Instability	0.24 (0.52)	0.52 (0.65)	0.07 (0.30)	0.14 (0.44)	<0.0001	0.28 (0.53)	0.21 (0.51)	0.090
Audible respiration	0.59 (0.81)	0.71 (0.84)	0.66 (0.82)	0.41 (0.76)	<0.0001	0.58 (0.82)	0.60 (0.81)	0.67

Data are shown as mean (standard deviation). *F test with analysis of variance, for differences between raters and between occasions, controlling for between-patient differences. GRBAS = grade-roughness-breathiness-asthenicity-strain scale

TABLE II

INTER- AND INTRA-RATER RELIABILITY ESTIMATES FOR GRBAS TOTAL SCORES, INDIVIDUAL GRBAS ITEM SCORES AND ADDITION ITEM SCORES

Parameter	Reliability estimate (95% CIs)	
	Inter-rater	Intra-rater
GRBAS total score	78.1 (66.6, 88.2)	81.8 (68.5, 90.6)
GRBAS item scores		
Grade	64.7 (45.7, 82.4)	69.6 (51.4, 85.1)
Roughness	45.3 (29.4, 61.7)	56.3 (41.1, 73.1)
Breathiness	52.8 (32.5, 69.4)	62.4 (44.6, 74.5)
Asthenicity	43.4 (26.9, 63.4)	49.6 (29.7, 69.5)
Strain	54.4 (35.7, 75.4)	68.7 (52.2, 85.1)
Additional item scores		
Instability	50.5 (24.7, 76.9)	72.2 (52.7, 85.3)
Audible respiration	70.2 (58.9, 80.7)	70.8 (57.4, 80.8)

GRBAS = grade-roughness-breathiness-asthenicity-strain scale; CIs = confidence intervals

intra-rater reliability, at 72.2 per cent. Audible respiration achieved 70.2 per cent inter-rater reliability and 72.2 per cent intra-rater reliability overall, although there was considerable variability among raters.

In fact, there was variability between the raters for many scale items, in terms of the level of intra-rater reliability achieved (Figure 1). Rater C achieved the highest degree of consistency for all scale items, dropping no lower than 62.3 per cent. Rater A, by comparison, was the least reliable on all but the strain item, with a low of 24.5 per cent for asthenicity. The CIs for individual raters' reliability estimates were wide, and the only differences between raters which reached statistical significance were between raters C and A on grade, asthenicity and audible respiration.

Separate analyses were undertaken for inter- and intra-rater reliability (for grade, roughness, breathiness, asthenicity, strain, instability and audible respiration assessment) for subjects with laryngeal abnormalities or functional laryngeal problems (see Table III). Results for this sub-group analysis showed a strong tendency to greater inter-rater reliability,

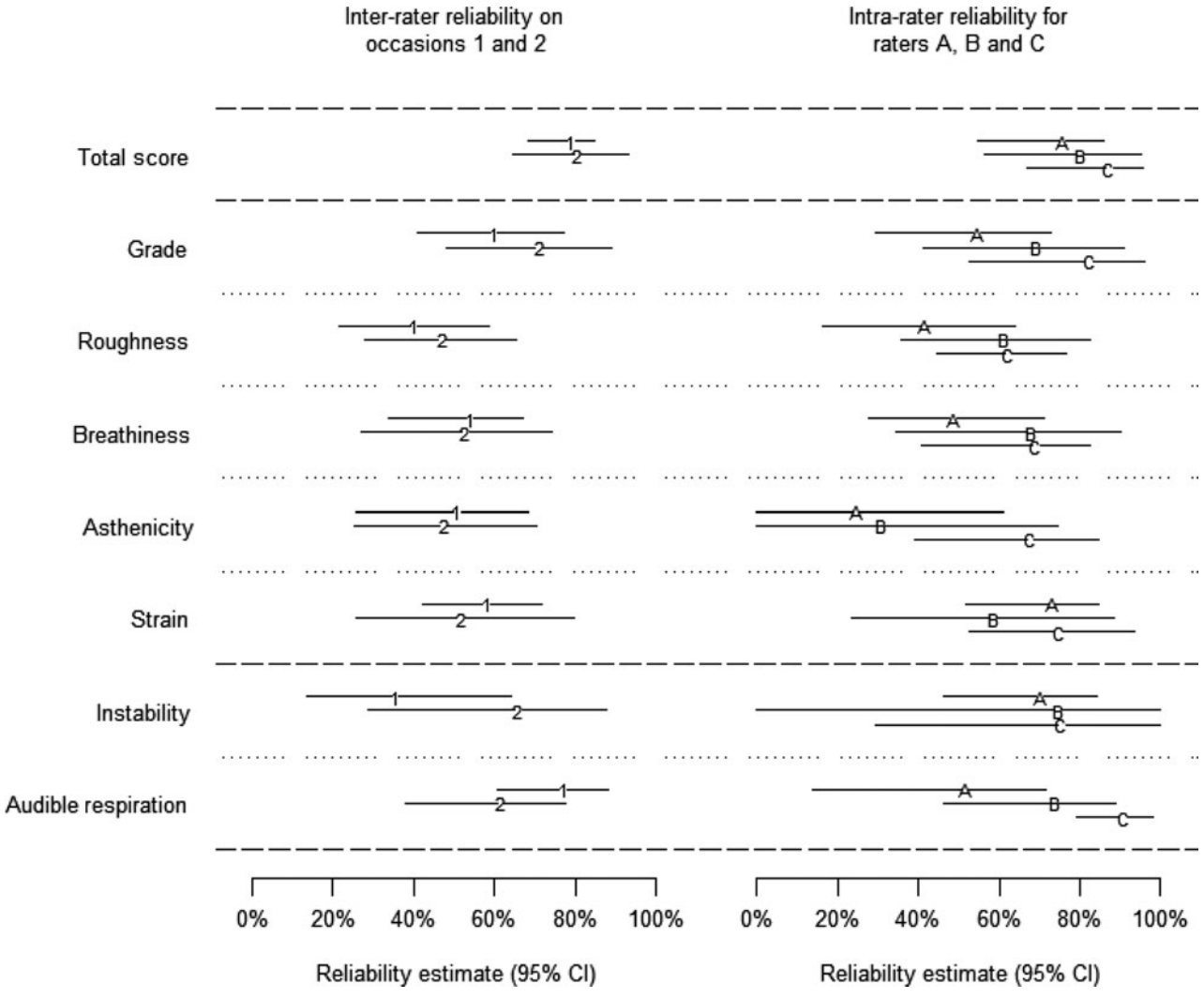

FIG. 1

Inter-rater reliability estimates on each rating occasion, and intra-rater reliability estimates for each rater, with 95% confidence intervals (CIs), for the GRBAS scale total score, individual item scores and additional item scores.

with improved consistency ranging between 5.5 per cent (for total score) and 29.5 per cent (for instability); however, no statistically significant differences were observed. The only exceptions to this trend were the reduced consistency rates for roughness and audible respiration (which were 5.6 and 1.7 per cent, respectively, in the other direction). Intra-rater reliability was significantly better for grade assessment in patients with any laryngeal abnormality, compared with patients with normal laryngeal appearance (78.8 versus 47.3 per cent, respectively; 95 per cent CI for difference: 6.2 to 59.1 per cent). A similar trend was apparent for other categories. In this instance, improved consistency ranged between 2.4 per cent (for strain) and 31.5 per cent (for grade). The only exception for intra-rater reliability was audible respiration (5.0 per cent in the other direction).

A contradictory, although inconsistent, trend was seen for poorer reliability of observations when patients with functional laryngeal problems alone were considered (versus patients with no laryngeal problems; see Table III). In such cases, inter-rater reliability agreement decreased by between 9.1 and 25.3 per cent, with the exception of asthenicity (20.5 per cent in the other direction). The picture is less clear-cut for intra-rater reliability, with poorer results of between 8.8 per cent (for audible respiration) and 43.5 per cent (for strain), but with three GRBAS scale items moving in the direction of improved consistency (grade, 2.5 per cent; breathiness, 2.7 per cent; and asthenicity, 9.5 per cent). A statistically significant difference was only observed for intra-rater reliability in strain assessment.

Other results

Perceived age. For 29 of the 43 patients, there was majority agreement ($\geq$4/6 observations across raters on both occasions) of perceived age. This was in agreement with the patients' actual age in only 13 (44.8 per cent) patients. On the two rating occasions, raters' perception of age was in agreement with actual age within a range of 39.5 to 46.5 per cent.

Perceived gender. The patient's perceived gender was incorrect on only six occasions (2.3 per cent), out of a total of 258.

TABLE III

LARYNGOSCOPIC FINDINGS

Finding	Pts (*n*)
Structure	
Normal	25
Abnormal	18 (42%)
– Mild/mod/severe laryngitis	10/4/1
– Miscellaneous (not specified)	3
Function	
Normal	31
Abnormal	12 (28%)
– Glottic chink	5
– Phonating with false vocal folds	5
– Reduced vocal fold mobility	2

Pts = patients; mod = moderate

Perceived pitch range. There was full agreement among the three raters on two ratings in 15 patients (34.9 per cent). Figure 2 illustrates the majority perception of pitch range, with 27 of the patients (62.8 per cent) considered as normal. No significant statistical relationship could be demonstrated between the perceived pitch range and the grade item score or total voice symptom scale score ($p = 0.17$ and 0.52, respectively; Mann–Whitney U test). Raters' perceptions of pitch range did not appear to distinguish particularly well between the different subgroups identified at laryngoscopy, although eight of the nine patients with reduced pitch range had some form of laryngoscopic abnormality (see Table IV).

Perceived pitch height. There was full agreement among the three raters on two ratings in 13 patients (30.2 per cent). Figure 3 illustrates the majority perception of pitch height, with 25 of the patients (58.1 per cent) considered as falling within a 'medium' overall pitch height. There was no significant difference between the low and medium pitch height groups in terms of their grade item score and total voice symptom scale scores ($p = 0.11$ and $p = 1.0$, respectively; Mann–Whitney U test). Perception of pitch height did not distinguish between the different subgroups identified at laryngoscopy.

Discussion

The findings of this project reported elsewhere[35–38] confirm the results of other studies regarding the presence of organic and functional laryngeal abnormalities with or without associated vocal consequences in an asthmatic population.[28–34] It is against this background of observed pathology (Table III) that the results of the present study must be considered.

Some authors have been able to report very high levels of reliability for the GRBAS scale. For example, Murry and colleagues have described reliability coefficients in a voice-disordered population ranging from 0.88 for strain to 0.98 for grade, and reliability coefficients in a normal population of 0.99 for all GRBAS items; Piccirillo and colleagues have reported coefficient α reliability estimates ranging from 0.92 (for asthenicity) to 0.96 (for

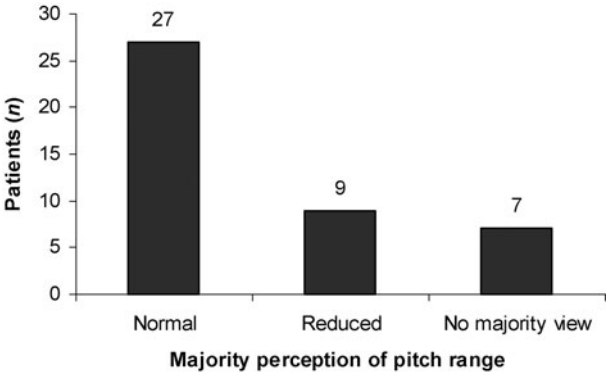

FIG. 2

Raters' perceptions of pitch range.

TABLE IV

MAJORITY PITCH RATINGS* COMPARED WITH LARYNGEAL ABNORMALITY FINDINGS

Pitch rating	Laryngeal abnormality		
	Structural	Functional	Any
Normal†	9	6	14
Reduced‡	7	4	8

*>3/6 agreements across raters and rating occasions. †$n = 27$; ‡$n = 9$.

grade and breathiness).[24,27] The reliability of the GRBAS scale in the present study has been shown to be fairly robust for total scores, both on an inter- and an intra-rater basis. However, total scores are not in common clinical use and may have little clinical relevance, although the present findings could be viewed as prima facie evidence for considering the use of total scores as a possible indicator in clinical practice. More commonly, grade is used as a measure of overall severity and has been generally reported as showing the best levels of agreement.[13,19,21,22] By comparison, in the current study the raters achieved a rather modest 64.7 per cent for inter-rater reliability and 69.6 per cent for intra-rater reliability for the grade item score. When one considers the asthenicity scale item, inter- and intra-reliability scores dropped as low as 43.4 and 49.6 per cent, respectively, a tendency that is repeated in other studies for the categories of both asthenicity and strain, usually, however, to the greater detriment of strain.[13,19,21,41] Moreover the reliability of the GRBAS scale remains open to question, as our findings also demonstrated a clear and consistent effect of the rater on the total score, the individual GRBAS scale item scores, and the instability and audible respiration scores, as well as a less consistent effect of rating occasion on total, asthenicity and strain scores (Table I). The latter effect also underlines the fragility of asthenicity and strain as reliable categories.

Although Dedivitis and colleagues do not report grade item scores, they observed, as did we, rather poor concordance rates for other GRBAS items in their patient population (smokers), which was in some respects similar to the current study population

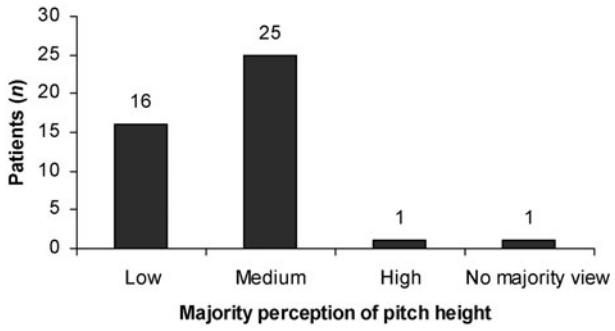

FIG. 3
Raters' perceptions of pitch height.

in being not specifically normal or voice-disordered but having clearly a significant proportion of vocal pathology. Furthermore, a tendency towards improved reliability in scoring, noted in the current study when subjects with laryngeal pathology were considered separately, was also found in Dedivitis' population of smokers.[44] This tendency also finds a partial resonance with Dejonckere and colleagues' finding that their raters tended to rate more severely for organic pathologies,[20] but is in contrast to the findings of Kreiman *et al.*, who reported improved agreement for listeners judging the voices of normal subjects.[14] Muñoz and colleagues did not use the GRBAS scale for perceptual ratings, and found that their raters were not able to distinguish unequivocally between recorded subjects with normal laryngeal status and those with identifiable pathology.[15] Such inconsistent findings across studies, possibly confounded by differing methodologies and different rating scales, underline some of the difficulties of achieving meaningful and reliable scores for perceptual evaluation of voice.

Despite the apparent attraction of an instability category to account for variability in perceived voice quality, this item too had a disappointing outcome, in line with the results reported by Dejonckere and colleagues.[21] It is therefore not surprising that this proposed addition has not generally been taken up in the literature. In the present study, the further addition of an audible respiration category (to account for the possible respiratory component in the perceived voice quality of a problem asthma clinic population, and scored zero to three as per the GRBAS scale) appeared to capture a feature recognised and agreed upon to a good degree by the raters (70.2 and 72.2 per cent overall inter- and intra-rater reliability, respectively). There is clearly a growing interest in the presence of vocal and laryngeal pathology in patients with respiratory disease, and this aspect of voice quality may need to be more closely defined and incorporated into any perceptual evaluation of this group, if not more generally.[28–34]

Various researchers and professional bodies have sought to establish the GRBAS scale as the best candidate for reliable perceptual evaluation of voice quality.[10,13,45,46] While there may be a case for using the GRBAS as the 'gold standard', this has not been accepted in the United States, where the American Speech-Language-Hearing Association has endorsed the use of the somewhat similar Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) scale. For all such scales, however, there must continue to be concern over the reported variability of rater reliability. It is of some significance that the present study revealed not only a consistent inter-rater effect (Table I) (with rater A tending to score higher on most scales) but also a considerable variability in intra-rater reliability across the raters (such that rater C achieved much greater consistency between rating occasions) (Figure 1). The latter finding in particular would suggest that some individuals may have better established internal listening standards. A similar phenomenon has been reported in two studies of the

application of cervical auscultation in determining the sounds of swallowing, with certain individuals proving much more consistent in their responses than their peers.[47,48] Leslie and colleagues reported intra-rater κ values ranging from −0.12 to 0.71, while Stroud *et al.* described one listener with 'almost perfect' intra-rater reliability. Such findings of superior consistency in listening performance lead one to speculate that close listening for perceptual rating purposes may be susceptible to the effects of training once standards have been established.[49] In the study by Karnell *et al.*, which employed both the GRBAS scale and the CAPE-V scale, the use of perceptual anchors in listening tasks may have contributed to the satisfactory correlation coefficients ($\geq$0.80) observed for both inter- and intra-rater reliability.[23] In a recent study, Eadie and Baylor, using perceptual categories derived from the CAPE-V scale, were able to demonstrate an effect of training in graduate speech language pathology students, in terms of improvements in intra-rater reliability (for overall severity of speech, $p = 0.015$) and in inter-rater reliability (more specifically for measures of breathiness, $p = 0.0167$).[50] Awan and Lawson investigated use of the CAPE-V scale and found that perceptual anchor modality may be critical in improving rater reliability, with a combination of textual and auditory anchors proving to be most effective.[51] These effects remain to be demonstrated for the GRBAS scale.

The remaining aspects of perceived voice considered in this study comprised age, gender, pitch range and pitch height. Of these categories, only gender reached good inter-rater agreement, an excellent 97.7 per cent, with one rater accounting for all the errors. Given the otherwise perfect agreement, this may reflect an element of rater fatigue or other form of inattention, factors which are not alluded to in other accounts of rater reliability in the voice literature.

- **Perceptual scales have been identified as an essential component in evaluation of the disordered voice**
- **The GRBAS scale has been recommended for perceptual evaluation of voice in voice-disordered populations**
- **The findings of this study do not support the routine use of this scale in characterising voice quality for clinical purposes**
- **Some scales for ad hoc perceptual analysis used in this study achieved superior rater agreement; further investigation of such scales may be indicated**
- **There are indications, supported by other published findings, that training in perceptual analysis of voice quality may improve rater reliability**

Perceived age, pitch range and pitch height reached poor levels of agreement among the raters. It is not surprising, therefore, that on the present evidence no correlation could be demonstrated between, on the one hand, the more specific, less well defined perceived categories of pitch range and height and, on the other, the GRBAS scale grade item and the total voice symptom scale score. There is some evidence in the literature to support the concept of a correlation between perceived attributes of disordered voice and the patient experience, but correlations of a similar nature in the present data could not be made.[27,52] Nevertheless, the process of voice perception remains subject to many poorly controlled factors, and the present negative results cannot be assumed to be the final word on the matter. Factors impinging on perception, and reported investigations thereof, may include listener experience, speaking task (e.g. sustained vowels versus connected speech) and experimental method adopted in the study (e.g. improving agreement by multiple presentation of the same stimuli).[16,25,53–59] Further work on these and other factors is required in order to be able to define the clinically relevant features of perceived voice, both normal and pathological, that can constitute a valid, reliable instrument for rating, triage and outcome purposes. Beyond this, however, the challenge may be to develop a robust, multi-dimensional (i.e. including perceptual, acoustic, laryngoscopic and self-reported data) evaluation to enable a truly complete characterisation of the voice, and to guide intervention for disordered voices.[8,9,11,12,16,60–62]

Acknowledgement

This work was supported by the Ritchie Trust Research Fellowship from the Royal College of Physicians and Surgeons of Glasgow.

References

1 Blood GW, Mahan BW, Hyman M. Judging personality and appearance from voice disorders. *J Commun Disord* 1979;**12**:63–7
2 DeGregorio NJ, Polow NG. Effect of teacher training sessions on listener perception of voice disorders. *Lang Speech Hear Serv Sch* 1985;**16**:25–8
3 Lallh AK, Rochet AP. The effect of information on listeners' attitudes toward speakers with voice or resonance disorders. *J Speech Lang Hear Res* 2000;**43**:782–95
4 Lass NJ, Ruscello DM, Bradshaw KH, Blankenship BL. Adolescents' perceptions of normal and voice-disordered children. *J Commun Disord* 1991;**24**:267–74
5 Ruscello DM, Lass NJ, Podbesek J. Listeners' perceptions of normal and voice-disordered children. *Folia Phoniatr* 1988;**40**:290–6
6 Davis CN, Harris TB. Teachers' ability to accurately identify disordered voices. *Lang Speech Hear Serv Sch* 1992;**23**:136–40
7 Bonet M, Casan P. Evaluation of dysphonia in a children's choir. *Folia Phoniatr Logop* 1994;**46**:27–34
8 Behrman A. Common practices of voice therapists in the evaluation of patients. *J Voice* 2005;**19**:454–69
9 Bhuta T, Patrick L, Garnett JD. Perceptual evaluation of voice quality and its correlation with acoustic measurements. *J Voice* 2004;**18**:299–304
10 Carding P, Carlson E, Epstein R, Mathieson L, Shewell C. Formal perceptual evaluation of voice quality in the United Kingdom. *Logoped Phoniatr Vocol* 2000;**25**:133–8

11 Eadie TL, Doyle PC. Classification of dysphonic voice: acoustic and auditory-perceptual measures. *J Voice* 2005; **19**:1–14
12 Speyer R, Wieneke GH, Dejonckere PH. Documentation of progress in voice therapy: perceptual, acoustic, and laryngostroboscopic findings pretherapy and posttherapy. *J Voice* 2004;**18**:325–40
13 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**:429–34
14 Kreiman J, Gerratt BR, Precoda K, Berke GS. Individual differences in voice quality perception. *J Speech Hear Res* 1992;**35**:512–20
15 Muñoz J, Mendoza E, Fresneda MD, Carballo G, Lopez P. Acoustic and perceptual indicators of normal and pathological voice. *Folia Phoniatr Logop* 2003;**55**:102–14
16 Kreiman J, Gerratt BR. Sources of listener disagreement in voice quality assessment. *J Acoust Soc Am* 2000;**108**: 1867–76
17 Hirano M. *Clinical Examination of Voice*, 1st edn. Berlin, Heidelberg, New York: Springer, 1981
18 De Bodt MS, Van de Heyning PH, Wuyts FL, Lambrechts L. The perceptual evaluation of voice disorders. *Acta Otorhinolaryngol Belg* 1996;**50**:283–91
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 Dejonckere PH, Obbens C, de Moor GM, Wieneke GH. Perceptual evaluation of dysphonia: reliability and relevance. *Folia Phoniatr Logop* 1993;**45**:76–83
21 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* 1996;**117**:219–24
22 Millet B, Dejonckere PH. What determines the differences in perceptual rating of dysphonia between experienced raters? *Folia Phoniatr Logop* 1998;**50**:305–10
23 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**:576–90
24 Piccirillo JF, Painter C, Haiduk A, Fuller D, Fredrickson JM. Assessment of two objective voice function indices. *Ann Otol Rhinol Laryngol* 1998;**107**:396–400
25 Lee M, Drinnan M, Carding P. The reliability and validity of patient self-rating of their own voice quality. *Clin Otolaryngol Allied Sci* 2005;**30**:357–61
26 Ma EP, Yiu EM. Multiparametric evaluation of dysphonic severity. *J Voice* 2006;**20**:380–90
27 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
28 Baker BM, Baker CD, Le HT. Vocal quality, articulation and audiological characteristics of children and young adults with diagnosed allergies. *Ann Otol Rhinol Laryngol* 1982;**91**:277–80
29 Barnes P. Corticosteroids. In: O'Byrne PM, Thomson NC, eds. *Manual of Asthma Management*. London: WB Saunders, 2000;173–96
30 Williamson IJ, Matusiewicz SP, Brown PH, Greening AP, Crompton GK. Frequency of voice problems and cough in patients using pressurized aerosol inhaled steroid preparations. *Eur Respir J* 1995;**8**:590–2
31 Crompton GK, Sanderson R, Dewar MH, Matusiewicz SP, Ning AC, Jamieson AH *et al.* Comparison of pulmicort pMDI plus nebuhaler and pulmicort turbuhaler in asthmatic patients with dysphonia. *Respir Med* 2000;**94**:448–53
32 Dogan M, Eryuksel E, Kocak I, Celikel T, Sehitoglu MA. Subjective and objective evaluation of voice quality in patients with asthma. *J Voice* 2007;**21**:224–30
33 Gallivan GJ, Gallivan KH, Gallivan HK. Inhaled corticosteroids: hazardous effects on voice – an update. *J Voice* 2007;**21**:101–11
34 Lavy JA, Wood G, Rubin JS, Harries M. Dysphonia associated with inhaled steroids. *J Voice* 2000;**14**:581–8
35 Stanton AE, Johnson MK, Carter R, MacKenzie K, Bucknall CE. Physiological evaluation of the upper airway in a problem asthma clinic. *Eur Respir J* 2004;**24**:P1712 http://www.ers-education.org/lr/abstract.aspx?idMedia=18512. Viewed 6/01/09
36 Stanton AE, MacKenzie K, Carter R, Bucknall CE. The spectrum of upper airway problems in a problem asthma clinic – the role of the larynx. *Eur Respir J* 2004;**24**:P1711 http://www.ers-education.org/lr/abstract.aspx?idMedia=18511. Viewed 6/01/09
37 Stanton AE, McGarry GW, Carter R, Bucknall CE. The spectrum of upper airway problems in a problem asthma clinic – the role of the nose. *Eur Respir J* 2004;**24**:1710 http://www.ers-education.org/lr/abstract.aspx?idMedia=18510. Viewed 6/01/09
38 Stanton AE, Sellars C, MacKenzie K, McConnachie A, Bucknall CE. Perceived vocal morbidity in a problem asthma clinic. *J Laryngol Otol* 2009;**123**:96–102
39 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 Allied Sci* 2004;**29**:169–74
40 Fairbanks G. *Voice and Articulation Drillbook*, 2nd edn. New York: Harper and Brothers, 1960
41 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* 1998;**119**:247–8
42 Gleser GC, Cronbach LJ, Rajaratnam N. Generalizability of scores influences by multiple sources of variance. *Psychometrika* 1965;**30**:395–418
43 Efron B, Tibshirani R. *An Introduction to the Bootstrap*. London: Chapman and Hall, 1993
44 Dedivitis RA, Barros AP, Queija DS, Alexandre JCM, Rezende WTM, Crazza VR *et al*. Interobserver perceptual analysis of smoker's voice. *Clin Otolaryngol Allied Sci* 2004;**29**:124–7
45 Carding P, Carlson E, Epstein R, Mathieson L, Shewell C. Re: evaluation of voice quality. *Int J Lang Commun Disord* 2001;**36**:127–34
46 Dejonckere PH. Perceptual and laboratory assessment of dysphonia. *Otolaryngol Clin North Am* 2000;**33**:731–50
47 Leslie P, Drinnan MJ, Finn P, Ford GA, Wilson JA. Reliability and validity of cervical auscultation: a controlled comparison using videofluoroscopy. *Dysphagia* 2004;**19**:231–40
48 Stroud AE, Lawrie BW, Wiles CM. Inter- and intrarater reliability of cervical auscultation to detect aspiration in patients with dysphagia. *Clin Rehabil* 2002; **16**:640–5
49 Wolfe VI, Martin DP, Palmer CI. Perception of dysphonic voice quality by naive listeners. *J Speech Lang Hear Res* 2000;**43**:697–705
50 Eadie TL, Baylor CR. The effect of perceptual training on inexperienced listeners' judgments of dysphonic voice. *J Voice* 2006;**20**:527–44
51 Awan SN, Lawson LL. The effect of anchor modality on the reliability of vocal severity ratings. *J Voice* 2008 http://www.sciencedirect.com/science/journal/08921997. Viewed 16/10/08
52 Pribuisiene R, Uloza V, Kupcinskas L, Jonaitis L. Perceptual and acoustic characteristics of voice changes in reflux laryngitis patients. *J Voice* 2006;**20**:128–36
53 Bele IV. Reliability in perceptual analysis of voice quality. *J Voice* 2005;**19**:555–73
54 Chan KMK, Yiu EM. A comparison of two perceptual voice evaluation training programs for naive listeners. *J Voice* 2006;**20**:229–41
55 Kreiman J, Gerratt B. Measuring vocal quality. In: Kent RD, Ball MJ, eds. *Voice Quality Measurement*, 1st edn. San Diego: Singular, 2000:73–101
56 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**:323–35
57 Shrivastav R. The use of an auditory model in predicting perceptual ratings of breathy voice quality. *J Voice* 2003; **17**:502–12

58 Shrivastav R. Multidimensional scaling of breathy voice quality: individual differences in perception. *J Voice* 2006; **20**:211–22
59 Zraick RI, Wendel K, Smith-Olinde L. The effect of speaking task on perceptual judgment of the severity of dysphonic voice. *J Voice* 2005;**19**:574–81
60 Behrman A, Sulica L, He T. Factors predicting patient perception of dysphonia caused by benign vocal fold lesions. *Laryngoscope* 2004;**114**:1693–700
61 Behrman A. Evidence-based treatment of paralytic dysphonia: making sense of outcomes and efficacy data. *Otolaryngol Clin North Am* 2004;**37**:75–104
62 Deary IJ, Wilson JA, Carding PN, Mackenzie K. The dysphonic voice heard by me, you and it: differential associations with personality and psychological distress. *Clin Otolaryngol Allied Sci* 2003;**28**:374–8

Address for correspondence:
Mr Cameron Sellars,
Dept of Speech and Language Therapy,
Glasgow Royal Infirmary,
Castle St,
Glasgow G4 0SF, Scotland, UK.

Fax: +44 (0)141 211 4821
E-mail: Cameron.Sellars@ggc.scot.nhs.uk

Mr C Sellars takes responsibility for the integrity of the content of the paper.
Competing interests: None declared