Perceptual Assessment of Tracheoesophageal Voice Quality With the SToPS: The Development of a Reliable and Valid Tool

*,[1]Anne Hurren, †Nick Miller, and ‡Paul Carding, *Sunderland, and †Newcastle upon Tyne, UK, and ‡Brisbane, Australia

Summary: Perceptual assessment of tracheoesophageal voice quality following total laryngectomy with surgical voice restoration is essential to investigate functional outcomes in relation to surgical procedure and rehabilitation regimes. There is no current tool with established reliability and validity to fulfill this purpose. This study describes the development of a set of new perceptual scales, in relation to core validity and reliability issues. These were investigated using voice stimuli from 55 voice prosthesis speakers and evaluated by 22 judges—12 speech and language therapists (SLTs), 10 Ear, Nose, and Throat surgeons—classified into experienced or not at assessing voice. SLT judges rated more parameters reliably than Ear, Nose, and Throat raters, and SLTs with specialist experience in laryngectomy and laryngeal voice attained the most parameters at an acceptable level of agreement. These scales are ready for clinical use, with the most optimal assessors being expert SLTs. Future studies are needed to ascertain precisely how reliability may relate to training, experience, voice stimuli type, and scale format.

Key Words: Laryngectomy−Alaryngeal voice−Surgical voice restoration−Voice assessment−Perceptual voice.

INTRODUCTION

Perceptual rating scales of voice quality continue to be considered a standard in measuring surgical and rehabilitation outcomes in voice quality over time. Although rating scales designed for use in evaluating laryngeal voice are well established, with good validity and reliability,[1] these scales are of minimal value for rating alaryngeal phonation.

Alaryngeal voice is produced by the vibration of reconstructed tissue in the neoglottis or neopharynx. Consequently, voice quality is fundamentally different from that produced by laryngeal vibration and even the key components used to describe voice quality (eg, breathiness, roughness) are of limited relevance to alaryngeal voice analysis. Furthermore, laryngeal voice quality rating scales commonly determine degrees of deviation (severity) away from "normal" (eg, Ref. 2). This concept is of minimal value when rating alaryngeal voices because there are no inherent or intuitive "normative" values of voice quality. Therefore, the continued practice of assessing tracheoesophageal speech "relative to the inherent characteristics of normal laryngeal voice and speech"[3] is fundamentally flawed. Selecting a normal laryngeal voice as a baseline for tracheoesophageal voice measurement (eg, Ref. 4) will result in scores that cluster at the severe end of the scale, thus artificially inflating reliability and compromising validity. Consequently, it is necessary to develop perceptual rating scales that are directly pertinent to the variables found in alaryngeal voice, uninfluenced by parameters that characterize only "normal" phonation.

There has been a proliferation of highly varied informal measures for use across many centers. Of these, a few have been reported in the literature.[5−10] However, all of these scales have significant problems with respect to validity and reliability. For example, unclear rationale of perceptual parameter selection compromises content validity. This includes both voice quality parameters (eg, "breathy,"[11] "rough"[10]) and nonvoice parameters (eg, "intelligibility,"[7] "pleasantness,"[6] and "acceptability"[12]). These descriptors are often poorly defined. In addition, some scales have been devised for use by clinicians (eg, Refs. 11,13), some by nonexperts or "naïve" listeners (eg, Refs. 6,8 or both, eg Ref. 14). Furthermore, the published scales also have limited reporting of inter- and intra-rater reliability. Some either did not use coefficients that calculate for chance agreement[5−7,10,11,13,15] or simply used percentage agreement or mean score calculations.[9,12] Other studies failed to assess inter- and intra-rater reliability at all (eg, Ref. 16).

Therefore, we aimed to develop a tool that (1) was based on perceptual features agreed to characterize variation in alaryngeal/tracheoesophageal voice; (2) employed scales with descriptors that captured those features and that showed acceptable levels of intra- and inter-rater agreement; (3) was easily applicable for clinician use; and (4) required no or minimal training for application.

METHODOLOGY

There were four stages in the development and trialing of this new tool: (1) scale design, including development of

Accepted for publication December 14, 2017.

Conflict of interest: Anne Hurren has undertaken consultancy work for PROTiP Medical and In Health Technologies and their UK distributor. Honoraria and sponsorship for training events have been received from In Health Technologies and their UK distributor. Sponsorship for training event has been received from Atos Medical/Platon Medical. This research was funded by a novice researcher grant from City Hospitals Sunderland NHS Foundation Trust. Nick Miller has no conflicts of interest Paul Carding has no conflicts of interest.

From the *Speech and Language Therapy, Sunderland Royal Hospital, Sunderland, UK; †Institute of Health and Society, Speech and Language Sciences, Newcastle University, Newcastle upon Tyne, UK; and the ‡Speech Pathology, School of Allied Health, Faculty of Health Sciences, Australian National Catholic University (Brisbane, Sydney, Melbourne), Brisbane, Australia.

[1]Present address: Speech and Language Sciences, City Campus, Leeds Beckett University, Leeds, UK LS1 3HE.

Address correspondence and reprint requests to Anne Hurren, Speech and Language Sciences, City Campus, Leeds Beckett University, Leeds, UK LS1 3HE. E-mail: a.hurren@leedsbeckett.ac.uk

Journal of Voice, Vol. 33, No. 4, pp. 465−472
0892-1997

https://doi.org/10.1016/j.jvoice.2017.12.006

guidance notes and pilot studies; (2) recruitment of participants and recording of alaryngeal voice samples that acted as voice stimuli for listeners; (3) judgment of voice stimuli by listener groups, including test-retest rating sessions; and (4) calculation of inter- and intra-rater agreement. These stages are described below.

The study was carried out following permissions from a British National Health Service ethics committee and research governance approval was gained from the appropriate National Health Service Trust.

Scale design

Selection of items for the scale and scale format was guided by a review of existing tools, a literature review of studies reporting validity and reliability of such scales, and consultation with clinicians experienced in the field. A panel of 20 experienced speech and language therapists (SLTs) from a variety of institutions participated in a workshop and discussion that aimed to form consensus on which parameters to rate to achieve clinical utility. The panel also agreed to trial the format of the rating scale and to be involved with the examination of rater agreement. Such consultation throughout the scale design process aimed to ensure maximal content and construct validity.[17,18]

The starting point for the tool developed in this investigation was a previously unpublished four-parameter (quality, acceptability, fluency, and intelligibility) scale,[19] which experienced clinicians advised showed promise but required significant improvements and additions. There was agreement that the parameters social acceptability, fluency, and intelligibility should be included but modified by the prefix "impairment of." The zero-baseline score for social acceptability was defined as representing the optimal level possible for a surgical voice restoration (SVR) speaker, fluency was defined in relation to normal laryngeal speakers in terms of the number of syllables per breath group, and intelligibility was defined in relation to a laryngeal speaker in a one-to-one speaking situation with no background noise. A subsequent detailed literature review suggested inclusion of additional features of alaryngeal phonation: overall grade, wetness, strain, stoma noise, whisper, and impairment of volume.[15,20−23] It was agreed that all of these parameters were to be judged from a baseline of optimal outcome (0 score), with increasing numerical values (scores 1−3) representing mild to severe impairment. A 0−3 equally appearing interval scale has been shown to facilitate high levels of agreement and sensitivity to change in previous studies of perceptual voice ratings,[24] and this strategy was adopted here.

There were two key additional components of alaryngeal voice that did not easily fit a 0−3 equal appearing scale, which were subsequently managed differently. The "quality" parameter from the original rating scale[19] was replaced with a more detailed "tonicity" scale. This was designed to reflect the spectrum of alaryngeal voice qualities that are qualitatively different from those of laryngeal voice, specifically that the tone of the alaryngeal vibratory mechanism can show marked interpatient variation and exists on a spectrum from high (hypertonic) to low (hypotonic) with a midpoint of neutral (optimal tone). A bipolar scale was selected to reflect this physiological continuum of neoglottal tone[21] where 0 represented normal tone, whereas deviations to higher or lower tone were rated 1−5.[25] Additionally, "stenosis" was judged as an all-or-nothing concept as an alternative branch of the tonicity scale. These ratings of tonicity and stenosis reflect Perry's work,[21] which demonstrated the physiology and morphology of the neoglottis as a continuum of tone with an alternative absence of a vibrating segment due to stenosis resulting in a strained, whisper quality.

This resultant tool,[25] the Sunderland Tracheoesophageal Perceptual Scale (SToPS) was taken forward as the version to trial further, and is the subject of development reported in this study. Prior to full trialing, clear definitions and textual reference points for each scale point per parameter were developed.[26] The rating scales were agreed by the SLT panel clinicians as incorporating clinically meaningful perceptual parameters of alaryngeal voice, having a feasible scale point allocation system and potential to differentiate between patients and treatment effects. It was considered easy to administer and score by clinicians and met the needs of clinical practice and outcome measurement.

Recruitment of participants and recording of voice stimuli

Two groups of participants were recruited: (a) people who had undergone SVR, who provided audio recordings for evaluation, and (b) clinicians to perform the rating tasks.

(a) SVR patient recruitment

A total of 73 SVR patients on the current SLT clinical caseload of a regional cancer unit in the North East of England were identified as potential participants. The only inclusion criteria were (1) ability to produce tracheoesophageal voice and (2) English as their first language. The exclusion criteria were (1) inability to read aloud; (2) presence or suspicion of persistent or recurrent cancer; (3) any further speech or language impairment in addition to total laryngectomy/pharyngolaryngectomy. Fifty-seven patients (78%) registered interest after initial invitation; two of these were identified as not meeting the inclusion criteria. The remaining 55 patients attended the voice recording session. The majority of the patients (89%) were born in the area covered by the unit and consequently spoke with a typical local accent. All patients had Blom Singer voice prostheses: duck bill (27.3%), 16-French gauge low pressure exdwelling (29.1%), 20-French gauge low pressure exdwelling (18.1%), and 20-French gauge indwelling (25.5%). The patient demographics are summarized in Table 1.

(b) Clinician recruitment

SLTs and Ear, Nose, and Throat (ENT) surgeons working in SVR in the North of England and Ireland were

TABLE 1.
Patient Demographics

Gender	
Male	49 (89%)
Female	6 (11%)
Age	
Mean (SD)	66 y (SD = 8.34)
Range	48–80 y
Time since operation (mo)	
Mean (SD)	59 mo (SD = 50)
Range	3 mo–17.4 y
Type of surgery	
Total laryngectomy: n	51 (93%)
Pharyngolaryngectomy + jejunum graft: n	4 (7%)
Type of voice prosthesis	
Blom Singer duck bill 16 French gauge: n	15 (27.3%)
Blom Singer low pressure exdwelling 16 French gauge: n	16 (29.1%)
Blom Singer low pressure exdwelling 20 French gauge: n	10 (18.1%)
Blom Singer indwelling 20 French gauge: n	14 (25.5%)

Abbreviation: SD, standard deviation.

invited to participate in the project. Eligible participants had to have worked in a head and neck multidisciplinary team and have experience of managing at least 40 tracheoesophageal speakers. The clinician demographics are summarized in Table 2. Expert SLT raters, classified as those who had worked in both alaryngeal and laryngeal voice rehabilitation at specialist level, had all received postgraduate level training in perceptual assessment of laryngeal voice quality, for example, Vocal Profile Analysis.[27] Expert ENT surgeons were defined as those who are employed in a Consultant grade post with experience of working in joint clinics with SLTs. None of the expert ENT surgeons had undergone any formal perceptual voice evaluation training but had been exposed to large numbers of tracheoesophageal voice speakers and had worked alongside experienced SLTs.

TABLE 2.
Rater Recruitment Demographics

Professional groups	**Experience range by profession (mean)**
Speech and language therapists (SLT), n = 12	SLTs 2–16 y (8 y 6 mo)
ENT surgeons (ENT), n = 10	ENTs 2–25 y (10 y 9 mo)
Total N = 22	
Level of expertise	**Experience range (mean) by expertise**
Expert SLTs[*], N = 5	Expert SLTs[*] 8–16 y (12 y 6 mo)
Nonexpert SLTs, N = 7	Nonexpert SLTs 2–10 y (5 years 6 mo)
Expert ENTs[†], N = 5	Expert ENTs[†] 9–25 y (15 years 10 mo)
Nonexpert ENTs N = 5	Nonexpert ENTs 2–10 y (5 years 10 mo)

[*] Expert SLT raters were classified as those who had worked in both alaryngeal and laryngeal voice rehabilitation at specialist level.

[†] Expert ENT surgeons were defined as those who are employed at Consultant grade with experience of working in joint clinics with SLTs.

Patients then attended for recording of the voice stimuli samples. A number of clinical checks were made for each patient prior to recording to ensure that a representative voice sample was obtained. The voice prosthesis was visualized and cleaned if required, and a new adhesive stoma cover was fitted if the seal had broken. Four patients used a hands-free tracheostoma valve on a part-time basis and were asked to replace them with a heat moisture exchange filter, given that tracheostoma valves cause additional stoma noise. Finally, each patient was asked to confirm that the current voice was typical of his or her usual tracheoesophageal voice. All patients were allowed to practice "The Rainbow" passage[28] aloud and any unfamiliar words were explained before recording commenced.

All recordings were carried out in an audiology soundproof room designed to undertake audiometric testing down to 10 dB in hearing level in the sound field, and all equipment was set up and function-checked by a consultant clinical scientist (audiology) prior to each session. A Sony Electret Condenser (Tokyo, Japan) microphone was attached to a microphone stand positioned exactly 1 m from the patient's mouth. The mouth to microphone distance is greater than that specified in the majority of both laryngeal and alaryngeal recording protocols to allow for realistic listener representation of any tracheostoma sounds emitted. Samples were recorded onto a Sony MD Walkman at maximum volume. A 70-dB sound pressure level warble tone (frequency modulated [FM] 1 kHz) was recorded as a calibration tone.

The anonymized Rainbow passage recordings were edited onto a master minidisc. The order of tracks was randomized using the minidisc system shuffle facility for initial rating and then reshuffled to produce a different order for a second master disc for use during the retest rating session.

Test-retest rating sessions

All raters (both SLT and ENT) underwent a 3-hour face-to-face training program with the first author, including instruction in the written guidance notes and trialing the scale with sample voice stimuli (not used in the study). Opportunity for discussion was an integral part of training.

All clinicians evaluated all 55 subjects' voice samples within 2 weeks of attending training. Judges could refer to the written guidance notes during the voice evaluation. The only information provided about each speaker was the gender. Raters heard each sample twice and could request a third repeat if desired. Raters could chose to listen to the voices in an organized session with other clinicians or individually with Sony headphones (Tokyo, Japan) (model

MDR-XD200) connected to their personal computer or laptop. No discussion about the task was permitted during the rating sessions or during breaks.

To measure intra-rater reliability the procedure was repeated in identical fashion 1 month later, with all the voices presented in a different random order. However, the training session was not repeated, but raters were asked to read the rating guidance information sheet again prior to starting the retest ratings. One ENT surgeon did not repeat the retest task due to workload pressure.

Inter- and intra-rater analysis

The raw scores assigned to each voice sample in both rating sessions were entered into *Cytel Studio 8* (Cambridge, USA). A separate database was created for each parameter. Each database was analyzed with the StatXact package to calculate quadratic weighted kappa coefficients for both intra- and inter-rater reliability. The range and mean of the raw kappa scores were calculated for all 22 raters from both professional groups. Further analysis involved categorizing the professional raters into seven groups: that is, all raters, all SLT, all ENT, expert SLT, expert ENT, nonexpert SLT, nonexpert ENT. The range and mean kappa coefficients were calculated for each of the seven groups.

RESULTS

The results of the evaluation of the 55 voice samples on all the tool's subscales are reported firstly for intra-rater agreement, then for inter-rater judgment before concluding with a summary of the overall reliability results for the SToPS.

Intra-rater agreement

The mean scores for intra-rater reliability of the seven subgroups of raters for each parameter are summarized in Table 3. Landis and Koch[29] define a "good" level of agreement as a mean weighted kappa coefficient of 0.61 or above. These are marked with an asterisk in Table 3. All SLT raters attained "good" agreement for all 10 parameters.

A profession-specific effect was apparent in terms of SLT raters achieving higher mean coefficient scores than ENT raters for all parameters. The all ENT subgroup achieved "good" intra-rater agreement for 8 of the 10 parameters, with the remaining 2 (tonicity and wetness) classified as "moderate" agreement. Wetness (mean 0.59) fell only 0.02 below a classification of "good" agreement.

The degree of expertise appears to show differences for SLT raters but not for ENT raters. The expert SLT subgroup attained higher kappa means than the nonexpert SLTs for eight parameters, with one parameter attaining an identical mean for both groups. This contrasts with the expert ENT subgroup who had higher coefficient scores than their nonexpert colleagues for only 2 of the 10 parameters (volume and stoma noise) and had equivalent agreement in one parameter (strain).

Inter-rater agreement

Table 4 shows the inter-rater agreement mean coefficients in relation to parameter and rater subgroups (profession and level of expertise). For all raters (as a whole group), the inter-rater coefficient means were lower than for intra-rater, with only 3 parameters (compared to 10 for intra-rater agreement) achieving "good" levels of agreement (overall grade, social acceptability, and strain). Analysis by profession type showed that the SLT group achieved "good" agreement for six parameters (overall grade, strain, volume, social acceptability, whisper, and fluency), which again reflected less agreement than for intra-rater judgments. The ENT group reached "good agreement" levels for only three parameters (overall grade, strain, and social acceptability). These outcomes suggest an effect of level of expertise.

The expert SLT group achieved "good" levels of inter-rater agreement in 9 out of the 10 parameters compared to only 5 for their less experienced SLT colleagues. Expert

TABLE 3.
Intra-rater Mean Weighted Kappa Coefficients for Each Parameter in Relation to Rater Groups

Intra-rater	All Professionals	All SLT	All ENT	Expert SLT	Expert ENT	Nonexpert SLT	Nonexpert ENT
Overall grade	0.78*	0.80*	0.77*	0.84†	0.71*	0.77*	0.81†
Tonicity	0.64*	0.70*	0.56	0.74*	0.53	0.68*	0.59
Strain	0.74*	0.75*	0.72*	0.79*	0.72*	0.72*	0.72*
Wetness	0.67*	0.73*	0.59	0.73*	0.57	0.73*	0.60
Volume	0.72*	0.76*	0.68*	0.77*	0.71*	0.76*	0.65*
Social acceptability	0.75*	0.77*	0.64*	0.78*	0.68*	0.76*	0.77*
Whisper	0.69*	0.73*	0.64*	0.69*	0.61*	0.76*	0.66*
Intelligibility	0.68*	0.72*	0.64*	0.73*	0.63*	0.71*	0.65*
Stoma noise	0.64*	0.66*	0.61*	0.70*	0.66*	0.64*	0.57
Fluency	0.68*	0.70*	0.65*	0.71*	0.64*	0.70*	0.65*

* Defined as "good" level of agreement as defined by Landis and Koch (1977).
† Defined as "very good" level of agreement as defined by Landis and Koch (1977).

TABLE 4.
Inter-rater Mean Weighted Kappa Co-Efficients for Each Parameter in Relation to Rater Groups

Inter-rater	All Professionals	All SLT	All ENT	Expert SLT	Expert ENT	Nonexpert SLT	Nonexpert ENT
Overall Grade	0.70*	0.70*	0.69*	0.77*	0.66*	0.66*	0.72*
Tonicity	0.40	0.51	0.40	0.63*	0.45	0.42	0.32
Strain	0.61*	0.62*	0.61*	0.74*	0.63*	0.54	0.55
Wetness	0.49	0.56	0.48	0.64*	0.42	0.53	0.54
Volume	0.56	0.62*	0.56	0.64*	0.64*	0.61*	0.49
Social Acceptability	0.68*	0.74*	0.63*	0.76*	0.57	0.74*	0.68*
Whisper	0.58	0.63*	0.54	0.62*	0.54	0.62*	0.56
Intelligibility	0.57	0.59	0.58	0.61*	0.60	0.55	0.52
Stoma Noise	0.51	0.55	0.47	0.56	0.43	0.55	0.49
Fluency	0.59	0.61*	0.58	0.68*	0.58	0.62*	0.60

* "good" level of agreement as defined by Landis and Koch (1977).

TABLE 5.
Parameters With "Good" (Landis and Koch 1977) Inter- and Intra-rater Agreement According to Profession and Expertise

	All Raters	Expert SLT	Expert ENT	Nonexpert SLT	Nonexpert ENT
Overall grade	✓	✓	✓	✓	✓
Tonicity		✓			
Strain	✓	✓	✓		
Wetness		✓			
Volume		✓	✓	✓	
Social acceptability	✓	✓		✓	✓
Whisper		✓		✓	
Intelligibility		✓			
Stoma noise					
Fluency		✓		✓	

SLTs also attained higher or equivalent mean kappa coefficients compared to the nonexpert SLT group for all 10 parameters. This effect was not seen for the two ENT groups, with little difference between expert and nonexpert results. Furthermore, expert ENTs only attained superior coefficients for five parameters compared to their nonexpert colleagues.

The overall summary of the parameters that attained agreement for both intra- and inter-rater judgments according to profession and expertise are in Table 5.

DISCUSSION

This study aimed to devise a perceptual rating scale for the assessment of alaryngeal voice quality that could be used in clinical settings. The aim was also to provide clarity about parameter selection and definition and investigate reliability and validity of the chosen scales. As an unreliable scale is inherently invalid, reliability will be considered prior to validity in the discussion below followed by consideration of the evidence that could support the clinical application of this scale.

Reliability

All rater groups

The results from this study demonstrated "good" (>0.60[29]) intra-rater reliability for all 10 parameters as assessed by all 21 raters (Table 3). This suggests listeners have a relatively stable internal baseline for these parameters against which a psychoacoustic evaluation can be made.[30] However, with the exception of intra-rater judgments for "overall grade," no parameter attained "very good" intra-rater agreement (coefficients over 0.81).

There is an expectation that intra-rater reliability will be superior to inter-rater reliability in perception of voice quality. This is because individual listeners may have a relatively stable internal benchmark for judgment of parameters, but because benchmarks are not identical across listeners, the extent of deviation across samples is not perceived identically.[30] Consequently, comparing a judge's individual perceptions to that of other raters will always have lower reliability. For the all rater group, only three parameters (overall grade, social acceptability, and strain) showed "good" inter-rater reliability. Although the remaining 7 parameters had limited inter-rater reliability when all raters were analyzed together, differences emerged when profession and expertise categories were considered (Table 5), and this will be discussed in more detail below. Previous studies in tracheoesophageal voice perception[10,11] reported higher inter- and intra-rater coefficients than those seen in this study. These apparent differences may be explained by the choice of statistical analysis in those studies, which failed to account for chance agreement and is therefore likely to inflate agreement coefficients.[18] Furthermore, one of these studies[10] had a baseline for the overall judgment scale that bears a relationship to normal laryngeal voice quality, which, with its three-point scale, could potentially polarize these atypical voices away from the optimal score, resulting in artificially greater reliability.

Profession

Inter-rater reliability was considerably different between SLTs and ENT surgeons. SLT raters attained "good" intra-rater agreement for six parameters (overall grade, social acceptability, strain, volume, whisper, fluency), whereas ENT surgeon judges only attained this level for the first three parameters. Therefore, SLTs appear to have more stable internalized representations that allowed them to reach agreement for more parameters of tracheoesophageal voice quality, including a greater number of the unidimensional parameters. This potentially relates to SLTs receiving preregistration training in perceptual analysis of the contributory components of communication impairments, that is, speech, language, voice, and resonance, plus the ongoing, routine use of these skills in clinical practice. In contrast, surgeons do not receive any formal auditory perceptual preregistration training.

Although previous investigations have included ENT surgeons as raters (eg, Ref. 31), none has examined profession as an independent variable. Kazi et al[4,32] reported on ENT surgeons (n = 2) ratings but used an invalid tool for assessing tracheoesophageal voice,[33,34] and the difficulties of comparison of data in this study are compounded by methodological and statistical analysis problems. Coffey[35] provides directly comparable inter-rater (but not intra-rater) data for SLTs using the tool described in this study reporting "good" inter-judge agreement for 9 of the 10 parameters listed in Table 5, but tonicity did not achieve "good" agreement. Coffey selected a different definition of expertise, that is, all three of her raters had specialist knowledge of SVR and had attended an advanced postgraduate laryngectomy training course. Further studies are needed to ascertain reasons why this did not translate into Coffey's group attaining "good" agreement for tonicity when they reached higher inter-rater agreement for the other parameters. This issue is discussed in the following "Expertise" section.

Expertise

Increased expertise was, in general, linked to higher inter- and intra-rater reliability for SLT raters, but not for ENT surgeons in this study. Again, this may relate to the absence of formal training for expert ENT surgeons who for the purposes of this study were classified as those who work alongside SLTs in joint clinics. In contrast, all expert SLTs in this study had undergone postgraduate training in voice analysis and had worked at a specialist level in the field of laryngeal voice disorders in addition to SVR. Training in, and repeated exposure to auditory perceptual rating of voice quality (at least for laryngeal voices) is a defined specialist skill[1] and could account for the findings in this study. The influence of rater expertise on inter-rater reliability has received little attention in the literature on alaryngeal voice evaluation. Our study indicates that expertise may be important. Expert SLTs attained "good" agreement for nine parameters compared to only five by their less expert colleagues.

Previous studies have compared SLT student ratings with more expert raters for both alaryngeal[7] and laryngeal stimuli.[36] In all cases, clinicians achieved better inter-rater reliability than inexperienced student SLTs (despite identical pretask training, practice rating and contextual parameters, and anchor reference samples). It would appear that raters who have limited experience of formal perceptual evaluation have been linked to a higher likelihood of increased variations in internal standards.[37]

Parameter types: global versus unidimensional

Finally, it is important to note that both intra- and inter-rater reliability were superior for global "overall severity" type parameters, that is, overall grade and social acceptability compared to others that require more complex discrimination (referred to as "unidimensional" parameters; Tables 3 and 5). This is not a surprising finding as this has been reported in previous investigations of tracheoesophageal voice perception.[7,13] However, several other studies have reported that the overall grade did not achieve the highest coefficients.[10,35] The possible reasons for this variance of findings are not clear but may pertain to the scale format/voice stimuli variations. It is interesting to note that superior inter- and intra-rater reliability for "overall grade" or overall "severity" judgments also holds for laryngeal perceptual assessment.[24]

Validity

The ultimate aim of validity testing is to determine the "inferentiality" of an instrument.[18] Content, criterion, and construct validity are considered as related components of the same core phenomenon.[17] Content validity remains a key tenet of inferentiality and is defined as a judgment as to whether a scale looks reasonable and samples all the relevant aspects. This is essentially a subjective judgment,[18] but careful planning during scale development strengthens claims for content validity.[38,39] The SToPS was based on a preexisting tool that was subsequently amended via a comprehensive literature review and discussions with a large expert panel. This ensured that each parameter that was selected was considered a key feature of tracheoesophageal voice and had maximal clinical (content) validity.

Concurrent or criterion validity typically requires a new tool undergoing development to be compared to a preexisting scale or "gold standard" measurement.[18] Unfortunately, there is no accepted criterion validity to act as a "gold standard" to determine the optimal baseline for tracheoesophageal voice. Furthermore, voice parameters are hypothetical constructs as the "true rating" for each voice is not a function of the voice per se but an interaction between the speech signal and the psychoacoustic perception of the listener.[40] Raters are required to agree on borders between qualities and assign a numerical value to the attribute perceived. This is the format (0−3 equal appearing scale) that is used for most of the parameters in this new tool. Construct validity testing cannot be proven definitively[17] because it is viewed as a continuous, cyclical process where testing helps move toward an understanding of the construct which in turn helps to set new predictions. This aspect of the scale's validity was partially addressed within the reliability section above. Certain parameters appear to have more construct validity in terms of them having a higher agreement coefficient score. A parameter that has low agreement is inherently invalid. Although agreement of judges does not amount to an assurance that the correct rating, and, consequently, construct validity has been attained, the patterns of sufficient agreement for SLT raters contribute to the cycle of validity testing for these hypothetical constructs.

CLINICAL APPLICATION

A clinical tool such as this newly developed scale can only be recommended for clinical application if it can demonstrate appropriate levels of validity and reliability. It would appear that this tool has achieved acceptable aspects of both content and construct validity testing particularly with respect to content validity. Other studies (eg, Ref. 35) provide additional supporting evidence of validity among the SLT clinical community.

The reliability results of both this study's and Coffey's data using the SToPS[35] have implications for future use in both clinical practice and research. These implications relate to which parameters may be selected and who should perform the rating judgments. The data presented in this paper suggest that expert SLT ratings are the most reliable and hence the most useful in determining surgical outcomes and informing decision about SVR longitudinal management within a multidisciplinary environment. The data also suggest that expert SLTs can proceed to use routinely all 10 parameters within the SToPS (overall grade, social acceptability, intelligibility, tonicity, stoma noise, strain, wetness, impairment of volume, fluency, and whisper). Tone of the neoglottis accounts for the marked variability of voice quality following total laryngectomy.[21] It is therefore particularly significant that the expert SLTs in this investigation were the only group able to demonstrate acceptable inter- and intra-rater reliability in judging tonicity. This contrasts with Coffey's[35] report that her SLT cohort failed to achieve acceptable reliability of this key parameter although her definition of what constituted "expert" was different from the present study. The study found that expert SLTs achieved acceptable agreement with similar training regimes and identical guidance notes. Caution may be required when SLTs with limited expertise in perceptual voice assessment assess tonicity.

A further important consideration for clinical application relates to all raters in this study and Coffey's investigation undergoing a three-hour training session. The laryngeal perceptual scales (GRBAS [Grade, Roughness, Breathiness, Aesthenia and Strain] and CAPE-V]Consensus Auditory-Perceptual Evaluation of Voice]) do not mandate a skill acquisition protocol and are freely available for SLTs to use. The SToPS' guidance notes developed for this study are available online (26) to assist others who wish to utilize the tool. However, further investigations are required to ascertain whether the same reliability can be achieved without training. If training is demonstrated to be an essential aspect of agreement, then development of an online tutorial, including anchor voice stimuli, may facilitate reliability development without the requirement for the travel, expense, and time needed to attend a formal course for skill acquisition.

The establishment of reliable and valid perceptual features of tracheoesophageal voice has significant research implications. Beyond the role of the tonicity of neoglottis, there is minimal knowledge about how alaryngeal anatomical and physiological components relate to specific perceptual characteristics, for example, how perceptual voice ratings relate to videofluoroscopic, tracheal, and esophageal manometry measurement. Reliable perceptual rating also allows comparisons between surgical technique and SVR valve variables. In addition, measures to improve voice quality (eg, pharyngoesophageal myotomy, botulinum toxin injection, hands-free valve, voice therapy) can be evaluated and quantified. The SToPS is in its infancy, and more studies are planned to include different voice stimuli and raters, aiming to investigate the effects of expertise, training, and anchor stimuli use in relation to agreement patterns. Future investigation will also include exploration of the guidance notes; some parameter definitions and scale points may require revision to provide more clarity for raters, for example, fluency to include total cessation of voicing due to spasm of the neoglottis. If training is demonstrated to be unnecessary or if it can be provided to the same effect online, then adopting this new tool[25] as a standard measure will allow comparison between centers and across publications, thus providing a consistency of approach to build a strong evidence-base behind surgical voice restoration and tracheoesophageal voice quality practice. It is hoped that future publications that utilize tracheoesophageal voice rating will routinely report inter-and intra-rater reliability to ensure that this perceptual (and therefore subjective) judgment had maximal methodological rigor.

CONCLUSION

There is evidence to support the validity, reliability, and clinical application of the SToPS[25] for 10 key parameters. Currently this conclusion is based on its utilization with expert SLTs (only) who have undergone a training protocol. As intimated, further research is required to establish if similar results are obtained without specialist training. Even though the study represented one of the largest, in terms of number of raters, to examine perceptual evaluation of alaryngeal voice, the size of the rater groups in the different fields of expertise in this study are small. Further research will be necessary across larger numbers of raters to see if levels of agreement are maintained. It will also be beneficial to examine the reliability and validity of this new tracheoesophageal voice assessment with additional cohorts of raters and voice stimuli samples. A final line of further inquiry relates to definitions used for the scale parameters and scale format. This appears especially pertinent for tonicity as the key determinant of tracheoesophageal voice quality. It is hoped that this new tool could provide the catalyst for reliable measures of alaryngeal voice quality, which in turn will enable the evaluation of the impact on multiple management variables that may determine voice outcomes.

REFERENCES

1. Carding P. *Evaluating the Effectiveness of Voice Therapy*. Compton: Oxford; 2017.
2. Hirano M. *Clinical Examination of Voice*. New York: Springer; 1981.
3. Doyle PC, Eadie TL. The perceptual nature of alaryngeal voice and speech. In: Doyle PC, Keith RL, eds. *Contemporary Considerations in the Treatment and Rehabilitation of Head and Neck Cancer: Voice, Speech and Swallowing*. Austin, TX: Pro-Ed Inc; 2005.
4. Kazi R, Kiverniti E, Prasad V, et al. Multidimensional assessment of female tracheoesophageal prosthetic speech. *Clin Otolaryngol*. 2006;31:511–517.
5. Eadie TL, Doyle PC. Auditory-perceptual scaling and quality of life in tracheoesophageal speakers. *Laryngoscope*. 2004;114:753–759.
6. Eadie TL, Doyle PC. Scaling of voice pleasantness and acceptability in tracheoesophageal speakers. *J Voice*. 2005;19:373–383.
7. Moerman M, Martens JP, Crevier-Buchman L, et al. The INFVo perceptual rating scale for substitution voicing: development and reliability. *Eur Arch Otorhinolaryngol*. 2006;263:435–439.
8. Nagle KF, Eadie TL. Listener effort for highly intelligible tracheoesophageal speech. *J Commun Disord*. 2012;45:235–245.
9. Nieboer GLJ, de Graaf T, Schutte HK. Esophageal voice quality judgements by means of the semantic differential. *J Phon*. 1988;16:417–436.
10. van As CJ, Koopmans-van Beinum FJ, Pols LC, et al. Perceptual evaluation of tracheoesophageal speech by naive and experienced judges through the use of semantic differential scales. *J Speech Lang Hear Res*. 2003;46:947–959.
11. Lundstrom E, Hammarberg B, Munck-Wikland E, et al. The pharyngoesophageal segment in laryngectomees—videoradiographic, acoustic and voice quality perceptual data. *Logoped Phoniatr Vocol*. 2008;33:115–125.
12. Finizia C, Lindstrom J, Dotevall H. Intelligibility and perceptual ratings after treatment for laryngeal cancer: laryngectomy versus radiotherapy. *Laryngoscope*. 1998;108:138–143.
13. Ward EC, Hancock K, Lawson N, et al. Perceptual characteristics of tracheoesophageal speech production using the new indwelling provox vega voice prosthesis: a randomized controlled crossover trial. *Head Neck*. 2011;33:13–19.
14. Delsupehe K, Zink I, Lejaegere M, et al. Prospective randomized comparative study of tracheoesophageal voice prosthesis: Blom-Singer versus Provox. *Laryngoscope*. 1998;108:1561–1565.
15. Eadie TL, Doyle PC. Direct magnitude estimation and interval scaling of naturalness and severity in tracheoesophageal (TE) speakers. *J Speech Lang Hear Res*. 2002;45:1088–1096.
16. O'Leary IK, Heaton JM, Clegg RT, et al. Acceptability and intelligibility of tracheoesophageal speech using the Groningen valve. *Folia Phoniatr Logop*. 1994;46:180–187.
17. McDowell I, Newell C. *Measuring Health: A Guide to Rating Scales and Questionnaires*. Third ed New York, NY: Oxford University Press; 2006.
18. Streiner DL, Norman GR. *Health Measurement Scales: A Practical Guide to Their Development and Use*. Second ed Oxford: Oxford University Press Inc.; 1995 Oxford Medical Publications.
19. O'Leary I. *A preliminary report on the use of the Groningen tracheoesophageal valve in laryngectomy patients in the UK, Head and Neck Oncology Conference*. Nottingham, UK.
20. Omori K, Kojima H. Neoglottic vibration in tracheoesophageal shunt phonation. *Eur Arch Otorhinolaryngol*. 1999;256:501–505.
21. Perry A. Vocal rehabilitation after total laryngectomy. *Leicester School of Speech Pathology*. 1989176 Leicester.
22. Silverman AH, Black MJ. Efficacy of primary tracheoesophageal puncture in laryngectomy rehabilitation. *J Otolaryngol*. 1994;23:370–377.
23. Singer MI, Blom ED, Hamaker RC. Pharyngeal plexus neurectomy for alaryngeal speech rehabilitation. *Laryngoscope*. 1986;96:50–53.
24. Webb AL, Carding PN, Deary IJ, et al. The reliability of three perceptual evaluation scales for dysphonia. *Eur Arch Otorhinolaryngol*. 2004;261:429–434.
25. Hurren A. The Sunderland Tracheosophageal Perceptual Scale, Leeds Beckett University Repository; 2017. Available at: http://eprints.leedsbeckett.ac.uk/4126/.
26. Hurren A. Guidance Notes for The Sunderland Tracheosophageal Perceptual Scale, Leeds Beckett University Repository; 2017. Available at: http://eprints.leedsbeckett.ac.uk/4126/.
27. Wirz S, Laver J, Mackenzie J. Vocal profile analysis scheme. *Folia Phoniatr (Basel)*. 1983;35:183–184.
28. Fairbanks G. *Voice and Articulation Drillbook*. New York: Harper Row; 1960.
29. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33:159–174.
30. Kreiman J, Gerratt BR, Kempster GB, et al. Perceptual evaluation of voice quality—review, tutorial, and a framework for future-research. *J Speech Hear Res*. 1993;36:21–40.
31. Bridges A. Acceptability ratings and intelligibility scores of alraryngeal speakers by three listener groups. *Br J Disord Commun*. 1991;26:325–335.
32. Kazi R, Kanagalingam J, Venkitaraman R, et al. Electroglottographic and perceptual evaluation of tracheoesophageal speech. *J Voice*. 2009;23:247–254.
33. Hurren A, Hildreth AJ, Carding PN. Can we perceptually rate alaryngeal voice? Developing the Sunderland Tracheoesophageal Voice Perceptual Scale. *Clin Otolaryngol*. 2009;34:533–538.
34. Schindler A, Ginocchio D, Atac M, et al. Reliability of the Italian INFVo scale and correlations with objective measures and VHI scores. *Acta Otorhinolaryngol Ital*. 2013;33:121–127.
35. Coffey M. *A comparison of Fiberoptic Endoscopic Evaluation and Videofluoroscopy in Post Laryngectomy Swallowing, and Swallow and Voice Evaluation with Different Voice Prostheses*. Imperial College London; 2013 Department of Medicine.
36. Bele IV. Reliability in perceptual analysis of voice quality. *J Voice*. 2005;19:555–573.
37. Kreiman J. Listening to voices: theory and practice in voice perception research. In: Johnson JW, Mullinex K, eds. *Talker Variability in Speech Processing*. San Diego: Academic Press Inc.; 1997:85–107.
38. Cronbach LJ. *Essentials of Psychological Testing*. 5th edition New York: Harper and Row; 1990.
39. Nunally JCJ. *Introduction to Psychological Measurement*. New York: McGraw-Hill; 1970.
40. Kreiman JG, Gerratt BR. Comparing two methods for reducing variability in voice quality measurements. *J Speech Lang Hear Res*. 2011;54:803–812.