An investigation of reliability of the Sunderland Tracheosophageal Voice Perceptual Scale

Abstract

Introduction: The consensus on how to effectively evaluate alaryngeal voice outcomes remains limited. The Sunderland Tracheosophageal Voice Perceptual scale (SToPS) was developed as a perceptual rating scale specifically for tracheosophageal voice [1]. Currently, it is the only tracheosophageal voice specific perceptual scale available and aims to address the limitations of previous scales.

Objective: To investigate inter rater reliability of the Sunderland Tracheosophageal Voice Perceptual Scale when analysing alaryngeal voice across a range of voice prostheses.

Methods: Prospective evaluation of inter rater reliability of the SToPS based on audio recordings of 230 voice samples from 41 laryngectomy patients rated by 3 experts. Interval data was analysed using Intraclass Correlation Coefficients (ICC*)* while categorical data was analysed using Kappa.

Results: ICC of above 0.6 was observed between raters for each prosthesis on a majority of parameters demonstrating a good level of reliability. Reliability was fair (ICC of between 0.40-0.59) on Q11 (Articulatory precision) and Q12 (Paralinguistics). Reliability was also fair (0.21-0.40) or slight (0.00-0.20) for Q2 (Tonicity), which was analysed using Kappa. Kappa of above 0.61 signified a good level of reliability.

Conclusions: This study demonstrates good rater reliability for the majority of parameters on the SToPS scale, supporting the use of this tool within the clinical realm. However further research is required to ascertain if any methods of increasing inter rater reliability on those parameters which did not reach good reliability can be identified.

Level of evidence: 2b Individual cohort study

Introduction

Laryngectomy involves the removal of the larynx in its entirety, usually as a treatment for advanced laryngeal cancer. As a consequence, this surgery profoundly affects the ability to communicate. The gold standard for communication rehabilitation after laryngectomy is surgical voice restoration (SVR) [2] [3] also known as tracheosophageal voice. This technique involves the placement of a one way valved voice prosthesis in a puncture between the trachea and oesophagus [4] [5]. The voice prosthesis shunts lung air from the oesophagus to a vibratory segment within the reconstructed throat to produce tracheosophageal voice. The ultimate objective of SVR is to provide the patient with the optimal voice possible without a larynx [1]. However, consensus on the most appropriate measure of voice outcome post laryngectomy is lacking..

Evaluation of post laryngectomy voice

Although most of the empirical research concerning laryngeal voice has focused on acoustic measures of frequency, intensity and duration, these measures do not necessarily indicate how well an individual communicates in a social situation. Auditory perceptual rating involves an expert listener judging a voice sample according to different parameters [6] which may include intelligibility, voice quality and acceptability [7]. Auditory perceptual evaluation of tracheosophageal voice quality has been posited as the most valid measure of SVR outcome [1]. There are a number of well-established voice quality rating scales which provide perceptual parameters for the patients with a larynx including the Buffalo Voice Profile [8], the Vocal Profile Analysis Scheme [9], Grade, Roughness, Breathiness, Asthenia, Strain

(GRBAS) scale [10], and Consensus Auditory Perceptual Evaluation of Voice [11]. Of these, the strongest validity and reliability has been established for the GRBAS [12,13]. The GRBAS has been used to assess auditory perceptual aspects of tracheosophageal voice in several studies [14-16] [17,18]. However, use of the GRBAS to measure perceptual aspects of tracheosophageal voice has been considered suboptimal due to the fundamental differences in tracheosophageal and laryngeal voice [1]. As the phonatory source of alaryngeal voice (vibratory segment) contrasts significantly with that of laryngeal voice (vocal folds), the use of a rating scale validated for the latter population poses limitations for post laryngectomy patients. Additionally, some perceptual features of alaryngeal voice such as tone and extraneous noise when covering the stoma to produce voice are unique and central to tracheosophageal voice quality and are not included in the GRBAS scale. Critically, studies which have used the GRBAS [14-18] or other perceptual scales [19,20] have failed to specify an anchor baseline so it is unclear whether raters have compared voice stimuli to that of normal laryngeal voice or optimal tracheosophageal voice.

STOPS

The Sunderland Tracheosophageal Voice Perceptual scale (SToPS) was developed as a perceptual rating scale specifically for tracheosophageal voice [1]. Currently, it is the only tracheosophageal voice specific perceptual scale available. The SToPS was developed as means of overcoming the major conceptual and methodological problems inherent in other studies of tracheosophageal voice, such as poorly defined terminology and impressionistic vocabulary [21]. The STOPS includes specific and clear guidance to define terminology used for each parameter. In addition, the SToPS crucially defines the anchor baseline for parameters as optimal tracheosophageal voice rather than normal laryngeal voice.

Reliability

Measurement is a way of understanding, evaluating and differentiating characteristics of people and objects [22] and forms the basis for making decisions or drawing conclusions in scientific research. A crucial prerequisite for clinical measurement is reliability. Reliability indicates the consistency and lack of errors in a tool [22,23]. As the ability to simply produce voice with a prosthesis following SVR is unlikely to be sufficient indication of functional ability to communicate in everyday situations, it is of clinical relevance to investigate the reliability of the SToPS. As intra rater reliability for expert raters had previously been established as good or above for all parameters of the SToPS except for accent, reading ability and articulatory precision [24], this study focuses on the investigation of inter rater reliability.

Aim

To investigate inter rater reliability of the Sunderland Tracheosophageal Voice Perceptual Scale

Hypothesis

Experts will not achieve a good level of inter rater reliability when they use the SToPS to rate alaryngeal voice. Should a good level of inter rater reliability be achieved, this will support the clinical relevance for the SToPS in identifying functional tracheosophageal voice for patients post laryngectomy.

Methods

SToPS

The Sunderland Tracheosophageal Voice Perceptual Scale (SToPS) for professional raters was originally developed as a 14-item auditory perceptual scale divided into two domains: (i) Six Voice quality parameters (perceptual voice tonicity, strain, wetness, impairment of volume, impairment of social acceptability of voice and whisper), and (ii) seven parameters not related to voice quality (impression of intelligibility, stoma blast, impairment of fluency, impairment of articulatory precision, positive features of articulation, accent and poor reader) and an overall score voice rating. The scale later underwent item reduction and now contains 10 parameters. Ref

Tone relates to the amount of pressure used to produce tracheosophageal voice. The perceptual voice tonicity parameter is measured on an 11 point bipolar semantic scale reflecting the continuum of tone [25] from hypotonic (too little tone) to hypertonic (too much tone)[1]. As stenotic voice occurs only in the absence of tone it is measured with a separate arm to the tone scale [1]. As stenosis is either present or absent, it is not rated along a graded continuum. For each individual voice sample, only one arm of the scale is chosen by a rater. Each of the remaining 5 items in the voice quality parameters domain are measured on a 4 point equally appearing interval scale 0 (optimal tracheosophageal voice quality), 1 (mild), 2 (moderate) and 3 (severe).

Each of the parameters not related to voice quality, with the exception of positive features of articulation is measured on a 4 point equally appearing interval scale 0 (optimal tracheosophageal voice quality), 1 (mild), 2 (moderate) and 3 (severe). Positive features of articulation are measured on an alternatively worded 4 point equally appearing interval scale 0 (neutral), 1 (good), 3 (excellent), and 4 (outstanding).

The parameter 'overall grade" is measured using a four point interval scale 0 = Excellent; 1 = Good; 2 = Adequate; 3 = Poor. This design is similar to the GRBAS scale [10] except that the value 0 represents optimal tracheosophageal voice quality as opposed to "normal" laryngeal voice quality.

Raters

Three Speech and Language Therapy raters were chosen. Each rater had at least five years experience specialising in the rehabilitation of communication post laryngectomy and other head and neck cancer patients and had completed advanced training in the field.

Training of raters

Each rater participated in three hours of training with the investigator in the use of the SToPS. This training took place during two conference calls of 90 minutes length and included practice ratings of ten anonymised audio samples of laryngectomy participants reading the Rainbow Passage. During training queries about individual items on the SToPS scale were raised. These parameters were discussed with the main author of the SToPS. Clarifications provided were passed onto all three raters regardless of how many raters had initially raised a query.

Voice stimuli

230 voice samples were elicited from 41 post laryngectomy participants. Please see table 1 for demographic details. Participants were recruited from the outpatient caseload of Head and Neck cancer patients at a large centre in ___. Exclusion criteria included participants without a voice prosthesis, less than 3 months post surgery or post operative oncological treatment. Each participant trialled up to 6 randomised voice prostheses over 2 appointments within a 72 hour period. Participants were blinded to prosthesis type and a voice sample was provided for each for each prosthesis. This data was used in a subsequent study investigating the differences between voice prostheses in terms of voice outcome.

For each prosthesis trial, participants had a Speedlink SL-8691-SBK spes clip on metal microphone (Speedlink, Weertzen, Germany) attached to their clothing 10 cm lateral to the stoma on the opposite side to the hand used to occlude the stoma during voicing. All subjects produced voice by occluding their stoma rather than depressing a humidification exchange device or using a hands free attachment. Subjects read a short version of the Rainbow passage [26], (see appendix). This was recorded onto a Sony ICD-PX820 Digital Voice Recorder with flash 2 GB (Sony, Weybridge, UK) in MP3 format to be rated later by experts.

Data analysis

Recordings of voice samples with individual voice prostheses were extracted in MP3 format and transferred to Final Cut Pro (Apple, California, USA) to allow titles to be added to indicate anonymised subject number and anonymised voice prosthesis letter. Voice samples were then exported to 3 Verbatim 4GB pinstripe

USB memory sticks (Verbatim, Surrey, UK). Raters were blinded to subject, prosthesis type, gender, type of laryngectomy surgery (extended laryngectomy or standard total laryngectomy) and history of radiotherapy and chemotherapy. Voice samples were posted to 3 expert Speech and Language Therapy raters along with blank numbered and lettered SToPS forms which corresponded to each voice sample for each subject.

Statistical analysis

Data was entered and analysed in IBM SPSS (Statistical Product and Service Solutions) version 23 (IBM Armonk, New York). The SToPS consists of 14 parameters, 13 of which (Q1, Q-Q14), are rated from 0-3 on an interval scale. A further parameter, Q2 of the SToPS is rated on an 11 point bipolar semantic scale which yielded categorical data.

Intraclass correlation coefficients (ICC) were used to analyse reliability of interval scale parameters. A 2 way mixed model was chosen as each subject was assessed by the same set of raters who have been purposely and not randomly selected [27,28]. 0.6 ICC has previously been indicated as signifying a useful [29] and good [30] level of reliability. ICC of between 0.40 and 0.59 has been defined as signifying a fair level of reliability [30]. This interpretation was used to benchmark inter rater reliability interval level data.

Cohen's kappa was used to analyse reliability of categorical data extracted from Q2 (Perceptual Tonicity – amount of pressure used to produce tracheosophageal voice) on the SToPS scale. In order to examine inter rater reliability for Q2, data were recoded into 4 categories as follows:

- Hypotonic 5, 4, 3, 2, 1 was recoded as 1
- Tonic 0 was recoded as 2
- Hypertonic 5, 4, 3, 2, 1 was recoded as 3
- Stenosis 5 was recoded as 4

Reliability was calculated using kappa to see whether raters agreed 2x2

- Rater 1x Rater 2
- Rater 1x Rater 3
- Rater 2 x Rater 3

Analysis was conducted for reliability by prosthesis type by splitting data by prosthesis type and then using cross tabs for kappa analysis by rater 2x2.

The Landis and Koch [31] classification of 0.61 as a good level of reliability, 0.41-0.60 as moderate reliability 0.21-0.40 as fair reliability and 0.00-0.20 as slight reliability was used to analyse categorical level data.

Results

Reliability of interval scale data

The majority of parameters (Q1,Q3,Q5,Q7,Q8,Q9,Q13,Q14) reached an ICC of 0.60 indicating a good level of reliability (table 2). Parameters, which did not reach an ICC of 0.60 are highlighted in greyscale. While reliability was not observed on Q4 ("Wetness" of voice quality) for the Blom Singer Low pressure voice prosthesis nor on Q10 (Impairment of fluency) for the Blom Singer Duckbill voice prosthesis, the ICC for both prostheses on both parameters approached good reliability. Reliability for Q11 (Impairment of articulatory precision) was fair as opposed to good except for the low-pressure prosthesis. Reliability was reached amongst raters for only three of the voice prostheses (Blom Singer Duckbill, Blom Singer Low pressure and Provox NID) but was fair for other prostheses on Q12 (Positive features of articulation – paralinguistics/diction).

Reliability of Q2 Bipolar Semantic Scale data from STOPS

Results of this analysis are outlined in table 2.1, 2.2 and 2.3

Reliability between raters was therefore only fair or slight for Q2 Tonicity across voice prostheses.

Discussion

Expert raters inter rater reliability on the SToPS

Reliability was investigated to ascertain whether there was a good level of agreement among all three raters when using the SToPS to perceptually judge voice. Parameters with poor reliability were Q2 – Perceptual Voice Tonicity, Q11-Impairment of articulatory precision and Q12 – Positive features of articulation (paralinguistics/diction). Q2 relates to tonicity of the vibratory segment or the amount of pressure used to produce alaryngeal voice. Clinically, a patient with a tonic voice will be able to produce fluent sound of adequate intensity without effort. A tonic voice has been defined as the ability to sustain /a:/ for 10 seconds and produce 10-15 syllables per breath [32] or to sustain /a:/ 8 seconds and count from 1-15 on one breath [33]. A previous study [1] examined inter rater agreement between 12 Speech and Language Therapists and 10 ENT surgeons for Q2 of the STOPS. While inter rater agreement was only moderate for the raters as a whole, it was good for the subgroup of Speech and Language Therapists with specific voice experience. Inter rater reliability was poor for three expert Speech and Language Therapist raters in this study, each of whom had demonstrated a strong understanding of tone within training sessions. The experience of Speech and Language Therapists in this study was primarily in head and neck cancer rather than specifically with laryngeal voice. This factor may account for the superior agreement achieved on Q2 in a previous study [1] However, the statistical methodology which involved recoding data from Q2 from an 11 point equally appearing interval scale into a four point categorical scale analysed with Kappa may have been a further factor in the poor reliability found in this study. Recoding data in this manner changes tonicity from a continuum to a categorical scale and thus may alter analysis. The use of Cohen's Kappa for analysis is based on absolute agreement. In examining a parameter such as tonicity, it may not be possible to attain absolute agreement within hypertonic and hypotonic aspects of the continuum. Both hypertonicity and hypotonicity contain a spectrum of variety.

Similarly the complexity of the scale used to measure Q2 may have influenced levels of reliability achieved.

Q11- Impairment of articulatory precision demonstrated fair rater reliability only. This parameter measures the degree of the lack of precision or "slurring" in speech. Lack of articulatory precision can be influenced by a number of factors including fatigue and sometimes accent. During training of expert raters, Q11 was not identified as one that needed further clarification. However, as the experience of the expert raters involved in this study was predominantly with head and neck cancer rather than with voice, it is possible that they were less familiar with the defined baseline, which used the Vocal Profile Analysis scale as a reference. This factor may have accounted for the fair rater reliability on this parameter. The final parameter to demonstrate fair rater reliability was Q12 Positive features of articulation (paralinguistics/diction). Positive features of articulation refer to diction, intonation or pause features that have an overall positive effect but are not part of the voice signal. Similarly to Q11, Q12 was not identified during training as one that required further definition. Fair rater reliability on this parameter and on Q11 may simply reflect the difficulties of assessing articulation and diction in laryngectomy patients, who present with an underlying disordered voice.

This study examined the reliability of the SToPS across a range of voice prostheses as part of the preparatory work for a later study examining differences between prostheses in terms of voice quality. Some voice prostheses notably differed in levels of reliability achieved on parameters 4, 10, 11 and 12 of the STOPS. The attributes of different types of prostheses may affect tracheoesophageal voice and therefore results of auditory perceptual analysis. This is an area that may warrant further research.

Measurement of reliability

This statistical methods used to analyse reliability in this study correspond with those conventionally used for measurement of categorical data (Cohen's kappa) [34]

[22] and interval data (ICC) [27] [22]. As a previous study [1] [24] utilised weighted kappa to evaluate reliability on all parameters of the STOPs, a possible limitation of this study was the use of ICCs rather than kappa to measure interval data. ICCs have been used extensively to measure reliability of pathological voice quality and for this reason were utilised in this study. However, the use of ICC is largely based on a framework of psychological testing. This framework substitutes listeners for test items and voices for test subjects and implies that a new set of raters would produce the same mean ratings for the same test voices. [35]. This approach has been challenged as neither representing patterns of reliability nor overall agreement for specific voice samples [35]. The alternative to ICC is weighted kappa. Weighted kappa addresses the issue of Cohen's kappa failing to take into account the degree of disagreement between raters by enabling greater weight to be assigned to some rater disagreements than others [36] However, kappa has been criticised as less informative when used with more than 2 raters and analysing exact agreement without accounting for "close" agreement [22]. In addition, with the use of kappa, variance of subjects may be an issue, as a homogenous group of subjects is more likely to show a high percentage of agreement, rather than a true reflection of reliability [22]. The lack of consensus and limited evidence regarding the optimal methodology to measure rater reliability in perceptual evaluation of both laryngeal and alaryngeal voice supports the need for further research in this area.

Conclusions

This study investigated inter rater reliability of the Sunderland Tracheosophageal Voice Perceptual Scale. The findings presented in this study supports the SToPS as a reliable tool for the auditory perceptual rating of alaryngeal voice. However, it is acknowledged that further research may be required to improve levels of agreement for parameters related to tonicity, articulatory precision and positive features of articulation.

References

1. Hurren A, Hildreth A, Carding P. Can we perceptually rate alaryngeal voice? Developing the Sunderland Tracheoesophageal Voice Perceptual Scale. Clinical otolaryngology : official journal of ENT-UK ; official journal of Netherlands Society for Oto-Rhino-Laryngology & Cervico-Facial Surgery 2009; 34:533-538.
2. Kazi R, Nutting C, Evans PR, Harrington K. A short perspective on the surgical restoration of alaryngeal speech. Southern medical journal 2009; 102.
3. Hancock K, Ward E, Lawson N, van As-Brooks CJ. A prospective, randomized comparative study of patient perceptions and preferences of two types of indwelling voice prostheses. Int J Lang Commun Disord 2012; 47:300-309.
4. Singer M, Blom E. An endoscopic technique for restoration of voice after laryngectomy. The Annals of otology, rhinology, and laryngology 1980; 89:529-533.
5. Blom E, Singer M. Surgical-Prosthetic Approaches for Post-laryngectomy Voice Restoration *Laryngectomy Rehabilitation*. Houston: College Hill Press, 1979.
6. Carding P, Carlson E, Epstein R, Mathieson L, Shewell C. Formal perceptual evaluation of voice quality in the United Kingdom. Logoped Phoniatr Vocol 2000; 25:133-138.
7. Doyle P, Eadie T. The perceptual nature of alaryngeal voice and speech. In: Doyle P, Keith R, eds. *Contemporary Considerations In The Treatment And Rehabilitation Of Head And Neck Cancer: Voice, Speech, And Swallowing*. Austin, Texas: Pro Ed, 2005:113-139.
8. Wilson D. Voice problems of children. Baltimore: Williams and Wilkins, 1987.
9. Laver J. The phonetic description of voice quality. Cambridge: Cambridge University Press, 1980.
10. Hirano M. Clinical Examination of Voice New York: Springer Verlag, 1981.
11. Kempster GB, Gerratt BR, Verdolini Abbott K, Barkmeier-Kraemer J, Hillman RE. Consensus auditory-perceptual evaluation of voice: development of a standardized clinical protocol. Am J Speech Lang Pathol 2009; 18:124-132.
12. 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.
13. 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; 13:508-517.
14. Omori K, Kojima H. Neoglottic vibration in tracheoesophageal shunt phonation. European archives of oto-rhino-laryngology : official journal of the European Federation of Oto-Rhino-Laryngological Societies (EUFOS) : affiliated with the German Society for Oto-Rhino-Laryngology - Head and Neck Surgery 1999; 256:501-505.

15. Kazi R, Singh A, Mullan Get al. Can objective parameters derived from videofluroscopic assessment of post laryngectomy valved speech replace current subjective measures? An e tool based analysis. Clinical Otolaryngology 2006; 31:518.

16. Kazi R, Singh A, Venkitaraman R, Sayed S, RhysEvans P, Harrington K. Is electroglottography-based videostroboscopic assessment of post-laryngectomy prosthetic speech useful? Journal of Cancer Research and Therapeutics 2009; 5:85-92.

17. Kazi R, Kiverniti E, Prasad Vet al. Multidimensional assessment of female tracheoesophageal prosthetic speech. Clinical otolaryngology 2006; 31:511-517.

18. Schindler A, Mozzanica F, Ginocchio D, Invernizzi A, Peri A, Ottaviani F. Voice-related quality of life in patients after total and partial laryngectomy. Auris Nasus Larynx 2012; 39:77-83.

19. van As CJ, Hilgers FJ, Verdonck-de Leeuw IM, Koopmans-van Beinum F. Acoustical analysis and perceptual evaluation of tracheoesophageal prosthetic voice. Journal of voice 1998; 12:239.

20. Finizia C, Dotevall H, Lundstrom E, Lindstrom J. Acoustic and perceptual evaluation of voice and speech quality: a study of patients with laryngeal cancer treated with laryngectomy vs irradiation. Archives of otolaryngology--head & neck surgery 1999; 125:157-163.

21. vanAs C. Tracheoesophageal speech: A multidimensional assessment of voice quality. *Institute of Phonetic Sciences*. Amsterdam: University of Amsterdam, 2001:199.

22. Portney L, Watkins M. Foundations of Clinical Research - Applications to practice. New Jersey: Pearson Education International, 2009.

23. Lachin JM. The role of measurement reliability in clinical trials. Clinical trials (London, England) 2004; 1:553-566.

24. Hurren A. The development of a new rating scale for the perceptual assessment of tracheoesophageal voice quality outcome following total laryngectomy *Institute of Health and Society*: Univeristy of Newcastle, 2014:307.

25. Perry A. Vocal rehabilitation after total laryngectomy *Leicester School of Speech Pathology*. Leicester: De Montfort University, 1989:176.

26. Fairbanks D. Voice and Articulation Drill book. New York: Harper & Brothers, 1960.

27. Shrout P, Fleiss J. Intraclass correlation: Uses in assessing rater reliability. Psychological bulletin 1979; 86:420-428.

28. McGraw K, Wong S. Forming inferences about some intraclass correlation coefficients. . Psychol Methods 1996; 1:30-46.

29. Chinn S. Statistics in respiratory medicine. 2. Repeatability and method comparison. Thorax 1991; 46:454-456.

30. Cicchetti D. Guidelines, Criteria, and Rules of Thumb for Evaluating Normed and Standardized Assessment Instrument in Psychology. 1994.

31. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977; 33:159-174.

32. Lewin JS, Baugh RF, Baker SR. An objective method for prediction of tracheoesophageal speech production. J Speech Hear Disord 1987; 52:212-217.

33. Blom ED, Singer MI, Hamaker R. Tracheoesophageal voice restoration following total laryngectomy. Singular Pub. Group, 1998.

34. Cohen J. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement 1960; 20:37-46.

35. Gerratt BR, Kreiman J. Theoretical and methodological development in the study of pathological voice quality. Journal of Phonetics 2000; 28:335-342.

36. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychological bulletin 1968; 70:213-220.