1.1 Introduction

Clinicians working with voice disorder take a multidimensional approach to voice evaluation. In adults this includes as a minimum: a full case history, laryngeal evaluation, aerodynamic evaluation of respiration and phonatory skill, acoustic and perceptual evaluation of voice quality and an understanding of the subjective impact of voice on quality of life (1). The same approach is recommended in the paediatric setting (2, 3). The extent to which voice clinicians manage to routinely gather all of this data varies depending on resources (such as access to specialised equipment) and the availability of multidisciplinary voice clinics – those that are staffed by both otolaryngologists and Speech and Language Therapists (SLTs).

Perceptual evaluation of voice quality is a clinical subjective skill, and clinicians rely on published guidelines and protocols to help them judge a presenting patient's voice quality. Commonly used protocols include the GRBAS (4) and the CAPE-V (5). The main difference between these two relates to the scaling used. While the GRBAS uses a 4 point ordinal scale and the CAPE-V a visual analogue scale, both provide evaluation of overall severity (CAPE-V) or grade (GRBAS), hoarseness (CAPE-V) or roughness (GRBAS), breathiness and strain (both CAPE-V and GRBAS). The CAPE-V additionally rates pitch and loudness while the GRBAS asthenia (vocal weakness). The CAPE-V provides standard stimuli. Different rating scale may impact on perceptual evaluation – for example, the GRBAS is reported to be faster to complete but with its small range of ordinal options it is less sensitive to detecting voice change than the CAPE-V (6, 7). In studies of adults with voice disorder there is reported to be a strong correlation between CAPE-V and GRBAS ratings suggesting that clinicians detect similar levels of dysphonia regardless of scale used (8). The GRBAS is recommended as the minimum scale for use in the UK (9) and since then has been more widely adopted into routine clinical practice. While there are no set stimuli for the GRBAS clinicians can use spontaneous speech, sustained vowel production and reading aloud and the stimuli from the CAPE-V offer a useful standard set of stimuli that can be rated using either the CAPE-V or the GRBAS

Previous studies have explored the intrarater (within judge) and interrater (across judges) reliability of perceptual analysis using one or both of these protocols. There is a high level of both intra and interrater reliability reported in adults with dysphonia using both the GRBAS and the CAPE-V (7, 8, 10) though some report that the CAPE-V shows a higher degree of reliability (10). While both Karnell et al (8) and Zraick et al (10) utilised retrospective clinical data including audio recordings of data of 34 and 74 (respectively) adult patients following the CAPE-V protocol, Nemr et al (7) prospectively gathered audio data from 60 patients, including 10 younger patients. Expert judges rated the audio data in each study, (four judges in Karnell et al (8), with one judge who was the original rating clinician at the time of the audio recordings being made, three judges in Nemr et al (7) and twenty-one in Zraick et al (10)). Repeat ratings of all the data was carried out in one study (8) and of a small sample of data in the other two studies. Zraick et al (10) noted that reliability was slightly higher for the CAPE-V than the GRBAS.

Various factors that might affect inter and intra-rater reliability are explored by Kreiman and Gerratt (11) who point out the importance of listener agreement in the clinical setting. A range of methods have been used to manage this such as level of training of listeners, what type of data is used, providing listeners with comparison anchor stimuli and the severity of voice disorder in the sample. Training that directs the listener to specific vocal characteristics and encourages groups of listeners to discuss and agree on ratings, regardless of rating system, improves reliability both within and across judges (12-14). Where listeners are able to access comparison stimuli there is less reliance on each listener's own internal subjective benchmarks and these, often referred to as anchor stimuli, increase interrater reliability (15). Some aspects of voice quality such as breathiness show a greater degree of interrater reliability over others (16). Type of stimuli may also effect reliability with conversational speech, counting and sustained vowels showing greater reliability than a standard reading passage (17, 18)

Furthermore, some retrospective studies have shown that perceptual evaluation of voice quality relates to known voice diagnosis (19, 20). In a large scale (n=254) study of vocal fold nodules in children (19), hoarseness, breathiness, strain and aphonia was evaluated using a 3 point scale, where 0 = no hoarseness and 3 = severe hoarseness. The authors explored the relationship between size of nodules and perceptual characteristics, with larger nodules correlated with increased hoarseness, breathiness, strain and aphonia. A similar pattern of correlation between nodule size and perceptual rating was found by Nuss et al (20) who used the CAPE-V protocol in their retrospective study. Neither study reports whether or not perceptual ratings were made 'blinded' to voice diagnosis and/or nodule size. This is important as prior knowledge of diagnosis might affect the subjective perceptual evaluation. For example, a clinician might be inadvertently biased by size of nodules in their perceptual ratings where this information is known to them.

Voice quality is problematic in other aetiologies. Children with airway problems present a unique challenge to the clinicians managing their care, for example the subgroup of children requiring surgical intervention (laryngotracheal reconstruction surgery- LTR) to ensure a safe airway following subglottic stenosis (SGS). Longer term outcomes for this relatively small population are improving in relation to general health related quality of life (21) and also in relation to voice outcome (22), though there is conflicting information in the published literature. Some studies report good outcomes (23) and others poor outcomes (21, 24). Good voice outcome seems to relate to the intactness of the laryngeal structures and use of a glottic phonation source (22). Some of the challenges in drawing together the literature relate to what measures are taken and how 'good' voice is defined, along with whether or not studies are gathered from retrospective (21, 24) or prospectively (22, 23) gathered data. Some studies explore voice related quality of life and health related quality of life along with laryngoscopy and perceptual evaluation of voice quality (21) while others also evaluated acoustic aspects of voice quality (22-24).

Reliability of perceptual rating of voice quality in children who have had LTR has been explored using the CAPE-V protocol (25). In this study of 50 children with LTR aged 4-20 yrs (32 female, 18 male), three experienced SLP judges, none of whom were involved in the management of the participants, rated recordings of the participants either reading aloud or repeating each of the CAPE-V sentences. Judges rated the samples following the CAPE-V protocol and additionally rated 34% of the sample a second time one week after their initial rating. This approach allowed the authors to measure interrater and intrarater reliability. They found high levels of interrater reliability in overall severity, roughness and breathiness and high levels of intrarater reliability in overall severity, roughness, breathiness, pitch and loudness with a lower reliability for judgements of strain. The authors outline that this latter parameter involves judging vocal effort through a combination of auditory evaluation and visual observation of muscular tension in the neck and vocal tract, and this might explain the lower reliability of ratings based on audio data alone in this domain. While this study shows that there is reliability using the CAPE-V protocol, with the GRBAS in more common use in the UK, it would be useful to explore the extent to which there is interrater and intrarater reliability of this particular clinician-rating tool for children with a history of LTR in the UK.

Clinician reporting of voice quality has also been compared with patient reports of impact of voice on quality of life. Clinically, this information is valuable so that intervention can be tailored to the individual and gathering patient opinion is considered an essential part of voice evaluation (1). Most studies that have compared clinician and patient report used a retrospective case study design. Different tools have been used to evaluate impact of voice on quality of life in different populations (e.g. adult and parent proxy questionnaires, different questionnaires suited to specific aetiologies) including the Voice Related Quality of Life (VRQOL) (26) questionnaire and the Iowa Patient's Voice Index (IPVI) (8), the Voice Handicap Index (VHI) (27) and the Pediatric Voice Handicap Index (PVHI) (28-30). In adults, Karnell et al (8) report a weak correlation between patient report (IPVI and VRQOL) and clinician report (GRBAS and CAPE-V) in their study of 103 patients with a range of different aetiologies. In children with vocal fold lesions, a fair correlation exists between CAPE-V

overall severity and PVHI (29) while this correlation is reported as weak in children who have a history of airway surgery (28).

There can be differences between how parents and children report subjective impact of voice problem using an adaptation of the Paediatric Voice Related Quality of Life (PVRQOL) questionnaire (31). From a clinical perspective, there is consensus that clinician and patient reporting both give unique insights. The voice therapist will take account of both of these in order to tailor a holistic intervention programme for the patient. In the paediatric context the clinician should consider any difference of opinion between the parent and child. Where for example the child is unconcerned about the impact of voice on quality of life, they may not be particularly motivated to make any behavioural changes that could improve voice quality and indirect intervention with a more concerned parent might be of more advantage at that time.

1.2 Aims

The aims of this study were to examine clinician and patient ratings of voice in children with a history of LTR following SGS. This includes examining the interrater and intrarater reliability of clinician rated voice quality using GRBAS; examining the agreement between GRBAS and CAPE-V rating scales and examining the relationship between parent proxy and child self-report of subjective impact with the clinician GRBAS rating of voice quality

2.1 Methods

2.2 Ethical permissions

Permission for the study was granted by the National Health Service West of Scotland Research Ethics Committee and the University of Strathclyde Research Ethics Committee. Information sheets and consent forms were designed so they were appropriate for younger children using pictures to describe the study, with written consent obtained from both the child and their parent/guardian.

2.3 Study Design

A prospective observational study design was used to analyse clinician ratings of voice quality using GRBAS and CAPE-V and to compare the GRBAS ratings to parent proxy and child self-report of voice related quality of life gathered at the same point in time. The data analysed was gathered in a previously reported research study and detailed information about how participants were recruited to that research study is available elsewhere along with information about acoustic analysis of voice quality and a description of laryngeal function for each child (22).

2.4 Participants

Eleven participants (4 female, 7 male) with a history of SGS and LTR attended for a range of voice evaluation measurements providing the voice recordings used in this study. Five had LTR at <12months with the other six before age 3. They were all aged between 5-14 years ($\bar{x}$ 8.5; SD 3.4) at the time of the study. Table 1 summarises the medical and surgical histories, including age at time(s) of surgery and at time of participation in the study along with the data gathered as part of the analysis presented in this paper. Four trained listeners (two final year speech and language therapy students and two recently qualified speech and language therapists) provided the clinician ratings of voice quality. Rating was also completed by an expert listener, a specialist speech and language therapist known to each of the patient participants.

2.5 Data

Audio recordings followed the CAPE-V protocol which included the sustained vowel sound [a], six sentences and a sample of conversational speech. Recordings were made using a Tascam DR-05 Version 2 Dictaphone Linear PCM Portable Recorder in a sound-treated room tested for ambient noise reduction to 14dB. The inbuilt stereo condenser microphone was used with a mouth-to-microphone distance of 21cm. Raw recordings were edited using Audacity 1.2 (Audacity Team, audacity.sourceforge.net) and exported as mp3 files (codec: mpga; channels: mono; sample rate: 16 kHz; bit rate: 128 kB/s.).

At the time of audio recordings, each participant and their parent/guardian completed the PVRQOL(30) questionnaire giving an indication of the extent to which voice or problems with voice currently impact on quality of life (30). A high score indicates high voice related quality of life and low impact of voice problems on daily living. Two of the youngest participants were not able to complete the child self-report version of the PVRQOL(31). Scores were calculated following the scoring guidelines giving a total score (PVRQOL T PARENT, PVRQOL T CHILD), physical functioning score (PVRQOL PF PARENT, PVRQOL PF CHILD) and socio-emotional score (PVRQOL SE PARENT, PVRQOL SE CHILD) for each participant.

M/F	**Age(s) at surgery**	**Medical and surgical history**	**Age at time of study (years)**	**Audio data**	**Parent proxy PVRQOL**	**Child self-report PVRQOL**
F	Neonate	Born at term. 22q11DS, cardiac surgery and LTR during neonatal period. Continues to attend for cardiology.	5	✓	✓	-
F	12 months	Pre-term 33 weeks. Ventilated in neonatal period. SGS. LTR at 1 year. Chronic reflux treated medically	6	✓	✓	✓
F	Neonate – 1 year	Pre-term 27 weeks, cardiac surgery, global developmental delay, SGS. Tracheostomy as a neonate, decannulated and LTR by 12 months. Continues to attend for cardiology.	6	✓	✓	-
F	2-3 years	Pre-term 25 weeks, SGS, neonatal cardiac surgery now discharged from cardiology. Tracheostomy followed by LTR and decannulation	12	✓	✓	✓
M	6 months	Pre-term 28 weeks. Ventilated during neonatal period. SGS. LTR.	6	✓	✓	✓
M	Neonate – 2 years	Pre-term 24 weeks. SGS. Neonatal tracheostomy. LTR at age 2 with a revision within one month. Tracheostomy in situ at time of study.	6	✓	✓	✓
M	Neonate – 2 years	Born at term. Pierre Robin sequence, cleft palate. Tracheostomy during neonatal period. LTR and decannulated at 2 years.	6	✓	✓	✓
M	3 years	Born at term. Laryngeal cleft and tracheo-oesophageal fissure. LTR.	9	✓	✓	✓
M	1 year	Born at term. Primary laryngeal atresia with glottic web and SGS. LTR.	11	✓	✓	✓
M	1 year	Pre-term 32 weeks. SGS. LTR.	13	✓	✓	✓
M	Neonate – 5 years	SGS. Tracheostomy in neonatal period. Decannulated and CTR at 2 years; residual stenosis LTR at 5 years. Atrophy of R vocal fold.	14	✓	✓	✓

Table 1 – Biographical, surgical and data gathered from the study participants.

2.6 Listener consensus training

Listeners attended a full day (7 hours) training session. During this training session the listeners, who were already familiar with the GRBAS protocol as part of their speech and language therapy studies

revised their understanding of the GRBAS protocol. This included a discussion of what each person understood was meant by the various GRBAS parameters along with a discussion about the severity ratings (normal = 0, mild = 1, moderate = 2, severe = 3). This was followed up with practice using a consensus rated CD that is currently used for classroom teaching. Discussion of each sample led to shared understanding of what normal, mild, moderate and severe means for each of the GRBAS parameters. This was repeated with discussion of the CAPE-V protocol and how the various parameters are rated on the visual analogue scale. Further practice was made using audio recordings taken from five children with voice disorders who were not involved in the study. These recordings consisted of the same stimuli used in the study (sustained vowel, 6 sentences and spontaneous conversation). Each listener individually rated each sample using GRBAS followed by group discussion leading to a consensus GRBAS rating for the sample before moving onto the next. This was repeated for all samples. From these, three anchor files were selected representing mild, moderate and severe voice quality, each with a consensus GRBAS rating. The three anchor files, with the consensus GRBAS rating were later accessed by the listeners prior to rating the experimental data. The same method was repeated with the samples using the CAPE-V protocol, ending up with three anchor files representing mild, moderate and severe voice quality each with a consensus CAPE-V rating.

2.7 Measurement / Statistical methods

Two dataset were created from the voice samples. The first dataset contained single files of one stimulus per child: the sustained vowel sound [a], the six sentences and a short sample of spontaneous speech, giving 8 files per participant. This dataset was used to examine interrater and intrarater reliability of the GRBAS ratings. The second dataset consisted of files containing one complete stimuli sample per child and was used to examine agreement between GRBAS and CAPE-V ratings.

Each dataset was duplicated and randomly ordered using a number randomisation generator (32). Dataset one contained 176 experimental files (11 x 8 x 2) for GRBAS rating and dataset two contained 22 experimental files (11 x 2) for CAPE-V and GRBAS overall rating.

Listeners may have been alerted to the duplication when rating the spontaneous speech samples, and this is a limitation of the methodology, however in an attempt to counter this, listeners were given the instruction: *"Each sample will be presented twice in a random order. If you recognise a sample, try to just rate it as if you were hearing it for the first time."* Listeners were instructed to play back the three anchor files from their listener training session, asked to listen to them and consider the provided consensus ratings from the training session before rating the experimental files. If listeners wished to take a rest break, they were able to re-access the anchor files before continuing to rate the experimental files. They were able to listen to each recording as many times as they wished and asked to note how many times they listened to each before settling on their final ratings.

Intrarater reliability and interrater reliability were tested using the Intraclass Correlation Coefficient (ICC) using SPSS statistical analysis software following Shrout and Fleiss (33). The expert listener rated the dataset following the same procedure so that comparison could be made with that of the trained listeners.

For the comparison of clinician rating of voice quality and parent/child reporting of subjective impact of voice quality, Spearman's Correlation Coefficient (using SPSS statistical analysis software) was used to correlate the median Grade rating from the four listeners from each audio file with the three PVRQOL PARENT and three PVRQOL CHILD scores. To allow for limitations of sample size and to reduce type 1 errors, a Bonferroni correction was used to determine the significance level (which was calculated as $p < 0.00625$).

3.1 Results

The findings are reported in the following sections organised in relation to the study aims.

3.1.1 Intrarater and interrater reliability of the GRBAS

Intrarater reliability of the expert listener was high. Table 2 below summarises the findings of the intra- and interrater analysis. ICC was significant (p<0.00625) showing reliability of GRBAS rating within and across all four trained listeners. Intrarater ratings for Grade and Roughness were higher than for Breathiness, Asthenia and Strain. Interrater ratings for Grade, Roughness and Asthenia were higher than for Breathiness and Strain. Intrarater reliability of the expert listener was significant (p<0.00625) and there was significant ICC between the median ratings from the trained listeners when compared to the expert listener.

Intrarater comparison of trained listeners					**Interrater comparison of trained listeners**			
	ICC	**95% CI**	***p value***			**ICC**	**95% CI**	***p value***
G	0.920	0.902, 0.936	0.000		**G**	0.945	0.923, 0.961	0.000
R	0.915	0.896, 0.931	0.000		**R**	0.942	0.920, 0.960	0.000
B	0.796	0.748, 0.835	0.000		**B**	0.802	0.724, 0.862	0.000
A	0.888	0.862, 0.909	0.000		**A**	0.911	0.877, 0.938	0.000
S	0.872	0.842, 0.896	0.000		**S**	0.868	0.809, 0.910	0.000
Intrarater comparison of expert listener					**Interrater comparison – Median ratings of trained listeners compared to expert listener**			
	ICC	**95% CI**	***p value***			**ICC**	**95% CI**	***p value***
G	0.907	0.858, 0.939	0.000		**G**	0.840	0.801, 0.914	0.000
R	0.833	0.745, 0.890	0.000		**R**	0.839	0.737, 0.899	0.000
B	0.779	0.663, 0.855	0.000		**B**	0.718	0.570, 0.815	0.000
A	0.896	0.842, 0.932	0.000		**A**	0.840	0.656, 0.915	0.000
S	0.939	0.908, 0.960	0.000		**S**	0.764	0.560, 0.863	0.000

Table 2. Results of the intrarater and interrater comparisons. ICC of average measures is shown, along with the 95% confidence interval and the significance level.

3.1.2 Agreement between GRBAS and CAPE-V

Figure 1 below shows the level of agreement between the median GRBAS ratings and their corresponding mean CAPE-V ratings (Grade / Overall Severity, Roughness, Breathiness and Strain) from the four trained listeners. In this small dataset there is a high level of agreement across these four parameters.

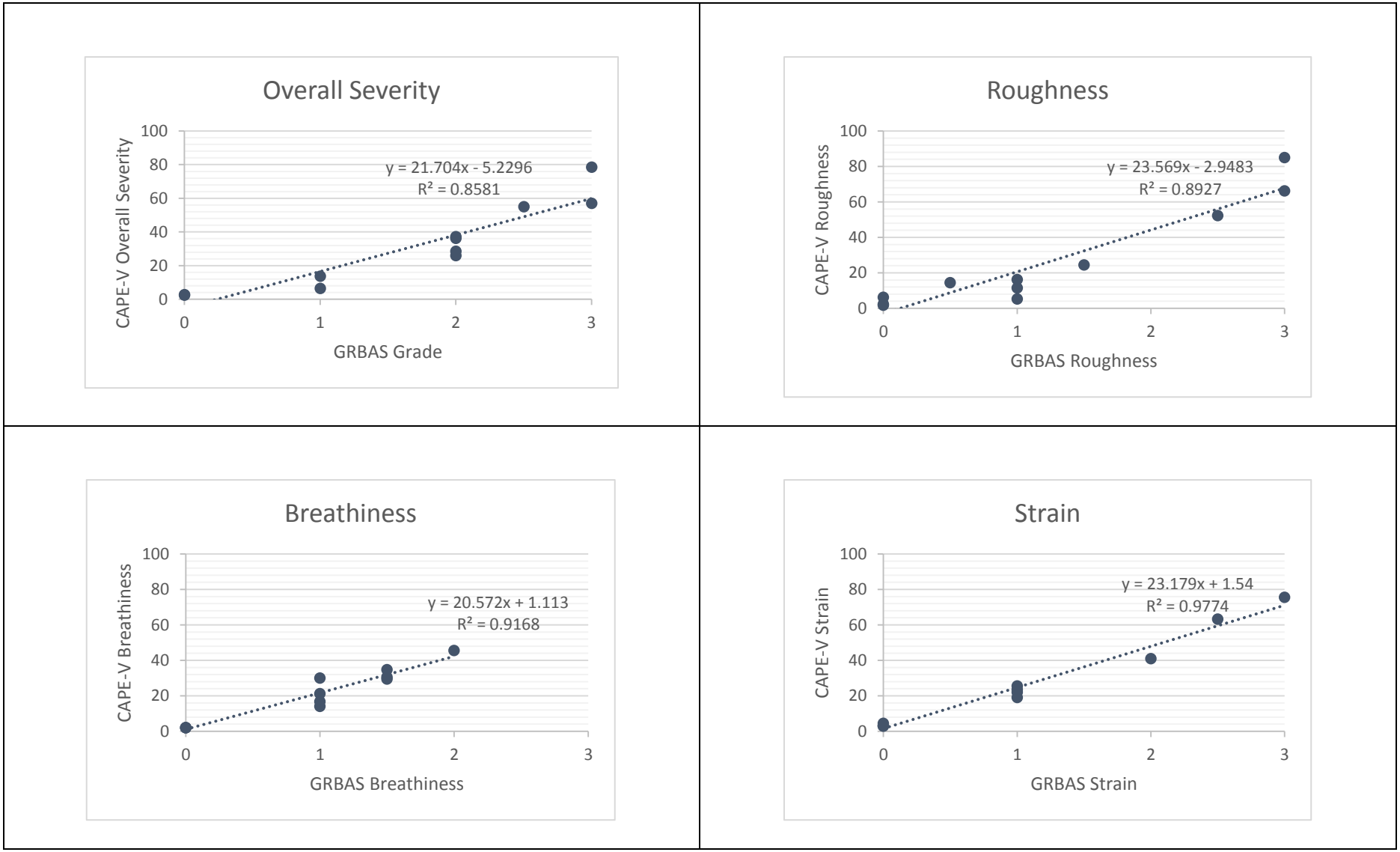

Figure 1. Agreement between GRBAS and CAPE-V ratings

3.1.3 Correlation between parent proxy and child self-report of PVRQOL and clinician GRBAS rating

A high score in the total (T), physical functioning (PF) or social emotional (SE) domains of the PVRQOL indicates high voice related quality of life (corresponding to a low impact of voice difficulties on quality of life). When scoring the parent proxy PVRQOL questionnaire, two parents had rated some questions as "not applicable". This led to missing values from the total scores as "not applicable" has no value attached to it. A decision was taken to consider these values commensurate with something being "not a problem" so that the T, PF and SE scores were not artificially lowered due to there being no value attached those responses. The scores are reported in table 3 below.

Parent Proxy			**Child Self-report**		
T	**PF**	**SE**	**T**	**PF**	**SE**
85	47.5	40.0	-	-	-
47.5	24.0	22.5	92.5	55.0	37.5
75	35.0	40.0	-	-	-
72.5	45.0	27.5	90	55.0	40.0
60	27.5	32.5	62.5	25.0	37.5
75	37.5	37.5	87.5	47.5	40.0
90	50.0	40.0	97.5	57.5	40.0
100	60.0	40.0	100	60.0	40.0
82.5	50.0	32.5	97.5	57.5	40.0
100	60.0	40.0	87.5	50.0	37.5
62.5	35.0	27.5	92.5	52.5	40.0

Table 3. Parent proxy and child self-report PVRQOL total (T), physical functioning (PF) and social emotional (SE) domains

Given that there was high interrater reliability of the perceptual judgements as outlined in section 3.1.1 above, correlation analysis of the median value from the four judges' Grade ratings was made with the parent proxy and child self-report PVRQOL data. Any correlation would be expected to be negative as a higher PVRQOL score means no impact of voice on quality of life (e.g. higher = 'more normal') while a higher clinician perceptual score relates to increased severity of perceived voice quality (e.g. higher = 'more abnormal').

Spearman's correlation coefficient revealed that there was a significant negative correlation (where significance is defined as $p<0.00625$) between *some* parent proxy PVRQOL scores and clinician perceptual rating of voice quality. Parent proxy total and physical functioning scores were significantly negatively correlated with clinician perceptual rating of the sentence "*the blue spot is on the key again*" ($p<0.005$ and $p<0.001$ respectively) and "*conversation*" ($p<0.003$) and $p<0.001$ respectively). Parent proxy physical functioning scores were also significantly negatively correlated with clinician perceptual rating of the sentences "*we were away a year ago*" ($p<0003$) and "*peter will keep at the peak* ($p<0006$). There was no correlation between child self-report and clinician perceptual rating.

4.1 Discussion

One of the aims of this study was to find out the reliability of clinician perceptual evaluation of voice in a cohort of children with a history of LTR. Previous study of a similar cohort found high interrater and intrarater reliability across three experienced judges using the CAPE-V (25).

Assumptions could be made on the basis of studies comparing the CAPE-V and the GRBAS scales in other voice disordered populations (7, 8, 10) that there would also be reliability using the GRBAS scale in children with a history of LTR following SGS. Our study provides data that supports this assertion. This validation is valuable for clinicians in the UK who tend to be more familiar with GRBAS that CAPE-V.

In our study we found a high degree of interrater and intrarater reliability across four consensus trained judges using the GRBAS. While our judges were in either the final stage or had recently completed their training, they were not as experienced as those in previous studies who had extensive experience in the assessment and treatment of voice disorders in both adults (8) and children (25). However, the level of training they had experienced appeared to be sufficient to

ensure a level of reliability of perceptual evaluation. It was encouraging to note that there was also a high degree of inter-rater reliability between the trained judges and the expert rater, though of course the expert rater was not blinded to the patient participants. A combination of factors in our listeners contributed to the high intra and interrater reliability. They had completed the same amount of training in voice theory; they participated as a group in the consensus listener training (12, 13) which led to them developing and agreeing the anchor stimuli, an approach which is also known to improve rater reliability (12-14). It is commonly that prospectively designed research studies involving subjective rater evaluation should ensure that raters are blinded to any diagnostic information that could affect their subjective rating. Perhaps the evidence from this small-scale study could open up some discussion of this given that there was a strong relationship between the expert listener's unblinded ratings and those of the trained listeners.

In our data, while the inter and intrarater reliability of Breathiness was significant, it had a slightly lower ICC than the other parameters, and this finding is at odds with that reported by Brinca et al (16) who suggest that Breathiness might be the easiest voice quality to evaluate. We also found that our listeners were more able to reliably rate Strain compared to those reported by Kelchner et al (25). While our listeners were able to reliability rate Asthenia (both within and across judges) this is contrary to observations in other studies of adult voice where inconsistency in the reliability of rating this parameter has remained even after anchor stimuli were provided (13). These may be facets of the presenting cases where Breathiness, Asthenia and Strain could be impacted by the otolaryngology status of the participants.

One of the biggest challenges in drawing comparisons between our data and previous studies relates to methodological differences amongst them. Differences include: which rating tool (e.g. aspects of GRBAS or using CAPE-V); what type of stimuli (e.g. sustained vowels, conversation and/or reading sentences or paragraphs); duration of listener training (one week before as in our study or over a longer period of time as outlined by Brinca at al (16)); and instructions over use of anchor stimuli

(use once only (16) or as frequently as required to help set and re-set an internal standard as in our study). We would propose that different voice conditions cue listeners to the discrete aspects of voice quality that relate to those specific aetiologies. For example there may be discrete differences in how Asthenia is perceived in children's versus adults' voices given that Asthenia relates to perceived weakness in a voice, and particularly in our specific population of children with a history of LTR.

A methodological strength in our subject design was using a standard set of stimuli for our participants. We selected the CAPE-V stimuli as it was relatively straightforward to prepare visual materials for our younger participants. This helped them to repeat sentences where they were unable to read the written form. While the CAPE-V is known to provide a more sensitive measurement of voice quality over the GRBAS, the GRBAS is quicker to carry out in the clinical setting and may be used more frequently for that reason (7, 10) so it is encouraging to find a high level of agreement in our sample between the two scales. We would recommend using these stimuli in the paediatric voice setting regardless of which rating system is being used.

Moreover the findings of our study, taken with previous research (8, 25), adds to the knowledge base about the reliability of clinician rating of voice disorders for any age of patient, child or adult as long as that clinician has either extensive experience in rating voice or has engaged in consensus training and makes use of agreed anchor data. We recommend that any service providing voice evaluation ensures there are suitable opportunities for clinicians to agree ratings across a range of voices to ensure ongoing reliability of this subjective clinical tool. Clinicians can choose to use either the GRBAS or CAPE-V with increased confidence in the reliability of both tools. This confidence could be enhanced with development of consensus materials in different languages and dialects that could be used in training the next generation of clinicians along with as continuing professional development opportunities for more experienced clinicians. Future work is now being planned to explore developing this type of consensus training material for the paediatric population in the UK.

In the paediatric context we currently advocate asking both the parent and the child, where the child is able to provide this, as there can be differences between what matters to parents and what matters to children when it comes to voice management (31). This is important in terms of how to approach management of a paediatric voice disorder – and whether the focus of any intervention should be on the parent or on the child themselves in the sphere of awareness raising. In our study, 9 of the 11 children were able to respond to our adapted child version of the PVRQOL. While there was a significant correlation between parent proxy report and some elements of clinician perceptual rating this was not the case with the child self-report. The lack of correlation across all parameters implies that there is continued strength in using a multidimensional voice evaluation approach that should ideally incorporate perceptual evaluation, subjective impact and also acoustic evaluation of voice quality and laryngoscopic evaluation. For the children presented in this paper, acoustic variables reported previously (consisting of F0, Jitter, Shimmer and Noise-Harmonic Ratio)(22) for four children were found to be within the normal range on all acoustic variables and in two other children within the normal range for three acoustic variables. One of these cases was perceived as having normal voice (GRBAS Grade = 0) with the others identified as mildly dysphonic (GRBAS Grade = 1). Two of the cases who were mildly dysphonic presented with vocal fold nodules. Both of these children presented with recent case histories that would suggest that these nodules were unrelated to the LTR.

This multidimensional approach means that any decisions taken in relation to intervention are made with a holistic understanding of the individual concerned. The lack of any significant correlation of child self-report with clinician rating in our sample might suggest that there is no added value to asking children for their views if the purpose of voice evaluation is simply to ascertain severity of voice disorder. However, the rationale of a multidimensional approach to voice evaluation takes into consideration severity, parental and child view of impact to assist in intervention planning. Given the small size of the cohort in this study there may be discrete individual differences in how children report subjective impact, and without a larger sample size to explore these relationships

these effects are unknown. In our sample it wasn't necessarily the case that those children with more severe GRBAS ratings also had lower voice related quality of life. Taken together with the other acoustic and laryngocsopic evidence then advice and management can be tailored to the individual – and in the cases presented here this included information about voice care for vocal fold nodules despite there be minimal concern noted of voice quality for these two children.

4.2 Implications for clinical practice

This study shows that there is intra and interjudge reliability in using the GRBAS in this clinical population. Reliability might be increased through training and the use of consensus agreed anchor stimuli and further research may be conducted to evaluate that specifically, however for those clinicians working with this population then ongoing development of perceptual analysis skills should be encouraged. That there was no correlation between child report and clinician rating is important from an intervention perspective. Clinicians need to be aware of this when planning management to ensure that any direct intervention approaches take into account the voice related impact on children.

4.3 Limitations

This was a small scale study of a unique clinical population and the duplication of conversation data can affect the true intrarater reliability given the potential recognition of conversational elements however there is no way to counter this unless conversational data is not rated. Replication and extension with a wider range of paediatric voice pathologies would be helpful. Future research exploring the extent to which listener training affects reliability would be beneficial.

4.3 Conclusions

There is a significantly high degree of intra- and interrater reliability using the GRBAS as part of a multidimensional evaluation of voice quality in children. The use of standard stimuli is to be recommended and these may well include those outlined in the CAPE-V.

There is a partial relationship between parent proxy of subjective impact of quality of life and clinician rating in this specific population. The lack of correlation across all parameters implies that there is continued strength in using a multidimensional voice evaluation approach. This approach assists the clinician in devising a holistic management plan suited to the individual.

Acknowledgements

This project was supported by generous grants from Action Medical Research, The Hugh Fraser Foundation and Jeffrey Charitable Trust. Thanks are due to the twelve participants and their families for giving their time generously to this study. Specific recognition is also due to the four listeners, now each furthering their careers as speech and language therapists. Their assistance in the study was invaluable.

References

1. Dejonckere PH, Bradley P, Clemente P, Cornut G, Crevier-Buchman L, Friedrich G, et al. A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques. Guideline elaborated by the Committee on Phoniatrics of the European Laryngological Society (ELS). Eur Arch Otorhinolaryngol. 2001;258(2):77-82.
2. Cohen W, Wynne DM, Kubba H, McCartney E. Development of a minimum protocol for assessment in the paediatric voice clinic. Part 1: evaluating vocal function. Logopedics, phoniatrics, vocology. 2012;37(1):33-8.
3. Cohen W, Wardrop A, Wynne DM, Kubba H, McCartney E. Development of a minimum protocol for assessment in the paediatric voice clinic. Part 2: subjective measurement of symptoms of voice disorder. Logopedics, phoniatrics, vocology. 2012;37(1):39-44.
4. Hirano M. Clinical examination of voice. Wien; New York: Springer-Verlag; 1981.
5. Kempster GB, Gerratt BR, Verdolini Abbott K, Barkmeier-Kraemer J, Hillman RE. Consensus Auditory-Perceptual Evaluation of Voice: Development of a Standardized Clinical Protocol. American Journal of Speech-Language Pathology. 2009;18(2):124-32.
6. 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. Journal of Voice. 1999;13(4):508-17.
7. Nemr K, Simões-Zenari M, Cordeiro GF, Tsuji D, Ogawa AI, Ubrig MT, et al. GRBAS and Cape-V Scales: High Reliability and Consensus When Applied at Different Times. Journal of Voice. 2012;26(6):812.e17-.e22.
8. 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. Journal of Voice. 2007;21(5):576-90.
9. Carding P, Carlson E, Epstein R, Mathieson L, Shewell C. Formal perceptual evaluation of voice quality in the United Kingdom. Logopedics, phoniatrics, vocology. 2000;25(3):133-8.

10. Zraick RI, Kempster GB, Connor NP, Thibeault S, Klaben BK, Bursac Z, et al. Establishing Validity of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V). American Journal of Speech-Language Pathology. 2011;20(1):14-22.
11. Kreiman J, Gerratt BR. Perceptual Assessment of Voice Quality: Past, Present, and Future. SIG 3 Perspectives on Voice and Voice Disorders. 2010;20(2):62-7.
12. Iwarsson J, Reinholt Petersen N. Effects of Consensus Training on the Reliability of Auditory Perceptual Ratings of Voice Quality. Journal of Voice. 2012;26(3):304-12.
13. Sofranko JL, Prosek RA. The Effect of Levels and Types of Experience on Judgment of Synthesized Voice Quality. Journal of Voice. 2018;28(1):24-35.
14. Eadie TL, Kapsner-Smith M. The Effect of Listener Experience and Anchors on Judgments of Dysphonia. Journal of Speech, Language, and Hearing Research. 2011;54(2):430-47.
15. Awan SN, Lawson LL. The Effect of Anchor Modality on the Reliability of Vocal Severity Ratings. Journal of Voice. 2009;23(3):341-52.
16. Brinca L, Batista AP, Tavares AI, Pinto PN, Araújo L. The Effect of Anchors and Training on the Reliability of Voice Quality Ratings for Different Types of Speech Stimuli. Journal of Voice. 2015;29(6):776.e7-.e14.
17. Lu F-L, Matteson S. Speech Tasks and Interrater Reliability in Perceptual Voice Evaluation. Journal of Voice. 2014;28(6):725-32.
18. Law T, Kim JH, Lee KY, Tang EC, Lam JH, van Hasselt AC, et al. Comparison of Rater's Reliability on Perceptual Evaluation of Different Types of Voice Sample. Journal of Voice. 2012;26(5):666.e13-.e21.
19. Shah RK, Woodnorth GH, Glynn A, Nuss RC. Pediatric vocal nodules: Correlation with perceptual voice analysis. International Journal of Pediatric Otorhinolaryngology. 2005;69(7):903-9.
20. Nuss RC, Ward J, Huang L, Volk M, Woodnorth GH. Correlation of vocal fold nodule size in children and perceptual assessment of voice quality. Ann Otol Rhinol Laryngol. 2010;119(10):651-5.
21. Geneid A, Pakkasjärvi N, Aherto A, Roine R, Sintonen H, Lindahl H, et al. Outcomes of early infancy laryngeal reconstruction on health-and voice-related quality of life. International journal of pediatric otorhinolaryngology. 2011;75(3):351-5.
22. Cohen W, Wynne DM, Lloyd S, Townsley RB. Cross-sectional follow-up of voice outcomes in children who have a history of airway reconstruction surgery. Clinical Otolaryngology. 2017:n/a-n/a.
23. Tirado Y, Chadha NK, Allegro J, Forte V, Campisi P. Quality of life and voice outcomes after thyroid ala graft laryngotracheal reconstruction in young children. Otolaryngol Head Neck Surg. 2011;144(5):770-7.
24. Krival K, Kelchner LN, Weinrich B, Baker SE, Lee L, Middendorf JH, et al. Vibratory source, vocal quality and fundamental frequency following pediatric laryngotracheal reconstruction. Int J Pediatr Otorhinolaryngol. 2007;71(8):1261-9.
25. Kelchner LN, Brehm SB, Weinrich B, Middendorf J, deAlarcon A, Levin L, et al. Perceptual evaluation of severe pediatric voice disorders: rater reliability using the consensus auditory perceptual evaluation of voice. Journal of voice : official journal of the Voice Foundation. 2010;24(4):441-9.
26. Hogikyan ND, Sethuraman G. Validation of an instrument to measure voice-related quality of life (V-RQOL). Journal of Voice. 1999;13(4):557-69.
27. Jacobson BH, Johnson A, Grywalski C, Silbergleit A, Jacobson G, Benninger MS, et al. The Voice Handicap Index (VHI)Development and Validation. American Journal of Speech-Language Pathology. 1997;6(3):66-70.
28. de Alarcon A, Brehm SB, Kelchner LN, Meinzen-Derr J, Middendorf J, Weinrich B. Comparison of pediatric voice handicap index scores with perceptual voice analysis in patients following airway reconstruction. Ann Otol Rhinol Laryngol. 2009;118(8):581-6.
29. Johnson K, Brehm SB, Weinrich B, Meinzen-Derr J, de Alarcon A. Comparison of the Pediatric Voice Handicap Index with perceptual voice analysis in pediatric patients with vocal fold lesions. Arch Otolaryngol Head Neck Surg. 2011;137(12):1258-62.

30. Boseley ME, Cunningham MJ, Volk MS, Hartnick CJ. Validation of the Pediatric Voice-Related Quality-of-Life survey. Arch Otolaryngol Head Neck Surg. 2006;132(7):717-20.
31. Cohen W, Wynne DM. Parent and Child Responses to the Pediatric Voice-Related Quality-of-Life Questionnaire. Journal of voice : official journal of the Voice Foundation. 2015;29(3):299-303.
32. Urbaniak GC, Plous S. Research Randomizer (Version 4.0) [Computer software]. . 2013.
33. Shrout PEaF, Joseph L. Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin. 1979;86(2):420-8.