The development of a new rating scale for the perceptual assessment of tracheoesophageal voice quality outcome following total laryngectomy

Anne Hurren

Thesis submitted in fulfillment of the requirements of the regulations for the degree of Doctor of Philosophy

Institute of Health and Society

April 2014

Abstract

Perceptual assessment of voice in people with surgical voice restoration (SVR) is essential to evaluate surgical and other interventions aimed at delivering optimal voice quality. Currently there are no tools to measure this that do not have issues of validity and reliability.

This work describes the development and trialling of investigatory versions of three scales to address this situation: a) the Sunderland Tracheoesophageal Perceptual Scale (SToPS) for professional raters, b) the Naïve Rater Scale for non-specialist raters and c) the Patient and Carer Scale.

In the final testing of the pilot version 55 speakers using tracheoesophageal voice were evaluated by twelve Speech and Language Therapists (SLT's) and ten Ear, Nose and Throat (ENT) surgeons, divided into experienced or not at assessing voice.

Ten naïve raters assessed the voice stimuli within a test-retest design. Forty tracheoesophageal speakers and thirty-seven carers attended an interview to rate their own or their relative's voice. Inter rater agreement was then calculated between SLT, ENT, naïve, patient and carer groups with weighted kappa co-efficients

Strength of agreement values (Landis and Koch 1977) were compared to profession and expertise. Expert SLT's achieved "good" agreement for nine of fourteen parameters. Naïve judges attained "good" levels of inter and intra-rater agreement for the parameters Overall Grade and Social Acceptability. The greatest inter group consensus was for patients and carers, with "good" agreement for Intelligibility, Volume and Wetness. The

only other “good” agreement was between naïve/ENT and naïve/ SLT groups for Overall Grade.

The scales are ready for clinical use with the proviso that future work will determine whether it is possible to enhance agreement so less experienced judges can achieve “good” levels of agreement for more parameters and examine which perceptual parameters might be more prominent or vital for outcomes for different groups.

Dedicated to my mother Irene Powell who in 1977 saved a magazine article about Speech and Language Therapy and suggested it may be a good career path to pursue.

Acknowledgements

I would like to express my gratitude to the patients and their families/friends as without them this study would not have been possible. Also to the ten members of the local community who willingly volunteered to judge the voices.

I am extremely grateful to the following Speech and Language Therapists and ENT surgeons who freely gave their time to attend the training course and rate the voices, Puveendran Arullendran , Lynn Bolden, David Brooker, Michael Caygill, Helen Cocks, Jayne Deakin, Catherine Farnon, Chris Hartley, Kathryn Johnson, Eimear McCrory, Emma McNeill, James O'Hara, Sarah Owen, Vin Paleri, Jo Patterson, Steven Powell, Barbara Preston, Shannon Rees, Lucy Truman, Murray Waldron. Also to the Speech and Language Therapist members of the South of England Head and Neck Cancer Special Interest Group who acted as the pilot study raters.

Many people have been involved in this thesis and its precursor stages and I would like to thank:

My supervisory team who generously gave their time and expertise with commitment, humour and enthusiasm to make this thesis possible: Professor Tony Hildreth who first encouraged me to register for a PhD and provided statistical training and guidance for the thesis. Professor Paul Carding, my main supervisor throughout the thesis who was the constant mentor and teacher from the planning stage to the final submission. Professor Nick Miller initially acted as an internal assessor providing many insights and recommendations before becoming a supervisor for the final 18 months of my registration; Nick provided ongoing assistance with statistics, structuring, content and writing to the final stage of submission.

Mike Drinnan, Suzanne Moffatt and Lindsey Pennington who were my internal assessors. Mike also provided a mock viva and statistical advice and support.

Ed Brown Consultant Clinical Scientist for his help and expertise in the voice stimuli recordings.

Natalie Eastwood in her previous role as a Speech and Language Therapy Assistant (prior to qualifying as a Speech and Language Therapist) for naïve rater and patient and carer recruitment and recording the voice stimuli.

Peter Samuel, Chris Hartley and Murray Waldron) ENT Consultant Surgeons) and Ralph Marsh (Consultant Radiologist) for sharing their expertise to further my knowledge of the anatomical and physiological aspects of tracheoesophageal voice. Chris also kindly gave much time to read the thesis in detail and provide a mock viva. Peter James, Statistician, who provided statistical expertise in the final stages of my PhD and in preparation for my viva.

Anne Brewis and Ruth Rayner for their encouragement and support in me being allocated time for my PhD in conjunction with my NHS clinical work.

Family and friends who provided so much ongoing understanding and support. The final acknowledgement is to my husband Richard as this thesis would never have been achieved without his unfailing love, support and encouragement which included him shouldering more than his fair share of domestic duties.

Financial Support

City Hospitals Sunderland NHS Foundation Trust kindly provided a Novice Researcher Award that funded the recording and editing of the voice stimuli and one day per week research time.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

DAHNO	Dataset for Head and Neck Oncology
DME	Direct magnitude estimation
EAI	Equally appearing interval
ENT	Ear, Nose and Throat
J	Jejunum graft
QoL	Quality of Life
RT	Radiotherapy
SLT	Speech and Language Therapist
SVR	Surgical Voice Restoration
TEP	Tracheoesophageal puncture
TL	Total laryngectomy
TPL	Total pharyngolaryngectomy
TSV	Tracheostoma valve
VA	Visual Analogue

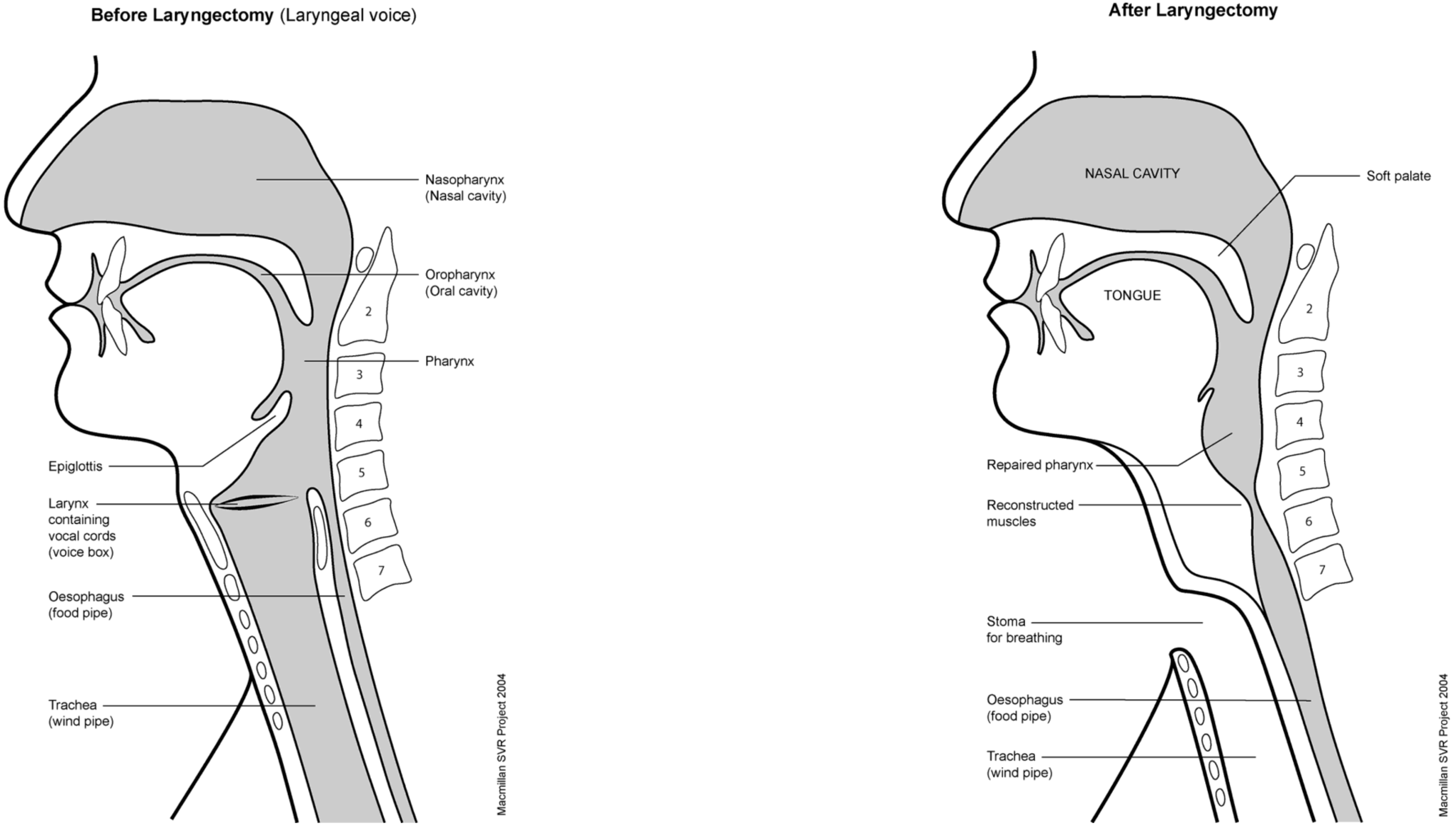
Chapter 1. Definition of Terminology and Concepts

This section outlines and defines the key concepts that form the basis of this thesis. These include the surgical resection of total laryngectomy (1.1), the surgical restoration of voice (1.2), the anatomy and physiology of alaryngeal phonation (1.3), the determinants of alaryngeal voice quality (1.4), the perceptual assessment of voice quality (1.5) and a theoretical overview of validity and reliability pertinent to the perceptual analysis of voice quality (1.6).

1.1 Total laryngectomy

Total laryngectomy refers to the surgical removal of the entire larynx; this is usually undertaken to resect a malignant tumour but may be required due to traumatic injury or a non-functional larynx as a sequel to successful (chemo) radiotherapy treatment. Singer et al (1986) described "the accepted wide field laryngectomy" as, the removal of the hyoid bone, pre-epiglottic space contents, cricoid ring, thyroid cartilage and cricoid insertions and one to four rings of the trachea. The constrictor muscles of the original pharynx are sectioned from their thyroid and cricoid cartilage insertions leaving a "gutter" configuration; this is surgically closed to form a reconstructed pharynx, known as the "neopharynx". More extensive resections involve total pharyngectomy and flap graft reconstruction as insufficient residual mucosa remains for primary closure. The tracheal remnant forms a permanent tracheostoma at the base of the anterior neck; consequently the pulmonary air stream is separated from the mouth, nose and neopharynx (Figure 1). This thesis will use the term laryngectomy and will not consider sub-total laryngectomy surgery.

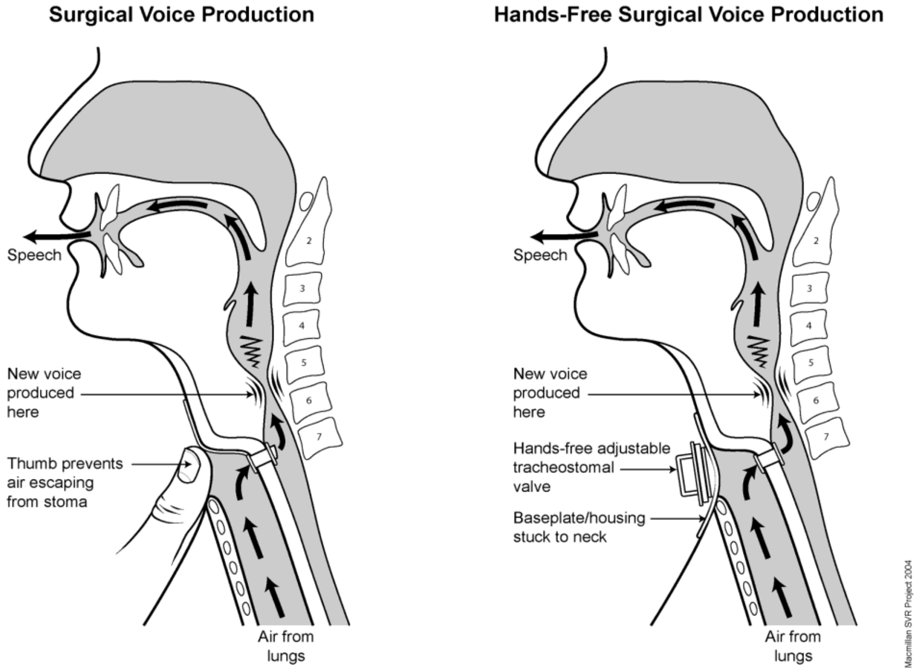
Figure 1. Pre-operative and post-laryngectomy anatomy. (courtesy of Yvonne Edels and Peter Clark, Charing Cross Hospital ©2004 Macmillan SVR project)



1.2 Surgical Voice Restoration

Surgical voice restoration (SVR) is one of five methods of alaryngeal communication after total laryngectomy; it is currently considered the “gold standard” for rehabilitation (Blom 2000). SVR involves the creation of a tracheoesophageal puncture (TEP) from the posterior tracheal wall of the tracheostoma to the anterior wall of the oesophagus (Figure 2).

Figure 2. Voice production after Surgical Voice Restoration. Courtesy of Yvonne Edels and Peter Clark, Charing Cross Hospital ©2004 Macmillan SVR project.


Surgical Voice Production
Speech
New voice produced here
Thumb prevents air escaping from stoma
Air from lungs
2
3
4
5
6
7
Hands-Free Surgical Voice Production
Speech
New voice produced here
Hands-free adjustable tracheostomal valve
Baseplate/housing stuck to neck
Air from lungs
2
3
4
5
6
7
Macmillan SVR Project 2004

A silicone voice prosthesis (valve) is fitted into the TEP to prevent the puncture from healing. The one-way valve prevents aspiration of diet into the trachea but allows the pulmonary airstream to be diverted into the neopharynx for phonation when the tracheostoma is occluded. This causes the opposing mucosal surfaces of the neoglottis tovibrateand produce a “husky or hoarse quality voice” (Blom 2000). Alaryngeal voice as a result of SVR will be referred to as “tracheoesophageal voice” in this study. Prior to the advent of SVR in 1979 (Blom 2000), alaryngeal voice could only be achieved by a method known as “oesophageal voice”. This involves transfer of air from the mouth to the upper oesophagus by one of three techniques (Edels 1983 p116), followed by controlled expulsion causing neoglottal vibration. It has been consistently demonstrated as inferior to tracheoesophageal voice (Nieboer et al 1988; Max et al 1996; Kreiman and Gerratt 1996; Finizia et al 1999a). As both methods concern neoglottal vibration, some seminal literature regarding oesophageal voice is applicable and is consequently included in this thesis.

1.3 What concepts/issues are pertinent when we consider the anatomical and physiological basis of alaryngeal voice, including tracheoesophagealvoice quality?

This section will outline the anatomical and physiological basis of tracheoesophageal phonation that causes the essential differences to voice quality in comparison to the laryngeal voice mechanism.

1.3.1 Overview of laryngeal versus alaryngeal voice

Laryngeal and alaryngeal voice involve considerably different mechanisms. Phonation in laryngeal speakers occurs due to volitional adduction of the vocal folds in coordination with an egressive pulmonary airstream. The opposing mucosal surfaces of the vocal folds are in contact and the airflow causes a periodic vibratory oscillation in the form of a regular mucosal wave. Fine motor control of the intrinsic and extrinsic laryngeal musculature permits changes in pitch, intensity and quality, thus allowing the huge potential for variability in human voice. The larynx has an acknowledged baseline for anatomy and physiology against which variance and pathology can be readily assessed (Figure 3).

Figure 3. Standard human larynx on voicing.

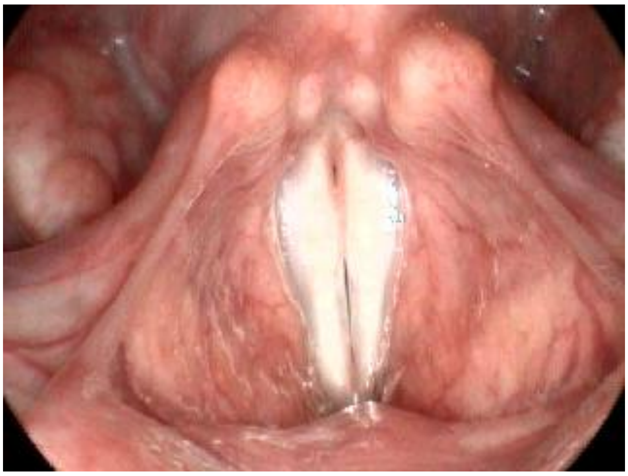
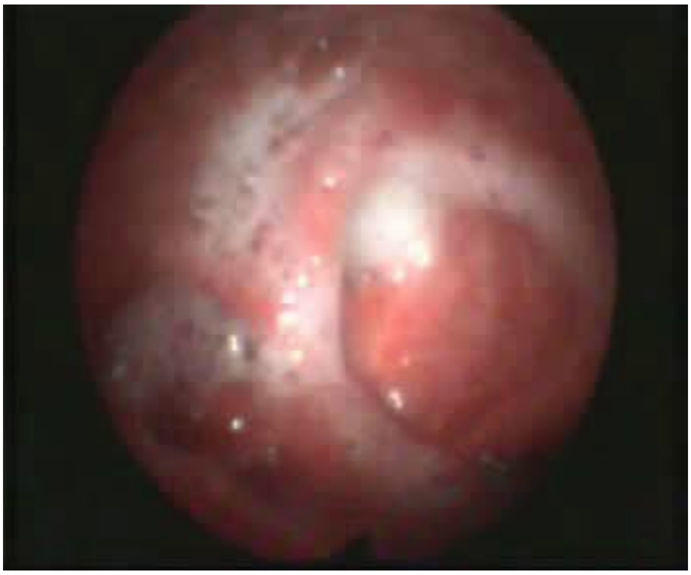


Figure 4. Fibre optic view of neoglottis on voicing.



In contrast post-laryngectomy phonation occurs from a non-standardised structure (Figure 4) due to the vibration of the neopharynx in oesophageal or tracheoesophageal voice.

Studies have consistently shown alaryngeal voice to be inferior to laryngeal voice (Green and Hults 1982; Cullinan et al 1986; Max et al 1996; Max et al 1997, Most et al 2000).These fundamental differences between laryngeal and alaryngeal voice quality have been attributed to the larger mass and reduced fine motor control of the neoglottis compared to the vocal folds (Blom et al 1995). Some auditory perceptual features are unique to alaryngeal voice and are not present in the laryngeal voice signal and will be discussed in section 1.4. These include, for example, the vibration of oesophageal secretions within the neopharynx as illustrated in Figure 4 (perceptually called "wetness") and extraneous noise from the tracheostoma during voicing (van As-Brooks et al 2005).

1.3.2 An overview of the alaryngeal phonatory mechanism

Seminal research has demonstrated that the alaryngeal phonatory source is a bar-like structure arising from the posterior neopharyngeal wall (Perry 1989; van As 2001). A variety of terms relate to this structure i.e. "neoglottis" (Omori et al 1994; van As 2001), "pharyngoesophageal segment" (Shipp 1970; Perry 1989; Koybasioglu et al 2003), "retropharyngeal bulge" (Singer et al 1986), "pseudoglottis" (van Weissenbruch et al 2000)and "retropharyngeal prominence" (Mohri et al 1994).This study will refer to the vibratory segment as the "neoglottis" and the generic reconstructed pharynx as the "neopharynx". The neoglottis is absent if patients undergo total pharyngectomy with flap reconstruction (van As-Brooks et al 2005).

The vibratory source is typically attributed to reconstructed cricopharyngeus (Diedrich and Youngstrom 1966; Simpson et al 1972), but studies have demonstrated thyropharyngeus (Perry 1989 p31; Omori, et al 1994) and the middle pharyngeal constrictor (Kirchner et al 1963) can also be involved. The neoglottis is not specifically surgically created but forms spontaneously. There is no consensus regarding how this occurs and four unsubstantiated hypotheses have been postulated to date:

a) the bulge forms from the resected thyropharyngeus remnant because it contracts anteroposteriorly due to its median posterior wall raphe and a surgical scar on the anterior wall; as cricopharyngeus has no raphe it contracts only concentrically(Omori et al 1994);

b) the anterior thyropharyngeus repair physically compresses the tissue within the tubular reconstruction with both the length of the retained muscle and the tightness of the repair determining neoglottic dimensions (Edels 2006);

c) unrepaired constrictors retract and bunch posteriorly to form a bulge (Kirchner et al 1963; Simpson et al 1972);

d) compensatory hypertrophy occurs in the newly repaired constrictors to form the bulge as normal pharyngeal peristalsis is disturbed during swallow (Perry 1989 p32).

The neopharynx is responsible for both swallowing and voice. The ultimate functional outcome is (a) a prompt and complete neoglottal bar opening for bolus passage on swallowing and (b) a neoglottal bar contraction on phonation that is neither too lax nor too tight (Perry 1989 p79; van As-Brooks et al 2005). These seminal studies also demonstrated that the reconstructed neopharynx was non-uniform, with wide inter-patient variation; consequently there is no “normal” baseline structure against which researchers can measure anatomy and physiology of voicing or swallow (Figure 5). This variability involves the physical dimensions, position in relation to cervical vertebrae, muscular tone and presence of more than one such prominence (Damste and Lerman 1966; Bentzen et al 1976; Wetmore et al 1985; Perry 1989; Isman and O’Brien 1992; Omori et al 1994; van Weissenbruch et al 2000; van As et al 2001; Lundstrom et al 2008).

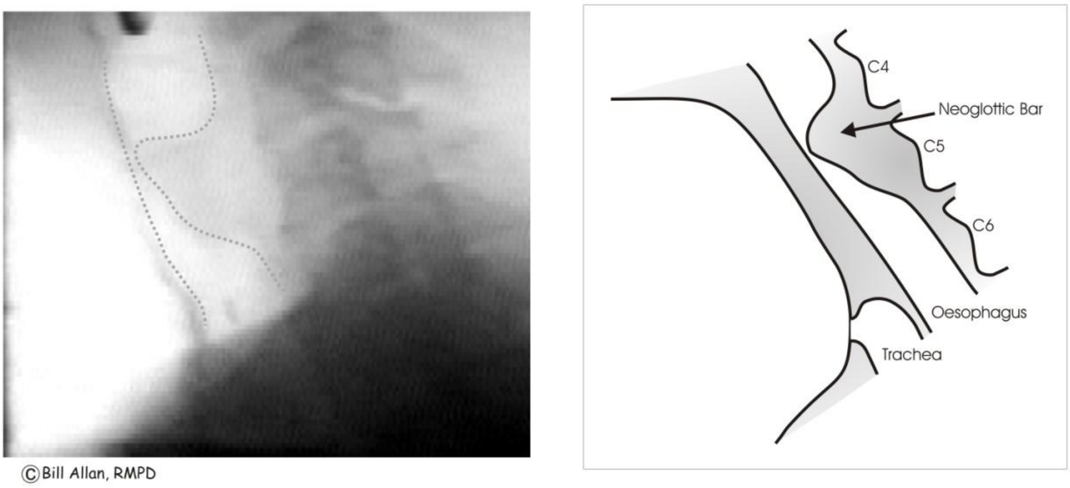
The neoglottis is a complex vertically multi-layered structure and only its most superficial layers can be seen on fibre optic examination (Meleca et al 2000). However the uppermost layer of the mucosal surfaces has been observed to approximate and vibrate in a similar manner to laryngeal phonation but with mild to moderate asynchronous vibratory patterns on stroboscopy (Dworkin et al 1999). Any regurgitation of frothy secretions from the oesophagus can also become a source of sound production independent of the neoglottis (Dworkin et al 1999).

1.4 The determinants of alaryngeal voice quality

Few studies have investigated differences *between* speakers with alaryngeal voice (Perry 1989; van As et al 2003,) and the way in which speakers' voices vary. However there is considerable evidence that primary neopharyngeal closure allows superior voice compared with flap reconstruction (Deschler et al 1994; Ahmad et al 2000; McAuliffe et al 2000; van As-Brooks et al 2005). The evidence base concerning the morphology and physiology has been achieved with a variety of instrumental assessment of neoglottal function (videofluoroscopy, intra-oesophageal manometry, endoscopy, electroglottography, electromyography, tracheal manometry, acoustic).

Although the post-surgical formation of the neoglottis and its neurophysiology are poorly understood (Doyle and Eadie 2005b p526), two important factors have been established as key indicators of optimal tracheoesophageal voicing. Firstly the neoglottis should be well defined in the form of a retropharyngeal bar (Perry 1989; van As-Brooks et al 2005) (Figure 5). This can only be fully visualised with lateral videofluoroscopic assessment.

Figure 5. A well defined neoglottal bar. This arises from the posterior neopharyngeal wall and is seen in lateral view on videofluoroscopy (left) and in diagrammatic representation (right).Courtesy of Bill Allan, Medical Physicist.

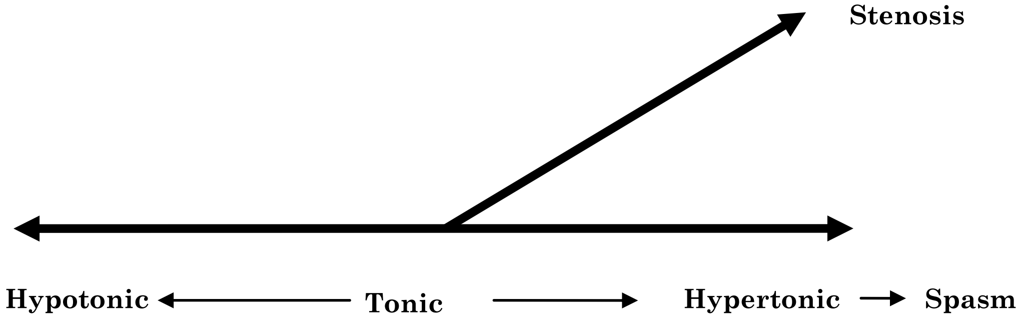


©Bill Allan, RMPD

Poor voice quality is associated with an absence of a definitive neoglottis. This occurs: a) after pharyngolaryngectomy with flap reconstruction causing "bubbly" and "whispery" voice quality (van As-Brooks et al 2005) and b) if fibrosis develops leaving just a "rigid but adynamic gullet (tubular pharynx)" (Perry 1989),associated with a "coarse whisper quality" (Singer et al 1986). The second feature that determines optimal alaryngeal voice is the muscular tone of the neoglottis (Singer et al 1986; Perry 1989; van As-Brooks et al 2005; Hurren et al 2009).

Tone was originally defined into five mutually exclusive categories of hypotonicity, hypertonicity, spasm, neutral tonicity and stenosis (Cheesman et al 1986; Perry 1989; McIvor et al 1990). However this was subsequently revised to tonicity existing along a continuum (Figure 6) (Perry 1989). This change in understanding was brought about by investigations using intraoesophageal manometry.

Figure 6. Perry's theory of tonicity.



Perry demonstrated neoglottal contraction on voicing occurred anteroposteriorly in neutral tonicity, hypertonicity and spasm. The neutral to hypertonic continuum is characterised by increasing neoglottal wall approximation resulting in a strained, effortful voice until total spasm occurs with severe sub-neoglottal ballooning and no voice. Conversely, the low tone (hypotonic) spectrum is related to neoglottal dilation and flattening on phonation. As tone decreases and antero-posterior wall separation increases there is reducing volume and increasing breathiness of the voice. This is linked to a "bubbly" or "wet" voice quality as oesophageal secretions regurgitate on phonation due to the lack of a seal from neoglottal closure (van As-Brooks et al 2005). The fifth category (stenosis) is separate to the hypo-hyper tonic continuum because tissue rigidity due to fibrosis results in no tone and no identifiable neoglottis. However fibrotic changes are not an "all or nothing" phenomenon. It would seem reasonable to hypothesise that tone and some degree of fibrosis could co-occur, but differentiating the influence of each factor would be challenging. The lack of a definitive neoglottic bar in stenosis is also likely to allow the regurgitation of secretions observed in hypotonicity but there is a lack of investigation into this category of neopharyngeal physiology to substantiate this hypothesis.

Evidence from electromyography has demonstrated that the neoglottis is a highly idiosyncratic structure and poor speakers have less control over muscle contraction of the neoglottis than better speakers (Shipp 1970): the neoglottal contraction has been attributed to a response to the passive stretch of the oesophageal wall on air bolus entry (Shipp 1970).

Although there is evidence that certain neoglottal features are linked to superior tracheoesophageal voice quality, there are no current definitive characteristics that constitute optimal anatomical and physiological baseline features of the neoglottis. This paucity of evidence is of course compounded by the lack of a valid and reliable voice quality perceptual rating scale for alaryngeal voices – and this represents the central issue of this thesis. Instrumental measures have been used but there is no accepted gold standard assessment against which voice quality or other assessments can be compared. This leaves clinicians and researchers with no baseline or set criteria against which to assess voice outcome.

1.5 The perceptual assessment of laryngeal and alaryngeal voice quality

Patients typically seek help and judge treatment efficacy on the basis of whether they perceive their voice to sound normal; consequently voice quality has been concluded to be "fundamentally perceptual in nature" (Kreiman et al 1993). Voice quality has been described as "an interaction between a voice stimulus and a listener" (Kreiman and Gerratt 1998). This means the acoustic signal does not possess a quality per se (Kreiman and Gerratt 1998). The perceptual assessment of voice relates to a rater's subjective response to the acoustic signal in a voice stimulus (Gerratt and Kreiman 2000). Voice evaluation requires (or at least challenges) the listener to separate the acoustic signal into pre-selected, perceptually putatively distinct parameters. An assessment tool must involve the design of a scale format to allow the rater to indicate their perception of a specified parameter. The majority require the rater to mark a scale point in response to a voice stimulus by judging the extent to which the voice deviates from an internalised psychoacoustic representation of the baseline. The key requisites for scales are sensitivity in differentiating between speakers and treatment effects and ease of use in clinical settings. Numerous scale designs have been used to measure voice parameters (Kreiman, Gerratt et al 1993):

1. categorical ratings; a parameter is measured as present or absent e.g. strain but without a scalar point to indicate the degree;

2. paired comparisons; two stimuli are compared as the same or different with a named parameter;

3. equally appearing interval (EAI); parameter evaluation extends from a zero baseline to a specified endpoint in intervals of one unit;

4. visual analogue (VA); a parameter is rated from zero to 100 with a non-calibrated 100 mm long line;

5. bipolar semantic scale (Osgood et al 1957); EAI or VA format may be used with opposing parameter endpoints marking each end of the scale e.g. strained/ not-strained;

6. adjectival; words but no numerical markers indicate scale points e.g. good, moderate, poor;

7. direct magnitude estimation (DME); the scale has no predetermined upper or lower limits. Raters assign points to each voice stimulus relative to a prior agreed baseline sample, (the “modulus”) which has a value of 100 e.g. half as good = 50.

Perceptual assessment can include raters listening to intermittent anchor stimuli. The theoretical benefit of using anchor stimuli is that it appears to improve both intra and inter-rater agreement. Direct magnitude estimation (DME) has been used in a small series of studies (Eadie and Doyle 2002b; Eadie and Doyle 2004; Eadie and Doyle 2005b). The lack of endpoints to the parameter continuum allows sensitivity to small changes and raters are not forced to decide between equally appearing intervals. Paired comparison has been used to allow raters to judge scale points of similarity/dissimilarity for each parameter for two voice stimuli at a time (Ward et al 2011). This format is highly likely to be sensitive to change but would be cumbersome to use in routine clinical settings.

A crucial aspect of perceptual tool design involves the specification of a baseline against which a parameter should be evaluated i.e. whether the stimulus is being assessed in relation to normal laryngeal voice or optimal tracheoesophageal voice. This facilitates consistent inter-rater application and thus reliability. An alternative baseline to normal laryngeal voice appears essential for the evaluation of alaryngeal speakers who can never achieve normal voice quality. The established practice of comparing tracheoesophageal to laryngeal voice quality has been described as "hindering the advancement of our understanding of alaryngeal voice quality" (Doyle and Eadie 2005a p115). It would seem appropriate to evaluate tracheoesophageal voice against a baseline of the most optimal tracheoesophageal voice it is feasible to achieve after laryngectomy.

Key aspects of laryngeal and alaryngeal rating tools are discussed in detail in sections 2.2 and 2.3 respectively.

1.6 A theoretical overview of validity and reliability

A fundamental issue in the design and implementation of rating scales concerns the establishment of the validity and reliability of such assessment tools. These issues of validity and reliability will therefore be crucial in this thesis as it centres on the development of a perceptual rating scale for alaryngeal voices. Validity and reliability are essentially interlinked because a scale lacking in stability and reproducibility must be deemed inherently invalid before any other factors are taken into consideration.

1.6.1 Reliability

Reliability or consistency refers to the degree of attainment of the same, or similar, results by different observers on different occasions or the same observer rating the same stimuli on a different occasion. Conversely, unreliability has been defined as the discrepancies that arise if a measurement is repeated many times (McDowell and Newell 2006 p40). There are many sources of such inconsistencies which are often referred to as measurement error. Error has been classified into two broad groups a) random error (or “noise”) and b) systematic error (or bias). Reliability theory has generally focussed on random errors leaving biases to be assessed under validity testing. Random errors have been defined as those that occur in unpredictable ways on every measurement (Schiavetti 1997) e.g. due to poorly standardised instructions (Nunally 1970), tiredness or inattention in raters (McDowell and Newell 2006 (p40) and mistakes (Kreiman and Gerratt 1998). As they can both under and over estimate agreement they have been assumed to cancel each other out if enough observations are made so the average score gives a reasonable assessment of the true score McDowell and Newell 2006 p41). In contrast systematic errors occur consistently over every repeated measurement (Schiavetti 1997). In the field of voice perception raters have been suggested to demonstrate perceptual biases (Kreiman and Gerratt 1998; Shrivastav et al 2005). These relate to the inherent challenge of perceptually partitioning continuous sound variables (such as pitch and loudness) into discrete scale points.

Indeed, in relation to this, subsequent research by Kreiman, Gerratt and co-workers has postulated an additional (third) type of error to account for inter reliability problems in voice perceptual assessment. This concerns four types of difficulty listeners experience in undertaking the rating tasks per se (Kreiman et al 2007): a) unstable internal standards of voice qualities, b) problems in isolating individual parameters in complex voice stimuli, c) the scale format and d) the magnitude of the attribute. This body of research concluded that random error only plays a small role in voice rating variance. Furthermore they criticise the theory that systematic biases are stable because expert raters judgments have been observed to "drift" in predictable ways i.e. a voice may be rated as moderately rough but on re-test several minutes later the same parameter can be judged as severe if mildly deviant voices have been assessed in the interim period (Gerratt et al 1993).

Gerratt and Kreiman conducted a series of experiments to provide evidence that the traditional two error model is not applicable to voice perception reliability (Kreiman et al 2007; Kreiman and Gerratt 2011). The crucial theoretical underpinning is that listener disagreements cannot be reduced to either random or systematic errors as the true rating for a voice stimulus is not entirely a function of the voice with the listener a "virtual acoustic analysis system". This alternative model considers differences between raters to be due to cognitive processes in mapping a complex auditory signal to a response rather than being an error per se. Kreiman and Gerratt (2011) provided an elegant analogy to explain their rationale i.e. two people can perceive a room's temperature differently, one can be hot and one can be cold. If the aim of measurement is to use the humans as a "virtual thermometer" to assess room temperature then differences in perception are classed as error and the best estimate of temperature would be to average out a number of human temperature estimates. However if the aim of measurement is to investigate how people perceive temperature then both responses are valid and the variation is not error but must be viewed as a perceptual process. Kreiman and Gerratt consider the latter aim to apply to voice perceptual analysis. Furthermore voice analysis causes listener error beyond random error and consistent systematic bias as more complex cognitive processes are involved than required to perceive temperature. Such complexity affects the raters' response and appears due to how attention is allocated to or focused on different aspects of the voice stimuli and interactions among the individual parameters (Kreiman et al 2007; Kreiman and Gerratt 2011). An additional complicating factor concerns a lack of objective criteria for voice measurement unlike temperature perception which can be compared to a reading from a thermometer; this relates to validity and will be discussed more fully in section 1.6.2.

A further core aspect of reliability testing concerns sampling procedures. These should be considered carefully and stimuli must be representative of the spread of behaviours present in the population undergoing investigation (Streiner and Norman 1995 p7). This is particularly important for the assessment of voice quality as there is more agreement about normal and severe voice qualities than for moderate (Kreiman et al 1993) i.e. overinflated levels of agreement between/within raters could occur simply due to over-representation of mild or severe stimuli. Reliability increases when true variation between the item being observed increases and when error variation is small (McDowell and Newell 2006 p40). Studies should therefore investigate an extremely heterogeneous sample because statistical calculations concern the ratio of variability between subjects to total variability.

Scale design can also influence reliability. There is for instance evidence people cannot discriminate beyond 7 scale points (Streiner and Norman 1995 p35). A further consideration is that points on equally appearing interval (EAI) scales or adjectival scales may not be equidistant psychoacoustically. Although high reliability co-efficients may be obtained this only implies judges rate in a parallel fashion and does not prove scale values have the same meaning (Kreiman et al 1993). Short EAI scales inherently allow a higher rate of chance agreement and selecting appropriate statistics that control for this is crucial.

A comprehensive summary of the laryngeal and tracheoesophageal voice quality perceptual studies, including the method of statistics selected to calculate reliability, is detailed in sections 2.2 and 2.3. Four major types of reliability calculations have generally been selected in these studies. All have been criticised for failing to account for chance agreement and have other limitations as outlined below:

a) percentage rater agreement to within plus or minus one scale point (Cullinan et al 1963; Kreiman et al 1993) which can inflate agreement even on scales which are 7points in length.

b) Pearson's correlation co-efficients which require interpretation with caution" when used in reliability studies as they quantify the association between two ratings i.e. they indicate how accurately one rating can be predicted from another not agreement per se (McDowell and Newell p36; Streiner and Norman p115). For example if one rater consistently rated one scale point higher than another judge the correlation would be perfect but agreement would be zero. Furthermore correlations are influenced by the range of the scale i.e. wider ranges increase correlations although agreement remains the same (Streiner and Norman 1995 p36). Such ability for Pearson's to exaggerate reliability scores has led to other statistical methods being preferred (Bartko 1991).

c) intraclass correlations have been criticised as they provide an average rating whereas clinical voice evaluations are more interested in individual rater behaviours (Kreiman et al 1993).

d) Cronbach's alpha is commonly used to calculate internal consistency of scale items (Streiner and Norman 1995 p64) but cannot represent patterns of agreement among raters nor indicate agreement for specific voice examples (Kreiman and Gerratt 1998; Kreiman and Gerratt 2000).

The use of the weighted Kappa co-efficient (Cohen 1968) is considered to be the optimal statistical calculation with nominal or ordinal rating scales if there is a high likelihood of agreement by chance alone (Streiner and Norman 1995 p116). It calculates the extent of agreement expected by chance and removes this from the estimation (McDowell and Newell 2006 p41); the quadratic forma allows credit for partial agreement. However quadratic weighted kappas provide statistically equivalent co-efficients to one sub-type of intraclass correlation (Streiner and Norman 1995 p126). This is a crucial consideration in relation to the blanket criticism of intraclass correlations outlined in c) above.

Another key factor in statistical testing concerns the criterion for acceptable reliability of a scale. This is expressed as a ratio of the variability between individuals to the total variability in the scores, in the form of a number from 0 to 1. Zero relates to no reliability and 1 is perfect reliability (Streiner and Norman 1995 p7). Varying authors have made different recommendations regarding the minimum acceptability but such suggested values have been criticised for being "at best, expressions of opinion" (McDowell and Newell 2006 p45). Acceptable values for Pearson's co-efficient has been suggested as those in excess of 0.85 (McDowell and Newell 2006 p45). The generally accepted guidelines for determining strength of agreement in relation to kappa co-efficient values are those specified by Landis and Koch (1977) i.e. <0.20 is "poor", 0.21-0.40 is "fair", 0.41-0.60 is "moderate", 0.61-0.80 is "good" and 0.81-1.0 is "very good". However the purpose of the measure influences the standard required; whilst Streiner and Norman (1995 p7) specify scale stability of 0.5 or above may be acceptable in some instances, clearly when crucial decisions ensue from the results much higher, even 100% agreement is demanded (e.g. has –does not have cancer, will-will not die).

Reliability only relates to whether a parameter is being measured in an acceptably reproducible fashion and does not consider whether it measures what was intended. Consequently a scale must also undergo validity testing.

1.6.2 Validity

The traditional definition stated a scale to be valid if it measures what is intended. However this classification was subsequently redefined to describe validity as the range of interpretations that can be appropriately placed on a measurement score (McDowell and Newell 2006 p 30). This new definition was considered to be significant because validity is then not "a property of the measurement" but the interpretation placed on the results (McDowell and Newell 2006 p30). Prior to the 1970's, the accepted evaluation of validity concerned "the 3 C's" (Landy 1986) i.e. Content, Criterion, and Construct validity. These were regarded as three separate attributes of measurement to be independently established. However newer trends in validity assessment no longer consider these to be disparate characteristics (Streiner and Norman 1995 p145) .The establishment of validity must include a critical evaluation of evidence to support the scale (Streiner and Norman 1995 p146) and the ultimate aim of validity testing is to determine a tool's "inferentiality" (Streiner and Norman p147) i.e. to ascertain what we can conclude from the measure. This issue in relation to the development of the scale in this thesis is broached in 6.2.2.

A key aspect of the inferentiality of a scale is content validity. This is a judgement as to whether the scale looks reasonable and samples all the relevant or important content or domains that theoretical and expert opinion in the field deem necessary in order to describe a phenomenon. Such assessment of the appropriacy of scale items is a subjective expert judgement and rarely uses empirical approaches (Streiner and Norman 1995 p5). Careful planning during a scale's development facilitates this aspect of validity (Nunally 1970; Cronbach 1990). This is often in the form of a literature review to facilitate evidence based selection of items that are characteristic of the attribute to be measured (Streiner and Norman p19). A subsequent review of a proposed scale by a panel of three to ten recognised experts has been suggested as the minimum prerequisite for a scale to be accepted (Streiner and Norman 1995 p5). However this method of content validity referred to as "validity by assumption" (Guildford 1954) is acknowledged to include subjectivity and the potential for bias with expert opinion (Streiner and Norman 1995 p20; Schiavetti 1997). Peer judgement is insufficiently robust to ensure a test actually measures what it anticipates to measure. Consequently two further methods for testing validity after content validity are essential (Streiner and Norman 1995 p8):

a) Comparison with either a similar, pre-existing scale or to some other form of gold standard to ascertain correlation. This is referred to in a variety of terms: convergent validity, criterion validity and concurrent validity. The authors highlight the risk of creating a circular argument where neither tool measures what is intended in spite of demonstrable agreement.

b) If a similar tool does not exist to enable comparison there must be a clear justification for development with verifiable construct validity. Demonstrating construct validity involves hypothesising how a measure should behave, what are the variables/parameters involved in its manifestation and variability and then gathering evidence to support the hypotheses (Schiavetti 1997). Consequently it is a complex process (Nunally 1970).

This thesis concerns tracheoesophageal voice quality but there is no basis for defining “the correct quality judgement for a given stimulus” (Gerratt and Kreiman 2000) in voice perceptual assessment. Firstly there is no universally accepted criterion validity to act as a gold standard as outlined in section 1.4. Secondly voice parameters are hypothetical constructs. The theoretical perspective of Kreiman and Gerratt (2011) with respect to the second issue was outlined in section 1.6.1 and specifies that the “true rating” for each voice is not entirely a function of the voice per se but about the mapping of the signal to a psychoacoustic response from the listener’s perspective. This relies on the ability of listeners to agree on borders between qualities and, in the case of numerical scales, assign a value to the voice attribute that is perceived. The test to investigate such abstract variables that cannot be directly observed is referred to as “construct validity” (Streiner and Norman 1995 p151).

In terms of this thesis the aim is to develop new tracheoesophageal voice rating scales as no scale exists that adequately encapsulates the key perceptual constructs of tracheoesophageal voice. Unlike criterion validity there is no one accepted experimental design or statistic to establish construct validity. For this reason construct validation has been described as part science and largely art form (McDowell and Newell 1996 p36). Validity cannot be proved definitively; instead it is a continuous process in which testing contributes to our understanding of the construct, which in turn enables new predictions to be proposed and tested. This type of validity does not differ conceptually to criterion and content validity; it is the basic meaning of validity and all validity has at its base some form of construct validity (Streiner and Norman 1995 p153).

1.7 Summary

This introductory chapter has outlined the concepts of total laryngectomy and Surgical Voice Restoration. The differences between laryngeal and alaryngeal voice were highlighted and included the anatomical and physiological determinants that constitute the phonatory mechanism following total laryngectomy. The perceptual analysis of voice quality was discussed including an overview of key scale format and design features. The final sections discussed reliability and validity aspects of scale development with reference to some of the theoretical aspects of laryngeal voice perceptual analysis. Before commencing the design of new tracheoesophageal perceptual scales it is important to review the relevance of this method of analysing tracheoesophageal voice. There is also a requirement to examine and critique the tools that have been developed to date for the perceptual analysis of both laryngeal and tracheoesophageal voice. The issue of rater perspective is also a key area to examine as multiple judge types will be recruited for this thesis. All these issues summarised above will be the subject of Chapter 2.

Chapter 2. Literature Review

This chapter involves a review of the published literature. The first section demonstrates the necessity of scales to measure alaryngeal voice (2.1). The field of laryngeal perceptual analysis is then discussed in relation to its application to tracheoesophageal voice (2.2) followed by a critique of tracheoesophageal scales utilised in studies to date (2.3). The subsequent section (2.4) outlines the advantages and challenges of assessing tracheoesophageal voice with different types of raters. The final sections summarise the issues and gaps in research to date (2.5) and conclude with the research aims that will be addressed in this thesis (2.6).

2.1 Why are perceptual voice quality rating scales necessary in SVR?

Perceptual rating scales in laryngeal voice are well established for the purpose of measuring surgical and therapy outcomes and change in voice quality over time (Carding et al 2000; Carding et al 2009; Oates 2009).These key issues are equally relevant in tracheoesophageal phonation. However outcome issues are more complex; neopharyngeal structures are responsible for voice and swallow and surgeons must balance these two key functions with cancer clearance and other morbidity and mortality considerations when planning surgery. Surgical and other types of management options are debated in the literature without consensus regarding which method optimizes voice and minimizes complications. These include surgical management of the pharyngeal constrictors (Simpson et al 1972; Singer et al 1986; Mahieu et al 1987; Olson and Callaway 1990; Clevens et al 1993; Wang et al 1997; Deschler et al 2000; Madeen et al 2011), determining the degree of tissue to permit primary closure (Hui et al 1996; Iwai et al 2003), myotomy (Singer and Blom 1981; Chodosh et al 1984; Blom et al 1986; Mahieu et al 1987; Milford et al 1988; Perry 1989 p.154; Olson and Callaway 1990; Op de Coul et al 2003; Albirmawy et al 2009), neurectomy

(Singer et al 1986; van Weissenbruch et al 2000; Koybasioglu et al 2003), reconstruction method for laryngopharyngectomy (Cumberworth et al 1992; Anthony et al 1994; Hilgers et al 1996; Jones et al 1996; Deschler et al 1998; Ahmad et al 2000; Iwai et al 2002; Ward et al 2002; Robb and Lewin 2003; Alam et al 2008; Murray et al 2008; Patel et al 2009; Yang et al 2011; Ho et al 2012), botulinum toxin regimes (Hoffman et al 1997; Terrell et al 1995; Hoffman and McCulloch 1998; Zormeier et al 1999; Lewin et al 2001; Hamaker and Blom 2003; Ramachandran et al 2003), and comparisons of voice prosthesis type (Heaton et al 1996; Delsupehe et al 1998; Blom 2003; Brown et al 2003; Vlantis et al 2003; Issing et al 2001; Ward et al 2011).

The requirement for research to develop techniques that "allow fine adjustments of neopharyngeal wall tension critical to effective sound production and vocal pitch" was highlighted with the initiation of SVR (Singer et al 1986). However only one study has investigated voice outcome in relation to surgery or reconstruction type with a scale that has some evidence of reliability (van As et al 2001 p44) resulting in a lack of robust evidence regarding best management.

A clinically relevant perceptual voice quality scale, with established reliability and validity would allow further research into surgical management that offers optimal voice quality in conjunction with the crucial surgical decisions of optimal survival and minimal morbidity e.g. salivary fistula prevention.SVR success rates vary and are difficult to define (Hillman et al 2005). Success rates have been described in relation to communication ability/voice quality (Donegan et al 1981;Blom et al 1986; Hilgers and Balm 1993; Ferrer-Ramirez et al 2001; Hotz et al 2002; Brown et al 2003), complication rates (Garth et al 1991; Camilleri et al 1992; de Raucourt et al 1998; Op de Coul et al 2000; Ferrer-Ramirez et al 2001; Karlen and Maisel 2001), percentage of time communicating with tracheoesophageal voice (Garth et al 1991; de Raucourt et al 1998; Ahmad et al 2000; Brown et al 2003), percentage of patients requiring permanent removal of the voice prosthesis (de Raucourt eta l 1998;Op de Coul et al 2000; Ferrer-Ramirez et al 2001; Chone et al 2005) and life span of the voice prosthesis (Hilgers and Balm 1993; Op de Coul et al 2000). Many perceptual scales are cited in the literature, but methodological flaws are common. This important issue will be addressed in detail in section 2.3.

2.2 Laryngeal voice quality research and its application to tracheoesophageal voice

Although some essential differences exist between laryngeal and tracheoesophageal voice, sufficient commonality of themes justifies a summary of key findings. There is a larger body of more robust publications for laryngeal voice quality scales in comparison to that of tracheoesophageal voice perception. The aim of examining this existing theoretical base is to facilitate the development of the most reliable and valid tool to assess tracheoesophageal voice from an appreciation of the limitations and strengths of previous mutually compatible research.

The theoretical underpinning of laryngeal voice perception is “fragmented” into diverse fields for example singing, computerised voice recognition, psycholinguistics and psychology, without inter-disciplinary collaboration (Gerratt and Kreiman 2000). An extensive literature review (Kreiman et al 1993) concluded it is difficult to summarise the field because study comparisons are hindered by variations in rater type/ number, training protocols, scale format and selection of reliability statistics. Methodological flaws were also prevalent e.g. reliability issues were ignored or addressed with inappropriate statistics. Although scales appear simple to use, a body of research details reservations about their validity. The key aspects of these factors are summarised below.

2.2.1 Baseline anchor point

The most common scale format in studies to date is equally appearing interval (EAI) scales that measure the psycho-acoustically perceived distance of a pathological voice from a baseline of normal voice quality (Hammarberg et al 1980; Laver 1980; Hammarberg et al 1986; Wilson 1987; Hirano 1989). However no accepted benchmarks perceptually define normal voice so the division between normal and abnormal voice is unclear (Carding et al 2000). Judging against such a subjective baseline causes difficulty in definition as there is enormous differentiation within voices that can be considered as normal; cultural aspects may also affect norms (Fex 1992). Conversely, other authors have suggested all listeners are likely to have similar, stable internal standards for normal voice (Kreiman et al 1992). These issues of anchor points are perhaps even more relevant when intuitive baselines do not exist in tracheoesophageal voice (see section 2.3 below).

2.2.2 Parameter selection and definition

A further difficulty in clarifying the evidence base is because perceptual scales to date have included a “plethora of vocabularies which are inconsistently used” (Gelfer 1988). Fifty seven scales were found in the USA alone (Kreiman et al 1993). The lack of consensus of terms makes reliable description of voice quality difficult (Carding et al 2000) yet producing “unambiguous descriptions for voice qualities” is an enormous challenge (Fex 1992). Parameter definition is crucial to consistent rating and parameters must also possess a “perceptual reality” for the listener (Kreiman et al 1993).

Some scales measure voice quality features at a laryngeal vibratory level e.g. GRBAS (Hirano 1989), whereas others e.g. Vocal Profile Analysis (Laver 1980) incorporate supralaryngeal features including resonance and lip, jaw and tongue movement patterns. A further fundamental factor concerns whether parameters are treated as individual features of voice or as "subordinate aspects" of "super-ordinate quality" as complex multidimensional structures (Gerratt and Kreiman 2000). Parameters that describe multi-dimensional features have been referred to as global parameters with their more discrete sub-divisions described as uni-dimensional parameters (Doyle and Eadie 2005a p13). The most common such Global parameter for laryngeal voice is "Overall Grade". "Pleasantness" has been included in studies of normal and dysphonic voice (Eadie and Doyle 2002b; Eadie and Doyle 2005c) but appears unlikely to measure purely laryngeal level features; furthermore the lack of a discernible baseline prevents it differentiating normal from abnormal voice quality and brings into question its clinical relevance. Clinically useful scales should focus on the key perceptual dimensions with justification of parameter selection (Kreiman and Gerratt 1998).Parameter selection may be even more problematic for tracheoesophageal voice because normal descriptors are not appropriate. Therefore clear definitions are essential - see section 2.3 below.

2.2.3 Reliability

Most reliability studies have involved investigation of the GRBAS scale (Dejonckere et al 1993; de Bodt et al 1997; Dejonckere et al 1998; Millet and Dejonckere 1998; Webbet al 2004), although other scales do exist e.g. Buffalo (Wilson 1987), Vocal Profile Analysis (Laver 1980), and CAPE-V (Kempster et al 2009). A common finding is that "Overall Grade" or "Overall Severity" are the most reliably judged (Dejonckere et al 1993; de Bodt et al 1997; Dejonckere et al 1998; Millet and Dejonckere 1998; Munoz et al 2002; Webb et al 2004; Zraick et al 2011). Other well known terms such as "Roughness" and "Breathiness" also seem to have good reliability within and across judges (Millet and Dejonckere 1998; Webb et al 2004; Lee et al 2005; Zraick et al 2011).

Different scale formats have been suggested to have superior reliability for Overall Grade. Eadie and Doyle (2002b) concluded this parameter could only be reliably measured with direct magnitude estimation (DME) rather than an equally appearing interval (EAI) scale. DME with an anchor stimulus was reported to increase reliability by reducing "variable internal representations" (Eadie and Doyle 2002b). However, only unanchored EAI scales were compared to DME and it is not possible to distinguish the effect of the single anchor modulus from the scale format. DME studies to date have included only naïve judges and the application of this research to professional judges is not clear.

Visual analogue (VA) scales have demonstrated good reliability for features of normal voice in professional users (Bele 2005), for the GRBAS (Dejonkereet al 1998) and CAPE-V (Kempster et al 2009). However investigations to compare EAI with VA scales have no clear conclusion. Superior inter-rater reliability has been demonstrated with the EAI (Wuyts et al 1999) but equally raters were observed to perform similarly with both scale types (Kreiman et al 1993). Although significant rater drift with the EAI scale was reported in the latter study this could be an artefact of the longer 1-7 scale selected for this investigation in comparison to the 0-3 used in for the GRBAS scale.

Certain parameters have been reported to be easier to evaluate, suggesting a parameter-specific aspect to reliability (Bele 2005; Webb 2005). Other methodological variations have also been attributed to reliability. Bele (2005) suggested increasing rater numbers leads to superior reliability whereas increasing the number of parameters per stimuli reduced inter-rater reliability. However when raters have used three rating scales (i.e. over 25 parameters) per stimulus simultaneously (Webb et al 2004), the GRBAS scale still demonstrated superior reliability in relation to the other two, thus suggesting format and parameter selection to be the important factors in reliability.

Several voice stimuli issues relate to reliability. Dysphonic voices are essentially complex due to great intra and inter-speaker acoustic variance and consequently are likely to be difficult to define psycho-acoustically (Gerratt and Kreiman 2000). Furthermore discriminating and rating Uni-dimensional parameters can be difficult (Kreiman and Gerratt 1998 and 2000). However some voices have been suggested as more likely to show inter-rater consistency (Kreiman et al 1993). The mix of voice stimuli can also impact on reliability as there is greater agreement for normal or severe voice quality, but mild to moderate qualities are less reliably measured (Kearns and Simmons 1988; Gerratt et al 1993; Kreiman et al 1993). However this was subsequently suggested to be a task-related issue as opposed to a problem with voice perception as perfect midrange agreement occurred when raters were asked to adjust a synthetic voice to match a natural voice stimulus (Gerratt and Kreiman 2000).

A major area of concern regarding reliability relates to statistics. These should be carefully examined (Kreiman and Gerratt 1998) as inappropriate application of models from other disciplines has contributed to a potential "theoretical dead end" (Gerratt and Kreiman 2000). The latter study criticised intraclass correlations and Cronbach's alpha because a) they can mask "large and predictable" differences in agreement of different voices and b) "high" reliability co-efficients (0.9 with Cronbach's alpha) may not exceed chance agreement. Cronbach's alpha is designed to measure internal consistency of test items in questionnaires and consequently is not an appropriate or accepted measure of agreement between judges (Steiner and Norman 1995 p64). However the criticism of intraclass correlations warrants further examination. Whilst weighted quadratic kappa statistical analysis will permit the investigation of individual rater behaviour no statistical packages permit calculations to determine whether these kappa co-efficients differ significantly for individual raters or according to rater type. One sub-type of intraclass coefficient will permit such analysis (Streiner and Norman 1995 p126). Reliability has also been suggested to relate to rater task. GRBAS scale scores for individual voices vary across raters, particularly for scale points 1 and 2 (Kreiman et al 1993). Such poor inter-rater reliability was related to mapping voice stimuli to unstable internal standards with EAI scales (Kreiman et al 1998). An alternative methodology involved raters comparing stimuli to an anchor rather than relying upon their unstable internal representations (Kreiman and Gerratt 1998). Several studies suggest this improves reliability (Gerratt et al 1993; Chan and Yiu 2002; Yiu et al 2007; Awan and Lawson 2009). However anchors are not a panacea to prevent poor agreement. Anchors may: a) be associated with improved inter but not intra rater agreement (Awan and Lawson 2009; Eadie and Kapsner-Smith 2011), b) relate to high listener variability if no training is included (Chan and Yui 2002), c) cause decreased agreement if they differ from the test item (Kreiman et al 2007) and d) cause rater difficulty if rating scalar points fall between anchor stimuli (Kreiman and Gerratt 2000).

Rater training would intuitively be expected to improve agreement. However the literature is not clear on this point. One study concluded that extensive training does not necessarily increase intra-rater reliability (Kreiman et al 1993 p25). Paradoxically it may teach raters to focus on different aspects of complex auditory stimuli and cause increased inter-rater variance (Kreiman et al 1993 p25). In contrast, several subsequent studies have suggested that rater training does result in improved reliability in voice perception (Chan and Yiu 2006; Chan et al 2012) and hypernasality perception (Lee et al 2009). Nevertheless, one investigation reported inconsistent patterns of improvement across stimulus types and intra/inter agreement (Eadie and Baylor 2006). Furthermore training does not appear to compensate for experience. Eight hours' training did not allow non-experienced SLT's to achieve the same reliability as experienced judges (Bassich and Ludlow, 1986), corroborating findings that reliability depends on careful rater selection (Abe et al 1986).

Finally, it has been suggested that any structured rating task can cause judges to behave differently to their typical perceptual processing in day-to-day situations (Kreiman et al 1990). This indicates research may alter the very factors it is seeking to measure. This poses many challenges in task design. Some solutions to overcome these challenges have included requesting clinicians to rate highly specific dimensions or employing paired sample preference rating as a heuristic for which features are perceptually salient. As regards the former, for instance one requests raters to evaluate in relation to one clearly defined and agreed variable, such as tonicity which has been demonstrated to constitute key variables in judgements of the target voice disorder to be rated. Tonicity would seem an optimal choice in rating tracheoesophageal voice as evidence to date suggests it is the major determinant of overall voice quality (Hurren et al 2009) and there is a strong evidence base that it can be reduced in hypertonicity/spasm with botulinum toxin injection (Terrell et al 1995; Crary and Glowasky 1996; Hoffman et al 1997; Brok et al 1998; Zormeier et al 1999; Meleca et al 2000; Ramachandran et al 2003).

As regards paired samples preference rating, this removes some of the influences on rater performance such as severity of the preceding voice stimuli. This is a useful methodology for analysing treatment effects especially in research settings but it would be a cumbersome and inappropriate method to use for routine outcome (as outlined in section 1.4). A key aspect of this thesis is to design a clinically relevant and useful scale. By use of further acoustic examination of what features differentiate well separated samples followed by regression analyses, this offers insights into the key sound variables that should be targets of scale development or listener evaluation.

The calculation of reliability will remain an important aspect in developing a tracheoesophageal voice rating scale. The most reliable tool to date for laryngeal voice perception (the GRBAS scale) has involved an EAI scale and there is no robust evidence that other formats offer superior reliability (Webb et al 2004). However evidence from the laryngeal perceptual literature suggests the importance of ensuring a sufficiently wide range of voice stimuli types are included in investigations when reporting on a scale's reliability. However reliability is not the sole consideration. Although Webb (2005 p125) demonstrated the GRBAS to be the most reliable scale expert SLT judges felt the parameter type range to be too simplistic for therapeutic purposes in clinical settings; this illustrates how reliable scales may be insufficiently sensitive to small differences in voice quality. Further investigation is required as this work is in its infancy but it is a key aspect of content validity which will be considered in the following section.

2.2.4 Rater variance and validity issues

Scale validity is of primary importance yet research (as discussed above) has focussed mainly on rater reliability. Robust reliability co-efficients do not preclude concerns about scale validity (Kreiman and Gerratt 1996, 1998, 2000). Kreiman, Gerratt and co-workers have been at the forefront of research and postulated a conceptual framework for laryngeal voice quality perception (Kreiman et al 1993). They expressed concern that most studies are based on the assumption that mapping a stimulus to a scale point is a constant linear process, with variations in rating classed as random rater error. The issue of error in voice perception rating was detailed in section 1.6.1. If raters used identical perceptual strategies, voice quality could be attributed solely to differences in voices and listeners would be interchangeable. Rating scale theory has tended to imply quality can be attributed to the stimulus as opposed to a psychoacoustic interaction between the voice and the rater's perception (Kreiman and Gerratt 1998). Consequently listener behaviour has not been the focus of research (Kreiman et al 1990). Studies of raters have demonstrated that highly experienced listeners disagree about which parameters are perceptually important in dysphonic voices (Kreiman and Gerratt 1996). Although systematic error in the form of raters showing individual biases and differing internal standards and sensitivities to parameters has been demonstrated (Kreiman et al 1993) the rating tasks per se are a further source of error (Kreiman et al 2007). This makes content validity challenging as selection of parameters will need to potentially encompass a range of idiosyncratic psychoacoustic preferences. Furthermore task design is the key factor. Kreiman, Gerratt and co-workers designed experiments focussing on one parameter at a time including the use of synthetically manufactured stimuli, methods which complement or feed directly into the strategies in variable identification mentioned at the end of section 2.3.3. Such procedures are crucial for developing the theoretical basis underpinning voice perception, though clearly they do not provide the final solution for how these analyses can be carried out in clinical settings. So, for instance, Webb et al (2004) demonstrated the GRBAS scale is sufficiently reliable and valid despite its limitations. Clinical utility of scales consequently differs from research experiments where the format of scale construction and task cannot be applied to routine clinical practice. However the ongoing theoretical research should be seen as providing vital steps in determining the parameters and scale types that constitute reliable and valid scales.

One solution to addressing this issue of content validity could be to introduce training of listeners prior to evaluation to agree on parameters and definitions. There is some evidence that training can make this aspect worse as discussed in section 2.2. The CAPE-V (Kempster et al 2009) has addressed the idiosyncrasy factor by including some blank scales for raters to nominate themselves but this aspect is for clinical use and has not been investigated in research settings.

The issue of criterion validity is equally problematic. The key tool against which voice perception has been measured is acoustic instrumentation. However this seemingly objective measure has considerable validity and reliability issues in its own right (Carding et al 2009; Maryn et al 2009). Acoustic parameters do not necessarily relate to perceptual ones. A meta-analysis of twenty-five investigations into the relationship between perceived overall laryngeal voice quality and their acoustic correlates highlighted marked reservations about the criterion validity and the clinical utility of acoustic measurement (Maryn et al 2009). Furthermore a review of all laryngeal outcome measures also specified concerns about the use of such instrumental measures due to: a) limited information about their sensitivity to change and b) concerns about their reliability and validity particularly in relation to moderate and severely dysphonic voice qualities (Carding et al 2009). These issues in relation to alaryngeal voice will be discussed later and are especially pertinent given the marked aperiodicity of neoglottal phonation. Instead the criterion validity of the GRBAS has been established due to the highly significant correlations between its five parameters and validated patient self report questionnaires (Webb et al 2004).

The complexity of assessing construct validity in relation to perceptual voice quality was outlined in section 1.5.1. The psychoacoustic interaction means there is no basis for determining the correct judgement (Gerratt and Kreiman 2000). Although Gerratt and Kreiman (2000) concluded that scales are probably adequate but not optimal for clinical or experimental purposes they continue to be widely used in both settings. This can be justified in terms of the revised theoretical basis of validity previously outlined in section 1.5.1 (Streiner and Norman 1995 p147) regarding what can be inferred from a measure. The validity of the GRBAS scale has been confirmed over time and use (Webb 2005 p122); this can be observed by its link to patient self report tools (Webb et al 2004), sensitivity to change (Steen et al 2008), utility (Carding et al 2009) and that most patients seek help for improvement in the sound of their voice see this is the key judgement of treatment success (Carding et al 2009).

The validity issues discussed above are all equally applicable to tracheoesophageal voice quality ratings. It would appear optimal to consider the validity of an alaryngeal voice rating scale using similar methods.

2.2.5 Rater type

There has been limited research to investigate the effect of rater type on reliability and patterns of laryngeal voice quality judgement. Furthermore concerns about statistical methodology have been documented as highlighted above (Gerratt and Kreiman 2000). There are many permutations of rater type comparison i.e. Speech and Language Therapist (SLT), Student SLT, Ear, Nose and Throat (ENT) Surgeon, Naïve and Patient but no studies have included the carer perspective.

With respect to professional raters it is difficult to compare ENT and SLT due to the paucity of research. De Bodt et al (1997) investigated both the effect of profession (SLT/ENT) and expertise; there was no statistical significance between raters, but average scores showed trends that SLT's were more reliable. This suggests profession could have more effect than just exposure to dysphonic voices and may relate to SLT's training and routine use of perceptual scales. The only other study reported ENT and SLT showed similar rating for post-thyroidectomy voices (Helou et al 2010).

When the expertise of just SLT raters is considered there is no clear consensus. Experienced qualified therapists have been observed to be more reliable than students (Bassich and Ludlow 1986; Bele 2005) but other studies have shown the two groups to be equally consistent in reliability and severity judgements (Damrose et al 2004; Eadie et al 2011).

Several issues emerge when naïve raters are compared to professionals. Naïve raters have been considered unable to rate dysphonic voices due to a lack of specific internal standards that only develop with exposure to pathological voices (Kreiman et al 1994). Furthermore naïve and expert focus on different aspects of normal and dysphonic voices when asked to compare voice stimuli (Kreiman et al 1990); naïve raters employ a consistently "inflexible perceptual strategy", using only a few parameters, whereas experts (SLT and ENT) demonstrate substantial differences in parameters considered important to judging voice stimuli similarity. Kreiman et al (1994) hypothesise experts' training facilitates psycho-acoustic awareness of a larger range of parameters. Studies to compare naïve and SLT raters demonstrated naïve listeners have lower agreement (Sofranko and Prosek 2012; Helou et al 2010). There is some evidence that naïve raters can achieve reliability with Global parameters (Eadie and Doyle 2002a; Eadie et al 2010a); however the former study used inexperienced SLT students who cannot be classed as truly naïve listeners and the latter selected statistics that did not account for chance agreement.

A recent study selected an alternative methodology whereby judges ranked dysphonic voice stimuli in order of severity using new software (NeAR) (Gould et al 2012). Naïve and SLT judges performed similarly which demonstrated they perceive Overall Grade in a similar manner. It is also important to consider whether Naïve judges are a homogeneous group; those with musical training demonstrated some aspects of higher reliability (Eadie, van Boven et al 2010b). The concept of naïve judges is potentially different in tracheoesophageal voice ratings as conceivably far more people would meet the criterion i.e. never having heard this type of voice. This is discussed below in section 2.4.

Few studies have compared how patients evaluate their voice in comparison to other rater types. Again there is no clear consensus. Investigations have demonstrated there is no difference between voice evaluations of SLT, Naïve or Patient rater groups although patients seem to be using different strategies (Eadie et al 2007; Eadie, Kapsner et al 2010a). However another investigation concluded patients rated their voice more severely than SLT's and showed significantly lower test-retest reliability (Lee et al 2005).

Further research is clearly needed to establish how rater type and expertise influences the severity and reliability of voice evaluation.

2.2.6 Summary

Many studies are methodologically flawed and comparisons are difficult due to lack of consensus regarding terminology. Concerns regarding scale use extend beyond reliability to validity and are rater, stimuli, task and scale format dependent. Few parameters have proven reliability in repeat investigations except for the equally appearing interval scale for GRBAS. Direct magnitude estimation research is in its infancy, but along with anchor stimuli may develop to address some of the concerns raised about the reliability of the equally appearing interval format. Raters' reliability is likely to depend upon a complex interaction of voice quality stimuli, scale format, rater experience/level of training and interactions between the task and listener (Carding et al 2000; Eadie and Doyle 2002c).It is difficult to ascertain how individual factors help or hinder reliability (Kreiman et al 1993). Rater type appears to be a key issue but investigations are limited to date. However there is some evidence that naïve rater agreement is limited to Global parameters due to their internal representation of voice. Scales exist for clinical purposes, and must be relevant. The GRBAS scale has been criticised for being too simplistic for planning clinical interventions (Webb 2005 p125); consequently further research is still warranted.

2.3 Current alaryngeal voice quality scales

A wide variety of scales have been used to rate alaryngeal voice quality. Most were designed to assess oesophageal voice (Robbins et al 1984; Trudeau 1987; Williams and Watson 1987; Nieboer et al 1988) or to compare tracheoesophageal to laryngeal voice and/or other methods of alaryngeal communication (Robbins et al 1984; Williams and Watson 1985; Cullinan et al 1986; Watson and Williams 1987; Williams and Watson 1987; Pindzolaand Cain 1988; Silverman and Black 1994). Tracheoesophageal perceptual scales to date are thus founded in tools designed for oesophageal voice measurement (Nieboer et al 1988; van As et al 2003).There are key similarities between tracheoesophageal and oesophageal voice as they share the same phonatory source. However Perry and co-workers (Cheesman et al 1986; Perry 1989; McIvor et al 1990) demonstrated that oesophageal voice was not achieved by patients with non-optimal tone or stenosis. Oesophageal speakers will not demonstrate the full range of qualities that are found in patients who have undergone SVR. Consequently scales devised for oesophageal voice are not likely to capture some key parameters found in tracheoesophageal voice e.g. whisper, wetness. Table 1 summarises the main studies. The key issues are discussed below.

2.3.1 Baseline definitions

In order for perceptual rating scales to be considered a valid measure of alaryngeal voice quality “clear definitions must be established” (Eadie and Doyle 2005). Perhaps the most fundamental aspect relates to the underpinning anchor baseline against which voices will be compared. The baseline for most laryngeal voice scales is “normal” voice quality (Hirano 1981; Wilson 1987; Hirano 1989) but a similar baseline for tracheoesophageal voice is more complex. Clearly the baseline can be either normal voice or to the optimal tracheoesophageal voice outcome.

Table 1. Summary of tracheoesophageal voice perceptual rating scales.

Authors and subject number	Investigation	Rating Scale Type and Parameters	Parameter and Baseline Definition	Intra/ Inter rater reliability	Type and number of raters
Blom et al (1986) N=47 consecutive	Prospective. To compare voice pre &post secondary SVR.	EAI 1-5 Acceptability	No	Test/Re-test for one group of 5 raters. No statistics.	Naïve N=80.Groups of 5 rated 3 patients each.
Bridges (1991a) N=12	Prospective. To compare oesophageal, tracheoesophageal and tracheoesophageal + TSV speakers according to rater type.	EAI 1-7 Fluency, Intelligibility, Pitch, Volume, Rate, Quality, Effort, Stoma Noise, General Acceptability	Written description of each parameter.	No. Calculated difference between 3 rater groups from overall mean scores.	SLT N=9, ENT N=5, Naïve N=10
Bridges (1991b) N=8 SVR, 4 oesophageal voice. All "superior speakers".	Prospective. To assess the perception and production of pitch contours in alaryngeal speech.	Patients imitated target word "key" in three tonal patterns. Raters marked tonal pattern as correct or incorrect.	No	Inter % correct. T test to compare to chance.	SLT N=6
Brown et al (2003) N=32 (16 of each type)	Retrospective. To compare primary and secondary SVR + prosthesis.	Visual analogue 100mm Generic perceptual rating.	No	No. All scores combined and averaged to compare primary and secondary.	SLT N=1, Naïve N=1, Carer and Patient
Cantu et al (1998) N=36	Retrospective. To examine long term success rates of functional communication in primary and secondary SVR and predictors of success.	Functional communication profile 1-5	Yes	No	SLT N=1 Patient Carer (only if patient deceased)
Cullinan et al (1986) N=5 (of each type)	Retrospective. To assess reliability of intelligibility rating and compare SVR and oesophageal voice.	EAI 1-5 Intelligibility	No	Inter and Intra. Test/re-test after 1month.	SLT N=9, Naïve N=9

Authors and subject number	Investigation	Rating Scale Type and Parameters	Parameter and Baseline Definition	Intra/ Inter rater reliability	Type and number of raters
Delsupehe et al (1998) N=116 (some had both prosthesis types)	Prospective RCT. To compare 2 voice prostheses.	EAI 1-5 Intonation, Intelligibility, Acceptability. EAI 1-3 Extraneous Noise. EAI 0-1 Loudness, Rate.	Normal voice baseline for Noise, Rate Loudness.	Inter (but only selected most consistent judge for detailed statistics).	Naïve=4, Expert=4.
Deschler et al (1994) N=11	Retrospective. To compare outcomes of total laryngectomy +/- free flap reconstruction	Visual analogue, Intelligibility, Communicative Effectiveness, Pitch, Volume, Rate, Pleasantness, Wetness, Fluency, Stoma Noise.	No	Inter measured with MANOVA.	SLT=6, Naïve=6.
Dworkin et al (1999) N=25 (only "effective" speakers selected)	Prospective. To study neoglottis of effective SVR speakers with stroboscopy.	EAI 1-7 Intelligibility, Fluency, Stoma Noise, Hoarse, Strain, Gurgly, Pitch, Pitch Stability, Volume, Volume Stability, Rate.	Definition for Intelligibility only.	% agreement for inter and intra	SLT N=2, ENT N=1
Eadie& Doyle (2002a) N=20 (male "better than average speakers")	Prospective. To determine validity of Overall Severity and Naturalness Parameters with DME and EAI.	EAI 1-9 and DME Overall Severity, Naturalness.	Written definitions. Baseline not overt but compared to normal voice.	25% test retest in same session. Coefficients used.	Naïve N=20
Eadie& Doyle (2004) N=28	Prospective. To assess SVR voice quality and correlation with QoL.	DME Overall Severity, Naturalness, Acceptability and Pleasantness.	No	25% test retest in same session.	Naïve N=15
Eadie& Doyle (2005a) N=20 (selected best speakers)	Prospective. To compare Acceptability and Pleasantness with EAI and DME.	EAI 1-9 and DME Acceptability and Pleasantness.	Yes	25% test retest in same session.	Naïve N=10

Authors and subject number	Investigation	Rating Scale Type and Parameters	Parameter and Baseline Definition	Intra/ Inter rater reliability	Type and number of raters
Finizia et al (1998) N=28 (14 of each type)	Retrospective. To compare SVR voice with laryngeal speakers who underwent radiotherapy.	Visual analogue 100mm. Intelligibility, Quality, Acceptability.	Yes. Baseline for Quality is normal voice.	Inter and intra % only. No true test-retest	Naïve N=10, SLT N=5
Finizia et al (1999) N=24 (12 of each type).	Retrospective. To compare SVR voice, laryngeal speakers who underwent radiotherapy and normal voice.	Visual analogue 100mm Intelligibility, Quality, Acceptability.	Assumed to be as per study above	Inter and Intra % only	Naïve N=10, SLT N=5
Fujimoto et al (1991) N=1	Prospective. To compare voice outcome with and without tracheostoma valve.	Visual analogue 100mm. Overall Voice Quality,	No	Inter – 3 way ANOVA	Naïve N=12, SLT N=12
Heaton et al (1996) N=20	Prospective. To compare acceptability of SVR voice with voice prosthesis type.	Overall Impression EAI 1-7.	Yes	Inter only. % agreement exact or within one point. Kappa only to see if chance scores.	ENT N=1, SLT N=1, Naïve N=1, Patient.
Kao et al (1994) N=166	Retrospective. To compare and evaluate primary and secondary SVR results including medical issues.	EAI 1-5 Volume, Pitch, Rate. (Combined SLT and Patient self rating score).	Yes. Pitch and Quality mixed definitions. Volume baseline is normal voice.	No	SLT N=1
Kazi, Kiverniti et al (2006a) N=20	Prospective. To compare male and female SVR speakers with EGG, acoustic, perceptual and QoL assessments.	GRBAS van As et al's(2003) Overall Voice Judgement scale.	Only for van As' et al scale.	Intra and Inter. Intra Class Co-efficient	Expert ENT N=2

Authors and subject number	Investigation	Rating Scale Type and Parameters	Parameter and Baseline Definition	Intra/ Inter rater reliability	Type and number of raters
Kazi, Singh et al (2006b) N=42	Prospective. To assess the neoglottis with videofluoroscopy and perceptual voice assessment.	GRBAS. van As et al's (2003) Overall Voice Judgement scale.	Only for van As' et al scale.	Intra and Inter. Intra Class Co-efficient	Expert ENT N=2
Kazi, Kanagalingam et al (2009) N=47	Prospective. To compare EGG and voice quality of SVR and laryngeal speakers.	GRBAS van As' et al (2003) Overall Voice Judgement.	Only for Van As' et al scale.	Intra (1 rater) and Inter Co-efficients used	Expert ENT N=2
Lundstrom et al (2008) N=9	To relate measurements of the neoglottis to acoustic and voice perceptual assessments.	Visual analogue 1000mm Gurgly, Hyperfunctional/Tense, Breathy, Rough.	"Tentative written and oral definitions"	Intra and Inter Co-efficients used.	SLT N=5
McAuliffe et al (2000) N=43 (30 Total laryngectomy, 13 Pharyngolaryngectomy and jejunum graft)	Retrospective. To compare laryngectomy with pharyngolaryngectomy and jejunum graft.	TOMS, Robillard-Schultz & Harrison, EAI 1-5 Effort, Stoma Noise, Pleasantness, Naturalness, Intelligibility.	Yes but EAI not defined.	No	Not specified.
Mahieuet al (1987) N=71	Retrospective. To research outcomes in oesophageal voice and primary SVR ±myotomy.	3 point adjectival (Good, Moderate, Poor).	No	No	Author plus unspecified number of SLT's.
Meleca et al (2000) N=5	Prospective. To assess voice quality pre & post botulinum toxin injection using stroboscopy.	EAI 1-7 Overall Intelligibility, Fluency, Strain.	Definition for Intelligibility. No baseline.	Intra test/re-test. % within 1 scale point	SLT N=2, ENT N=1
Moerman et al (2004) N=53 (SVR, oesophageal voice and partial laryngectomy).	Prospective. To assess the best method of acoustic analysis to support voice perceptual analysis.	Visual analogue 100mm Tonicity, Fluency, Overall Grade, Voice Onset, Stoma Noise, Tempo, Intonation, Intelligibility	No	Inter only. Co-efficient used.	Semi-professional (SLT students) N=10

Authors and subject number	Investigation	Rating Scale Type and Parameters	Parameter and Baseline Definition	Intra/ Inter rater reliability	Type and number of raters
Moerman et al (2006) N=68	Prospective. To assess inter rater reliability of a SVR perceptual scale.	Visual analogue Overall Grade, Fluency, Stoma Noise, Voice Quality.	Definitions but no baselines.	Inter only. Co-efficient used.	Semi-professional (SLT students) N=24, SLT N=6 (2 rated 30-40 voices each in 3 centres)
Most et al (2000) N=15 (5 in each group).	Retrospective. To compare normal, oesophageal and tracheoesophageal voice.	EAI 1-6. Acceptability.	No	No	Naïve N=25
Nagle & Eadie (2012) N=18.	Prospective. To determine whether naïve raters can judge Listener Effort and if this parameter is relevant to intelligibility and acceptability.	Visual analogue 100mm judgement of similarity with paired comparison task. Speech Acceptability, Listener Effort.	Written definitions	Intra (10% of stimuli only) and Inter. Co-efficients used.	Naïve N=20
Nieboer et al 1988 N=18 SVR	Prospective. To apply a rating scale designed for normal voice to alaryngeal voice. To differentiate oesophageal and tracheoesophageal voice quality.	Semantic differential 1-7. 13 parameters adapted from a normal voice quality perceptual scale	Yes. Baseline is not to compare to normal voice.	Inter=mean correlation between raters.	Naïve N=34 SLT students N=51
O'Leary et al (1994) N=9	Retrospective. To replicate Tardy-Mitzell et al (1985).	EAI 1-7 Acceptability, Intelligibility	Yes. Baseline not discussed.	No. Mean rating range and mean for group.	Naïve N=8 SLT N=1
Pindzola & Cain (1988) N=15 (5 from each group)	Prospective. To compare SVR, oesophageal and normal voice.	EAI 1-7 Fluency, Pitch/Quality, Inflection, Rate, Acceptability (mean score).	No	Inter - Mean scores.	Naïve N=16
Robillard-Shultz & Harrison (1992) N=24	Retrospective. To research success of SVR, use of valve, voice quality and care of valve.	EAI 1-5. Overall Severity score of quality of combined intelligibility, fluency and occlusion ability	SVR voice - well defined levels.	No statistics. Inter & Intra=% agreement of 2 raters	Authors. Not blinded & familiar with patients.

Authors and subject number	Investigation	Rating Scale Type and Parameters	Parameter and Baseline Definition	Intra/ Inter rater reliability	Type and number of raters
Sanderson et al (1993) N=10 (4 oesophageal, 6 SVR)	Prospective. To compare oesophageal and SVR voices.	EAI 1-5. Intelligibility, Rate, Acceptability, Pleasantness Pitch. (Combined score=20).	No	No	SLT N=1. Not blinded
Shipp (1967) N=33	Prospective. To assess acceptability of oesophageal voice in relation to fundamental frequency.	EAI 1-5. Acceptability and Stoma Noise	Not defined.	Mean rating for each speaker for each parameter.	Naive N=116
Singer et al (1986) N=25	Retrospective. To investigate tracheoesophageal voice outcome after neurectomy.	Fluent versus Dysfluent. No rating scale.	No	No	Authors.. Not blinded & familiar with patients.
Tardy-Mitzell et al (1985) N=15	Retrospective. To evaluate SVR voice acceptability & intelligibility.	EAI 1-7. Acceptability, Intelligibility.	Yes. Baseline not discussed	No. Mean, median and mode for group only.	Naïve N=46
van As et al (2003) N=40 .	Prospective. To compare voice to acoustic, videofluoroscopic and high speed digital imaging instruments.	3-point adjectival Overall Voice Judgement, Semantic differential 1-7. 20 items	Overall voice judgement only.	Yes. Test/re-test for 50% in same session. Repeated unrandomised. Co-efficients used.	SLT N=4 Naïve N=20
van den Hoogen (1998) N=105	Prospective. To compare voice outcome with Groningen, Nijdam and Provox voice prostheses.	3 point adjectival. Availability of Sound, Fluency, Stoma Noise, Intelligibility, Voice Quality (i.e. Hypertonic/tense, Normotonic, Hypotonic/Lax)	Some definitions.	Inter only=% agreement.	SLT N=1 Experienced listener (non-SLT) N=1
van Weissenbruch et al (2000) N=40 (20 from each group)	Prospective. To investigate SVR ± Myotomy&myotomy + neurectomy with videofluoroscopy	3- point adjectival. Overall Grade	No	Intra (Unclear test/re-test protocol - 20 re-tested) and Inter. Correlation co-efficients.	SLT N=2

Authors and subject number	Investigation	Rating Scale Type and Parameters	Parameter and Baseline Definition	Intra/ Inter rater reliability	Type and number of raters
Vlantis et al (2003) N=17	Prospective to compare exdwelling 16fg to indwelling 22fg voice prostheses.	3 point adjectival. Voice Availability, Fluency and Intelligibility.	No	Inter Co-efficient used.	SLT N=2
Ward et al (2011) N=21	Prospective. To compare two voice prostheses.	Paired comparisons EAI 0-+3 or -3 similar/dissimilar. Steadiness, Pitch, Fluency, Intelligibility, Strain, Vocal Effort, Overall Grade.	No.	Intra=% exact agreement (10% retest) Inter=Co-efficient	SLT N=4
Watson and Williams (1987) N=43 (11 of each type plus 10 normal)	Prospective. To compare types of communication (electrolarynx, tracheoesophageal, oesophageal and normal voice).	EAI 1-7. Quality, Pitch, Loudness, Rate, Intelligibility, Extraneous Noise, Overall Communicative Effectiveness.	Written definitions. Unclear baseline.	Inter Co-efficient used. Comparison of judges within each rater type group	Naïve (SLT students) N=3 Post graduate SLT (some laryngectomy experience) N=3 Expert SLT N=3 Patient panel N=4 (3 electrolarynx, 1 oesophageal voice)
Wetmore et al (1981) N=18	Prospective. To compare intelligibility pre &post secondary SVR.	4 point adjectival. Intelligibility.	Yes (% words intelligible).	No	Unclear type or number
Williams and Watson (1987) N=43 (11 of each type plus 10 normal)	Prospective. To compare speaking proficiency of oesophageal, tracheoesophageal and electrolarynx speakers with normal, laryngeal speakers.	EAI 1-7. Quality, Pitch, Loudness, Rate, Intelligibility, Extraneous Noise, Overall Communicative Effectiveness.	Written definitions Unclear baseline.	Inter Co-efficient used.	Naïve (SLT students) N=12

Section 1.4 concluded it appears more appropriate for scales to be compared to the latter. Most studies have failed to address this core issue, or at best used inadequately defined parameters. It should be noted, however, that a baseline of normal laryngeal voice is appropriate for some types of investigation e.g. comparing voice outcomes of total laryngectomy with laryngeal conservation treatments and normal controls (Finizia et al 1998; Finizia et al 1999).

Some authors have used the GRBAS scale to assess tracheoesophageal voice (Omori and Kojima 1999; Kazi, Kiverniti et al 2006a; Kazi, Singh et al 2006b;Kazi et al 2009) but have failed to address the underpinning issue that GRBAS uses a baseline of normal voice quality. This has the potential to cause scores to cluster at the severe end of the scale because raters judge the markedly atypical tracheoesophageal voice against a normal voice quality baseline. This fundamental error is likely to artificially inflate reliability and compromise validity.

The alaryngeal voice scales of Nieboer et al (1988) and van As et al (2003)form the main research base in this area and thus warrant more detailed discussion here. Nieboer et al's scale was designed to compare oesophageal and tracheoesophageal voice by adapting a tool originally constructed to rate normal voices (Blom and Koopmans-van Beinum 1973; Blom and van Herpt 1976; Fagel et al 1983). Their rationale for this adaptation related to laryngeal scales having "proved to yield results at least with normal, healthy voices" despite such parameters being unlikely to be key features of tracheoesophageal voice. Furthermore the investigation required naïve and SLT students to judge voices in relation to alaryngeal not laryngeal voice but it is unclear how such untrained assessors could be expected to have an established internal baseline of alaryngeal voice quality.

Van As et al subsequently adapted Nieboer's scale to create the most comprehensive tracheoesophageal scale to date. The Overall Judgement scale definitions are assessed in relation to normal laryngeal voice as outlined in Figure 7. However no guidelines are provided for the 21 bipolar scales (Figure 8) although the study methodology highlighted that baseline formulation was included in the pre-rating training.

Some bipolar parameters are particularly subjective/descriptive i.e. ugly-beautiful, pleasant-unpleasant and have no readily discernible baseline even for normal laryngeal speakers. They appear not to measure voice quality alone and could covertly assess accent and aesthetic voice properties. The parameters slow-quick and low-high would appear to have a midpoint score of 4 as the normal baseline as neither scale endpoint is optimal or alternatively the end scale points may be interpreted as representing optimally high or optimally low judgements. This contrasts with parameters 1, 4, 8, 9, 13, 14, 15, 16, 17 and 20 which require a score of 7 to be within normal limits. Scale variability regarding the numerical and comparator baseline are not discussed or adjusted for in any later statistical analysis. The parameters that measure neoglottal tonicity (18 and 19) do not occur in laryngeal voice and thus must have a baseline of optimal tracheoesophageal tone. However as low tone (hypo) and high tone (hyper) are separate 1-7 scales, tonicity is in effect measured on a 14 point scale.

Figure 7. van As et al (2003) Overall Voice Judgement scale definitions.

PhD Thesis Definition	Overall Severity Parameter	Journal Definition
almost similar to normal	Good	most similar to normal voice
somewhere between both extremes.	Reasonable	in between extremes
very deviant from normal voice	Poor	least similar to normal

Figure 8. van As et al (2003) Tracheoesophageal perceptual scales.

a) Overall Severity°[2]	Good
	Reasonable
	Poor
b) Selected Parameters Bipolar Semantic Scale 1-7	1. Deviant- normal[1]
	2. Unpleasant-pleasant
	3. Ugly –beautiful
	4. Noise- no noise
	5. Monotonous-melodious
	6. Expressionless- expressive
	7. Weak-powerful
	8. Unsteady-steady
	9. Jerking-fluent
	10. Slow- quick
	11. Low-high
	12. Bubbly-not bubbly[1]
	13. Breathy-not breathy
	14. Rough-not rough[1] [2]
	15. Creaky-not creaky[1] [2]
	16. Tense-relaxed*
	17. Dull –clear[2]
	18. Hypertonic-not hypertonic°[1]
	19. Hypotonic –not hypotonic°[1]
	20. Unintelligble - intelligible

*only for naïve raters
° only for expert raters
[1] added to original scale by Van As.
[2] Reliability not achieved in van As et al (2003) study

From an examination of the literature to date it would appear optimal for tracheoesophageal scales to have parameters with clearly specified textual references at each scale endpoint to facilitate uniform use of scales.

2.3.2 Parameter selection and definition

The selection of parameters for a rating scale is obviously a key issue. Its relationship to content validity was highlighted in section 1.6.2. The first and most fundamental concern is whether tracheoesophageal voice quality is best judged by parameters traditionally selected to rate laryngeal dysphonia or whether different parameters are required due to the characteristically different voice qualities produced by the reconstructed neopharynx. The second concern is whether global parameters are sufficient or if more specific uni-dimensional parameters would better encapsulate the core psychoacoustic features of this complex voice quality.

With respect to the first issue, key researchers in tracheoesophageal voice have criticised the practice of assessing tracheoesophageal speech "relative to the inherent characteristics of normal laryngeal voice and speech" (Doyle and Eadie 2005a p.115). The authors concede Intelligibility is the exception if the research aims to compare tracheoesophageal and laryngeal speakers or to compare methods of alaryngeal speech.

Optimal parameter selection has had limited consideration because only six previous studies have specifically aimed to design scales to measure tracheoesophageal voice quality (Nieboer et al 1988; Eadie and Doyle 2002b; van As et al 2003; Eadie and Doyle 2004; Eadie and Doyle 2005a; Nagle and Eadie 2012). None of these investigations has adequately addressed content validity to ascertain whether the selected parameters assessed the key features of alaryngeal voice. Parameters included in studies to date are summarised in Table 1 column 3. As highlighted above, Nieboer et al and van As et al formulated their scales from parameters designed to assess normal voice quality resulting in unclear baselines and subjective parameters. Although van As et al added parameters (12, 18, 19) as "features occurring in alaryngeal voice" supporting evidence from the literature was not provided.

The remaining studies that set out to develop tracheoesophageal perceptual parameters (Eadie and Doyle 2002a; Eadie and Doyle 2004; Eadie and Doyle 2005a; Nagle and Eadie 2012) focussed almost exclusively on global parameters with the following rationale: a) they measure the social perspective of the individual's voice, b) they measure the speech signal proficiency as a whole - which makes them a more appropriate measure of "communication rehabilitation outcomes" and c) they are clinically useful as they assess the speaker performance within a communicative context better than uni-dimensional parameters. However all the studies cited above included only naïve judges. Parameters selected for scales designed for only naïve raters are likely to differ from those developed for expert SLT's. It is unclear if such over-arching parameters have sufficient sensitivity to measure change and/or distinguish between different voice types and hence whether they are clinically useful. However, uni-dimensional parameters have also been proposed as important (Doyle and Eadie 2005a p134) because: a) we do not know which uni-dimensional parameters most influence listeners' ratings of the global parameters, b) they permit better differentiation of speakers, c) investigating the correlation and interaction of these parameters will allow for more targeted rehabilitation. On this basis it would seem optimal to include both global and uni-dimensional parameters in the measurement of tracheoesophageal voice quality.

Although a wide range of parameters have been included in investigations to date most have failed to select those that assess the unique features of tracheoesophageal voice that are universally acknowledged as of clinical importance (e.g. extraneous noise as air escapes from the stoma (Shipp 1967; Perry 1989 p84; Silverman and Black 1994; Moerman 2006), wetness/gurgliness as oesophageal secretions vibrate on phonation (Blom et al 1986; Omori and Kojima 1999; Robb and Lewin 2003; van As-Brooks et al 2005), hypo or hypertonicity (Singer et al 1986; Perry 1989; van As-Brooks et al 2005; Hurren et al 2009) phonation from a rigid stenosed fibrotic neopharynx (Singer et al 1986; Perry 1989; Hurren et al 2009). It is important to note that there is some overlap of features between laryngeal and tracheoesophageal voice e.g. fluency, strain, volume, pitch but simple transference of scales from laryngeal to tracheoesophageal voice will omit to measure core outcomes and compromise validity. Furthermore most studies have not defined their parameters as summarised in Table 1 column 4. Clear definitions are an important aspect of validity and enhancing reliability.

2.3.3 Scale format

The key requisites for scale construction are sensitivity in differentiating between speakers and treatment effects and ease of use in clinical settings. Scale formats included in studies to date are summarised in Table 1 column 3. No studies have undertaken research to determine whether their tool is able to differentiate tracheoesophageal voices but several have investigated treatment effects such as surgical option (Singer et al 1986; Mahieu et al 1987; Deschler et al 1994; McAuliffe et al 2000;) or type of voice prosthesis (Delsupehe et al 1998; Vlantis et al 2003; Ward et al 2011). Unfortunately all used unvalidated scales and the majority failed to assess reliability adequately or included other methodological flaws e.g. recruiting only two raters (Vlantis et al 2003; Kazi et al 2006a; Kazi et al 2006b; Kazi et al 2009). Consequently there is no evidence to suggest that any of the scale formats used to date i.e. equally appearing interval (EAI), visual analogue (VA), adjectival, direct magnitude estimation (DME) or paired comparisons can offer superior sensitivity.

The other scale requirement (ease of clinical use) would appear to be met by the EAI or adjectival scale formats. Other scale formats may be more sensitive i.e. VA, paired comparisons or DME. However each has other drawbacks. VA scales of 100mm are more time consuming as they require measurement to determine the result of each parameter and there is no evidence that raters can judge 100 points in tracheoesophageal perceptual analysis. Paired comparison scales appeared to be sufficiently sensitive to allow raters to differentiate between two types of voice prosthesis (Ward et al 2011). This format would be highly likely to be sensitive to change but would be cumbersome to use and interpret in routine clinical settings. It is also noteworthy that van As was a co-author for this study but there was no reference to why her previously published scales (van As et al 2003) were not selected for this study. It is possible this was due to concerns about scale sensitivity.

Finally with respect to DME, even the main proponents of this scale format advised it is not yet at a stage where it can be used clinically as expert consensus is first required to determine suitable moduli against which to allocate judgements (Eadie and Doyle 2005).

2.3.4 Reliability

The most serious research limitation in the literature concerns the omission of inter and intra- rater reliability assessment. Studies have aimed to investigate issues such as SVR success rates, medical complications or surgical reconstruction outlined in section 2.1 using idiosyncratic scales without addressing scale development methodology. Scale reliability is crucial as a lack of stability and reproducibility renders it inherently invalid (Streiner and Norman 1995 p6).

A robust methodology to examine reliability should consider at a minimum:

a) test-retest for the whole stimuli sample over a reasonable period (for example several weeks) as voice stimuli recall can extend to several days (Kreiman 1997),

b) inter rater calculation to compare raters,

c) employment of statistical design that provides a robust indication of the odds of chance agreement,

d) an adequate number of raters that represent a breadth of the group concerned,

e) an adequate number of voice stimuli that represent the range of severities and types of variability,

f) employment of rater blinding e.g. such that the rater is not aware of the names of the patients if they are likely to be familiar with them or raters are blind to type of voice restoration, time of recording and so forth.

The reliability methodology, design methods and results of statistical analysis for previous studies are summarised in Table 1 and 2 respectively. It is clear that none of the published studies have robust evidence of reliability and many failed to include any agreement calculations. Test-re-test issues include inadequate design where raters reassessed only a limited sample of voices and within the same session (Eadie and Doyle 2002; van As et al 2003; Eadie and Doyle 2004; Eadie and Doyle 2005; Ward et al 2011; Nagle and Eadie 2012). Reliability is often inadequately calculated using percentage agreement between raters or mean score calculations for inter-rater reliability (Nieboer et al 1988; Pindzola and Cain 1988; Robillard-Schultz and Harrison 1992; O'Leary et al 1994; Heaton et al 1996; Finizia et al 1998; van den Hoogen 1998; Meleca et al 2000) or utilising co-efficients that do not calculate for chance agreement (van As et al 2003;Eadie and Doyle 2002a; Eadie and Doyle 2004; Eadie and Doyle 2005a; Kazi et al 2006a; Kazi et al 2006b; Moerman et al 2006; Lundstrom et al 2008; Kazi et al 2009; Ward et al 2011; Nagle and Eadie 2012).

Outcomes of the few studies that have included reliability measures are summarised in Table 2 and include the co-efficients for reliability in relation to parameter type.

Table 2.Inter and intra rater reliability of tracheoesophageal perceptual rating scales.

Authors	Rater type	Parameter and Reliability Co-efficient	Statistics
Blom et al (1986)	Naïve N=80	Acceptability- 0.885 (Inter) Intelligibility - 0.998 (Inter)	Methodology of statistics not stated. Groups of 5 judges rated only 3 patients each.
Cullinan et al (1986)	Naïve N=9 SLT N=9	Intelligibility - Naïve 0.85 (Intra), 0.96 (Inter). SLT 0.84 (Intra) 0.96 (Inter).	Intra and Inter reliability = Pearson's co-efficient.
Eadie and Doyle (2002a)	Naïve N=20 (Graduate SLT students)	DME: Overall Severity - 0.62 - 0.98 (Intra), 0.87 (Inter), Naturalness - 0.49-0.99 (Intra), 0.95 (Inter). EAI: Intra- mean is 86% within one scale point and 96% within 2 points for both parameters. Naturalness -0.96 (Inter), Overall Severity 0.97 (Inter).	DME. Intra = Pearson's co-efficient. Inter = Cronbach's alpha. EAI. Intra = % agreement within one scale point. Inter = Cronbach's alpha.
Eadie and Doyle (2004)	Naïve N=15 (Graduate SLT students)	Overall Severity - 0.76 (Intra) 0.97 (Inter), Naturalness - 0.77 (Intra) 0.92 (Inter), Acceptability - 0.85 (Intra) 0.96 (Inter), Pleasantness 0.74 (Intra) 0.96 (Inter).	Intra = Pearson's co-efficient, Inter = Cronbach's alpha.
Eadie and Doyle (2005a)	Naïve N=10 (Graduate SLT students)	EAI: Acceptability - 0.71 (Intra) 0.71 (Inter), Pleasantness - 0.68 (Intra), 0.73 (Inter). DME: Acceptability - 0.83 (Intra) 0.70 (Inter), Pleasantness – 0.77 (Intra) 0.70 (Inter).	Intra = Spearman's for EAI, Pearson's for DME. Inter = avoided group mean data and each rater looked at as an individual. Examined relationships among listeners' judgments. Intercorrelational matrix selected to report on listener by listener basis to avoid masking listener variance with group mean data.
Kazi, Kiverniti et al (2006a)	ENT N=2	van As Overall Voice Judgement over 0.80 (Intra)and 0.86 (Inter) Intra GRBAS (over 0.8) Inter GRBAS (range 0.89-0.96)	Inter and Intra Intra Class Co-efficient
Kazi, Singh et al (2006b)	ENT N=2	van As Overall Voice Judgement 0.88 (Intra)and 0.86 (Inter) Intra G (0.9), R (0.8), B (0.8), A (0.5), S (0.5) Inter G (0.88), R (0.81), B (0.88), A (0.45), S (0.5)	Inter and Intra Intra Class Co-efficient

Authors	Rater type	Parameter and Reliability Co-efficient	Statistics
Kazi, Kanagalingum et al (2009)	ENT N=2	van As Overall Voice Judgement 0.9 (Intra and Inter) Intra G (0.9), R (0.8), B (0.8), A (0.5), S (0.5) Inter G (0.9), R (0.6), B (0.7), A (0.7), S (0.7)	Inter and Intra (1 rater) Intra Class Co-efficient
Lundstrom et al (2008)	SLT N=5	Intra: Hyperfunctional (0.69-0.96), Breathy (0.75-0.94), Rough ("low reliability" reported, no data), Gurgly ("low reliability" reported, no data) Inter: Hyperfunctional (0.85-0.90), Breathy (0.93-0.94), Rough (0.30-.66), Gurgly (0.30-0.66).	Intra = Pearson's, Inter = Cronbach's alpha
Moerman et al (2004)	Semi-professional (SLT students) N=10	Inter rater only. Hyper/Hypotone 0.54, Fluency 0.68, Voice Onset 0.57, Additional Noise 0.55, Intonation 0.55, Tempo 0.70, Intelligibility 0.75, General Impression 0.78.	Inter = Pearson's
Moerman et al (2006)	SLT N=6	Inter rater only. Overall Grade 0.68, Intelligibility 0.68, Stoma Noise 0.57, Fluency 0.67, Voice Quality 0.58.	Inter = The mean was calculated for each unit, then the mean for all units, then Kendall's Tau. NB not all voices were rated by all SLT's.
Nagle and Eadie (2012)	Naïve N=20	Listener Effort - 0.78 (Intra), 0.71 (Inter). Speech Acceptability - 0.78 (Intra) 0.66 (Inter).	Intra = Pearson's, Inter = compare each rater's scores with group mean with Intra Class Correlation.
Shipp (1967)	Naïve N=116 (SLT students)	Acceptability - 0.75 (Intra). Stoma Noise 0.73 (Intra)	Co-efficient not specified. Inter rater calculated by mean agreement.
van As et al (2003)	SLT N=4 Naïve N=40	SLT: Range for 21 parameter scale 0.64-0.93 (Intra), 0.57-0.92 (Inter). Naïve: Range for 19 parameter scale 0.23-0.75 (Intra) 0.87-0.94 (Inter)	Intra = Pearson, Inter = Cronbach's alpha
van Weissenbruch et al (2000)	SLT N=2	Intra 0.46 - 1.0 (mean 0.66) Inter 0.51-1.0 (mean 0.68) Videofluoroscopy and voice scores not separated consequently difficult to interpret.	Inter and Intra - Kappa
Vlantis et al (2003)	SLT N=2	Inter rater only. Two co-efficients relate to two different types of voice prostheses. Availability - 0.75, 1.0, Fluency - 0.65, 0.73, Intelligibility - 1.0, 1.0	Inter = Kappa

Authors	Rater type	Parameter and Reliability Co-efficient	Statistics

Ward et al (2011)	SLT N=4	Intra: Percent exact 72% (range 72-82%) and percentage close agreement 89% (range 77-100%)
		Inter: Overall Grade 0.83, Steadiness 0.79, Pitch 0.52, Fluency 0.80, Intelligibility 0.74, Strain 0.84, Vocal Effort 0.84
			Intra = % exact agreement (10% re-test)
			Inter = Cronbach's alpha

Watson and Williams (1987)	SLT N=3, Naive N=3, Post-graduate SLT N=3, Patient N=4	0.67-0.87 range of co-efficients per group. No comparisons between groups.	Inter = Pearson's
Williams and Watson (1987)	SLT Students=12	States highly significant p <.001. No other information.	Inter = Pearson's

The gold standard for reliability and validity assessment was addressed in section 1.6. Streiner and Norman (1995 p121) summarised the difficulty in establishing when reliability scores are sufficiently high because recommendations reported in the literature are generally subjective, brief and without justification. Van As et al (2003) considered Cronbach's alpha co-efficients in excess of 0.7 to indicate sufficient reliability for their perceptual tracheoesophageal scales reflecting the value also recommended by Nunally (1978). Alternatively Landis and Koch (1977) considered kappa co-efficients over 0.61 to be "good" and Streiner and Norman (1995 p7) specified scale stability of 0.5 or above is acceptable but higher scores are necessary if fatality could feasibly occur as a result of employing a scale.

The majority of studies summarised in Table 2 have selected Cronbach's alpha, intra class correlations and/or Pearson's correlation to measure rater agreement and the majority reported high reliability. However Cronbach's alpha, as previously outlined (1.6.1), is a measure of internal consistency of scale items (Streiner and Norman 1995 p64) and does not represent patterns of agreement among raters nor indicate agreement for specific voice samples (Kreiman and Gerratt 1998; Kreiman and Gerratt 2000). The suggestion that Cronbach's alpha produce artificially high reliability results in voice research (Gerratt et al 1997) was outlined in 2.2. Similarly, Pearson's correlation has been considered to be a non-optimal statistic for the calculation of rater reliability (McDowell and Newell p36) especially as it provides a more liberal measure of reliability unless the predominant source of error is random error (Streiner and Norman 1995 p115). The work of Kreiman and Gerratt (2011) demonstrated that this error pattern is not the key issue in perceptual voice analysis. No tracheoesophageal voice studies have used weighted Kappa co-efficients which account for chance agreement. Although one subtype of intraclass correlation will provide identical co-efficients to weighted quadratic kappas as outlined in 1.6.1, the three studies that included intraclass correlations (Kazi et al 2006a; 2006b;2009) failed to specify which subtype was utilised. Consequently it is not possible to conclude whether the co-efficients in the Kazi et al series were optimal for the assessment of rater reliability.

Despite the limitations in methodology, several valuable themes emerge. Intra rater reliability is generally inferior to inter rater (Cullinan et al 1986; van As et al 2003; Eadie and Doyle 2004a; Lundstrom et al 2008). This is in contrast to laryngeal perceptual studies which have reported the opposite effect. Overall Grade was either the most reliably assessed parameter (Vlantis et al 2003; Moerman et al 2006; Ward et al 2011) or achieved high levels of reliability (van As et al 2003).

Many studies have compromised reliability due to rater factors. These relate to the inclusion of only one or two judges (Robillard-Schultz and Harrison 1992; Sanderson et al 1993; Kao et al 1994; Heaton et al 1996; Brown et al 2003;Vlantis et al 2003; Kazi et al 2006a; Kazi et al 2006b; Kazi et al 2009), failure to specify rater type or numbers (Wetmore et al 1981; Mahieu et al 1987) and/or failure to blind judges who are likely to be familiar with the patients (Robillard-Schultz and Harrison 1992; Sanderson et al 1993; Heaton et al 1996).

Reliability has been investigated in relation to scale format i.e. equally appearing interval (EAI) scales versus direct magnitude estimation (DME) (Eadie and Doyle 2002a; Eadie and Doyle 2005a). The authors concluded Acceptability and Naturalness are measurable with an EAI scale as they can be reliably psycho-acoustically intervalised in contrast to Overall Grade and Pleasantness which require DME as raters cannot reliably sub-divide these into equal intervals. However several issues need to be considered in relation to this assertion. The studies only included naïve raters who were in fact SLT students and voice stimuli from "better than average" speakers rather than a representation of the spectrum of tracheoesophageal voices. Furthermore the inferior reliability of the EAI scale could relate to the selection of a nine point scale; this is longer than typically used in voice analysis and exceeds recommendation of optimal scale length (Streiner and Norman 1995 p35).

The issue of rater type in relation to reliability is a key consideration and consequently will be addressed in a separate section (2.4).

2.3.5 Validity

Validity aspects of perceptual voice analysis were outlined previously in section 1.5 in terms of how they broadly relate to content, criterion and construct validity. No studies to date have adequately addressed the validity of a perceptual tracheoesophageal rating scale.

The only detailed and investigated scale to date (van As et al 2003) has not adequately addressed parameter selection as discussed above. This potentially compromises content validity. Furthermore van As' scales had no history of prior use in clinical practice, there was no pilot data and no evidence of clinical consensus regarding parameter selection which would meet the lowest level required for content validity (Streiner and Norman 1995 p5). Further potential compromise of content validity relates to the translation from Dutch to English. Nieboerand Van As undertook research in Dutch with publication in English. This may be especially pertinent for relatively subjective perceptual terminology where the meaning may not be equal across languages (Streiner and Norman 1995 p24).

A further validity concern relates to criterion validity because there is no gold standard for tracheoesophageal voice assessment (as highlighted in section 1.6.2). Although researchers have compared tracheoesophageal perceptual voice to measures gained from manometry (Perry 1989), videofluoroscopy (van As et al 2003; Kazi et al 2006b; Lundstrom et al 2008) and acoustic analysis (Kazi et al 2009; van As-Brooks et al 2005) these instrumental measures all have inherent limitations of reliability and validity themselves.

Construct validity is also problematic in perceptual voice analysis as discussed in relation to the theoretical basis of Kreiman and Gerratt in section 2.2. The psychoacoustic interaction means there is no basis for determining a correct judgement (Gerratt and Kreiman 2000) and voice parameters are consequently considered as hypothetical constructs i.e. a psychoacoustic interaction between the voice stimulus and the rater's internalised memory of voices. Certain parameters could potentially relate to more instrumental measures e.g. pitch to fundamental frequency, volume to decibel level, fluency/rate to syllables per minute. Tonicity may also be investigated in relation to intra-oesophageal manometry. However this interlinks with the problems of criterion validity where there is no robust alternative assessment against which these constructs can be evaluated. Future research to compare these assessments may be beneficial but it is likely this would need to be considered in relation to the limitations. It is however a circular argument since, in order to establish construct validity, an alaryngeal perceptual rating scale needs to be devised.

2.3.6 Subjects

The number of patient speakers included in studies to date are summarised in Table 1, column 1. As can be seen, the number of speakers is mostly small and recruitment procedures and inclusion/exclusion criteria are frequently unclear (Tardy-Mitzell et al 1985). Sometimes only the "best" speakers are specifically recruited (Dworkin et al 1999; Eadie& Doyle 2002 and 2005) for purposes that are not adequately explained. This renders the application to clinical practice as unclear. Only one study recruited a consecutive series of patients of various vocal proficiencies and qualities (total = 47) (Blom et al 1986). Patient consent is essential in research but self-selecting patients may restrict or bias the cohort in a variety of ways, for instance the possibility that only the best rehabilitated speakers who are often keen to volunteer come forwards but who may not represent a clinically representative sample.

2.3.7 Summary of the limitations of current rating scales and criteria for a future robust scale

Current research into tracheoesophageal voice perceptual analysis has many methodological flaws. Although some scales have demonstrated good reliability, none has adequately addressed validity. Deficiencies include failing to specify baselines, address parameter inclusion (or definition) and/or to assess reliability with adequate test-re-test design and optimal statistical soundness. Ideally, a robust scale should:

a) demonstrate intra and inter rater agreement with statistics that provide robust indication of the odds of chance agreement,

b) address content, criterion and construct validity,

c) include both global and uni-dimensional parameters,

d) be sufficiently sensitive to be able to document inter and intra-patient variation,

e) the scale format should be easy to implement in clinical situations (have maximum utility value),

f) use equally appearing interval scales (in the absence of evidence that visual analogue scales can offer superior reliability).

2.4 Tracheoesophageal rating scales: from whose perspective and why should it matter?

The types of rater included in studies to date are listed in Table 1, column 6. Investigations have focussed mainly on SLT and naïve raters judgements. Few studies have included ENT surgeons or patients and carers have been included in only one study to date. It is essential to investigate the perceptions of professional, naïve, patient and carer raters to assess the impact of tracheoesophageal voice quality comprehensively; such comparisons of rater type effect are extremely limited. The unique features of each rater sub-group are outlined below with reference to the published studies from the literature.

2.4.1 Patient self rating

The patient's self rating of voice is perhaps the key outcome measure as this is the only group who directly experience the altered voice. They may be more likely than outside observers to be sensitive to parameters such as strain and vocal fatigue (Meleca et al 2000). However, a cancer diagnosis leads to anxiety regarding survival and poor voice quality may be more readily tolerated when cure is the key preoccupation. Such issues of rater perspective have been reported in other relevant areas of speech and language pathology where patients' judgements have been compared to those of SLT's i.e. dysarthria (Walshe et al 2008); spasmodic dysphonia (Sapir et al 1986) and dysphonia (Lee et al 2005). All of these studies concluded that SLT and patient ratings are unlikely to concur due to the different context of each group and because patients also judge their voice/speech via kinaesthetic awareness and bone conduction.

Several studies have included patients' views as one of multiple research aims but this has resulted in only brief discussion of this key aspect of results (Ackerstaff et al 1994; Kao et al 1994; Silverman and Black 1994; Brown et al 2003; Kazi et al 2005). An absence of validated tools has led to studies: a) using non-validated questionnaires (Silverman and Black 1994; b) using tools validated for laryngeal voice self-rating (Moerman et al 2004; Schuster et al 2005; Evans et al 2009; Day and Doyle 2010) or c) developing new questionnaires concerning general voice prosthesis issues but not focussing specifically on voice quality aspects (Silverman and Black 1994; Kazi, Singh, de Cordova et al 2006c). These assessments were designed solely for patient use and consequently cannot be used to compare patients' views to those of other rater types.

There are very limited insights into patients' perceptions of tracheoesophageal voice from investigations to date. Patients have reported difficulty speaking over background noise (Silverman and Black 1994; Op de Coul et al 2005) and considered the most negative feature to be stoma noise (Silverman and Black 1994). One study reported the "majority" of patients rated their intelligibility as "fair or good", sixty percent were content with volume, two-thirds were happy with fluency and age was unrelated to voice perception. (Ackerstaff et al 1994).

The fundamental issue of lack of agreement between rater types can be observed in studies that demonstrated laryngectomy patients' attitude to communication was unrelated to their speaking proficiency (Williams and Watson 1988).

A final omission in research concerns whether patients can reliably rate uni-dimensional voice parameters in their own voice. This is an important consideration as a study in the related field of laryngeal conservation surgery reported male patients preferred a "rough" quality and females more "breathy" voices (Doyle 1997). Consequently patient scales that include only global parameters are likely to be omitting key features that could allow patients to describe and rate their own voice quality in more detail and enable more patient centred rehabilitation focussed on the aspects that cause them the most concern.

2.4.2 Carer

This group represents the viewpoint of those who spend the most time with the patient and are likely to act as their support and confidante. Carers have unique insight into functional intelligibility and social acceptability, witnessing reactions to the patient's voice in a variety of settings. They may be privy to a third party's negative perceptions of the patient's voice of which the patient remains unaware, in denial or prefers not to disclose in a voice rating scale. Carer views have been minimally investigated to date. This is potentially a major limitation of previous studies when carers can provide the unique perspectives outlined above. Brown et al (2003) included carers' ratings for their own relative only; however this included only the mean score for an "overall impression" parameter and did not assess inter rater agreement in comparison to scores from the patient, SLT and naïve judges.

2.4.3 Naïve

The ratings of naïve listeners are crucial because they represent members of the community the speakers will encounter in their daily lives outside their immediate family. In this context, naïve raters are commonly defined as "listeners who have not previously encountered alaryngeal voice". A unique feature of this group is their ability to determine the "social penalty" of tracheoesophageal voice (Eadie and Doyle 2004).

However many studies have used SLT student raters as naïve raters (Watson and Williams 1987; Williams and Watson 1987; Nieboer et al 1988; Pindzola and Cain 1988; Eadie and Doyle 2002a; Eadie and Doyle 2005a; Eadie and Doyle 2004; Moerman et al 2004) but their training means they do not fulfil the definition above. They are consequently unlikely to be representative of the view of the local community. Other studies have used non-SLT undergraduate students (Blom et al 1986; O'Leary et al 1994; Most et al 2000) but such recruits do not represent the general population in terms of education and age.

Only two studies have combined the recruitment of non student naïve raters with more relatively robust reliability assessments (van As et al 2003; Nagle and Eadie 2010). However the good inter and intra rater reliability should be considered in light of the statistical concerns raised above as both used statistics that do not account for chance agreement. Van As et al investigated multiple uni-dimensional parameters with this rater group and reported inferior intra rater agreement to SLT raters; this reduced agreement was attributed to their judgments shifting over time as naïve judges become accustomed to the "deviant" voice quality. This study also found naïve scores clustered in the more severe scale points for all parameters with failure to use the uni-dimensional scales to differentiate between voice stimuli. These factors could have contributed to the high reliability reported. Nagle and Eadie's (2010) paired comparison task for Acceptability and Listener Effort was simpler as judges do not need to rate against an internalised baseline of a quality and this would be expected to afford more reliability than an equally appearing interval scale.

2.4.4 Professional (SLT and ENT)

SLT's experienced in tracheoesophageal voice rehabilitation should theoretically be able to rate stimuli in comparison to an internalised reference baseline of optimal tracheoesophageal voice. However it has been suggested experts may be too "desensitised" to the impact of tracheoesophageal voice to accurately assess social acceptability (Eadie and Doyle 2004). The SLT viewpoint is important as their judgement of voice quality is a key component in their treatment planning. SLTs are responsible for rehabilitating voice, selecting voice prosthesis type, assessing voice with videofluoroscopy and other instrumentation and motivating patients whilst helping them to adjust to their new communication status. SLT raters have been involved in a large percentage of the studies to date. In contrast, few studies have included ENT surgeon raters and there is currently no evidence about their reliability in rating or how their judgement relates to that of other sub-groups. Although surgeons do not carry out formal perceptual assessments, their informal judgements may covertly or overtly influence how they may evaluate and select future options for surgery.

Only two studies have discussed the potential for professional rater bias (Heaton et al 1996; Cantu et al 1998). Both studies reported SLT's rated voices as more superior but both only included one SLT rater who was not blinded to the patients who were well known to the unit. It was postulated that SLT's might rate the voices more highly because low scores may reflect on their rehabilitation therapy. Similarly, SLTs may be accustomed to working with more severe communication impairments (Heaton et al 1996) and hence rate tracheoesophageal speakers more favourably. However one ENT surgeon was included in Heaton et al and bias was not considered for this group. Further more robustly conducted research is required before conclusions about professional bias can be addressed.

2.4.5 Comparisons of rater type

The methodological flaws discussed in 2.3 prevent evidence-based conclusions regarding rater-type influence. This is compounded by studies using varied scale type for different rater groups (Kao et al 1994;Silverman and Black 1994; Delsupehe et al 1998;McAuliffe et al 2000; Olthoff et al 2003).

There are no definitive conclusions regarding patient perspective in relation to other rater types. There have been few studies (Heaton et al 1986; Watson and Williams 1987; Cantu et al 1998; Brown et al 2003) and these have failed to include inter rater reliability statistics (Cantu et al 1998; Brown et al 2003), failed to blind professional raters familiar with the patients (Heaton et al 1996; Cantu et al 1998) or included patient judges as a panel of oesophageal and electrolarynx speakers rather than tracheoesophageal speakers' self rating (Watson and Williams 1987).

There is no robust evidence regarding how SLT and ENT raters compare in their judgement of tracheoesophageal voice. Some preliminary investigations have contrasted the SLT and Naïve perspectives but there is no clear consensus. All the studies have statistical issues as discussed previously (Cullinan et al 1986; Bridges 1991a; Finizia et al 1998; van As et al 2003). One study reported naïve listeners as rating all parameters more severely than SLT's (van As et al 2003). The authors attributed this to SLT's "internal standard" being tracheoesophageal speech which caused them to rate patients more highly due to familiarity with "deviant voice quality". Conversely with respect to Intelligibility Naïve judges have been observed to rate both a) more highly than SLT's (Finizia et al 1998) and b) similarly to SLT's (Cullinan et al 1986; Bridges 1991a).

2.5 Summary

There is a lack of research regarding carer, patient and ENT perspectives on tracheoesophageal voice outcome. Most studies have focused on SLT and naïve raters but the evidence base has been hindered by methodological issues and there is little consensus regarding the influence of rater type. Different raters bring different contexts and biases to the rating process which will influence their judgment. There is a need for more detailed studies to enable clinicians to understand how patients and carer perspectives relate to those who provide treatment for them or who will encounter them in their local community.

An ideal scale to measure tracheoesophageal perceptual voice quality by professional raters would include well-defined global and uni-dimensional parameters with specified baselines that relate to optimal tracheoesophageal voice quality where appropriate. Content, criterion and construct validity would be addressed and reliability would be established. The scale would be clinically relevant and sensitive to change or to differentiation of voice between patients. No scale to date has met these criteria and consequently there is no validated tool to meet the clinical and research needs outlined in section 2.1. As other types of raters are also crucial to include in outcome studies, scales should be devised to enable other types of judges to rate tracheoesophageal voices and to allow inter rater comparisons. A triangulated view of different rater types will require some commonality of scale items but with a more comprehensive scale for professionals to use.

2.6 Research aims

This thesis will aim to address these issues with the following five research aims:

1. To devise a reliable and valid perceptual rating scale for professional (SLT and ENT) raters to assess the complex parameters of tracheoesophageal voice.

2. To examine the inter and intra-rater agreement and reliability of the professional scale according to rater type and expertise.

3. To examine the inter and intra-rater reliability of naïve raters in the perceptual assessment of tracheoesophageal voice using a modified form of the expert scale.

4. To examine the patient and carer perspective of SVR voice outcome with a modified version of the naïve rater scale.

5. To examine the relationship between SLT, ENT, naïve, patient and carer raters of tracheoesophageal voice.

Chapter 3. The development and design of the tracheoesophageal perceptual rating scales

This chapter will outline the development and design of three separate rating scales to assess tracheoesophageal voice outcome from five different rater perspectives i.e. SLT, ENT, Naïve, Patient and Carer. The first two sections (3.1 and 3.2) concern the scale for professional (SLT and ENT) raters. Key background issues and rationale for the initial development of the scale are detailed (3.1) including the preliminary clinical application (3.1.2) and a subsequent pilot study (3.1.3). This is followed by a detailed description of the revised, final version of the professional scale (3.2). The third and fourth sections outline the design and development of the rating scales for Naïve listeners (3.3) and Patients/Carers (3.4). The three rating scales are detailed in Table 3. These scales were subsequently used to collect performance data in a number of separate studies detailed in Chapters 4 and 5.

Table 3. Overview of scales developed in relation to rater type.

Rater Group	Rating Scale used
Professional: SLT's ENT surgeons	Sunderland Tracheoesophageal Perceptual Scale (SToPS)
Naïve	Naïve Rating Scale (for people with no prior experience of tracheoesophageal voice)
SVR patients	Patient and Carer Rating Scale
Carers of SVR patients	Patient and Carer Rating Scale

3.1 The design and development of a new perceptual scale for professional raters

The first consideration concerns the requirement of a scale that focuses on the key aspects of tracheoesophageal voice quality. There is some commonality of features between laryngeal and alaryngeal phonation but careful consideration is warranted.

The aim was to design a perceptual scale that addressed the criticisms of the tools previously outlined (in section 2.3) and to provide an optimal scale that meets the key criteria summarised in section 2.3.7. The aspects that were essential to consider at this stage of the scale development were:

1. Parameter selection – Each parameter should have a clear rationale to demonstrate clinical relevance i.e. permit accurate description of tracheoesophageal voice quality. Both global and uni-dimensional parameters should be included.

2. Scale Format – This should be clinically practical in design and in length i.e. in terms of the number of parameters to be assessed. It should also be sensitive to change or to differentiation between patients.

3. Parameter baselines and definitions – The baseline or "standard" against which each parameter should be measured should be clearly defined and justified. Clear guidance notes for the definitions of each parameter and scale point will facilitate scale reproducibility, training and reliability.

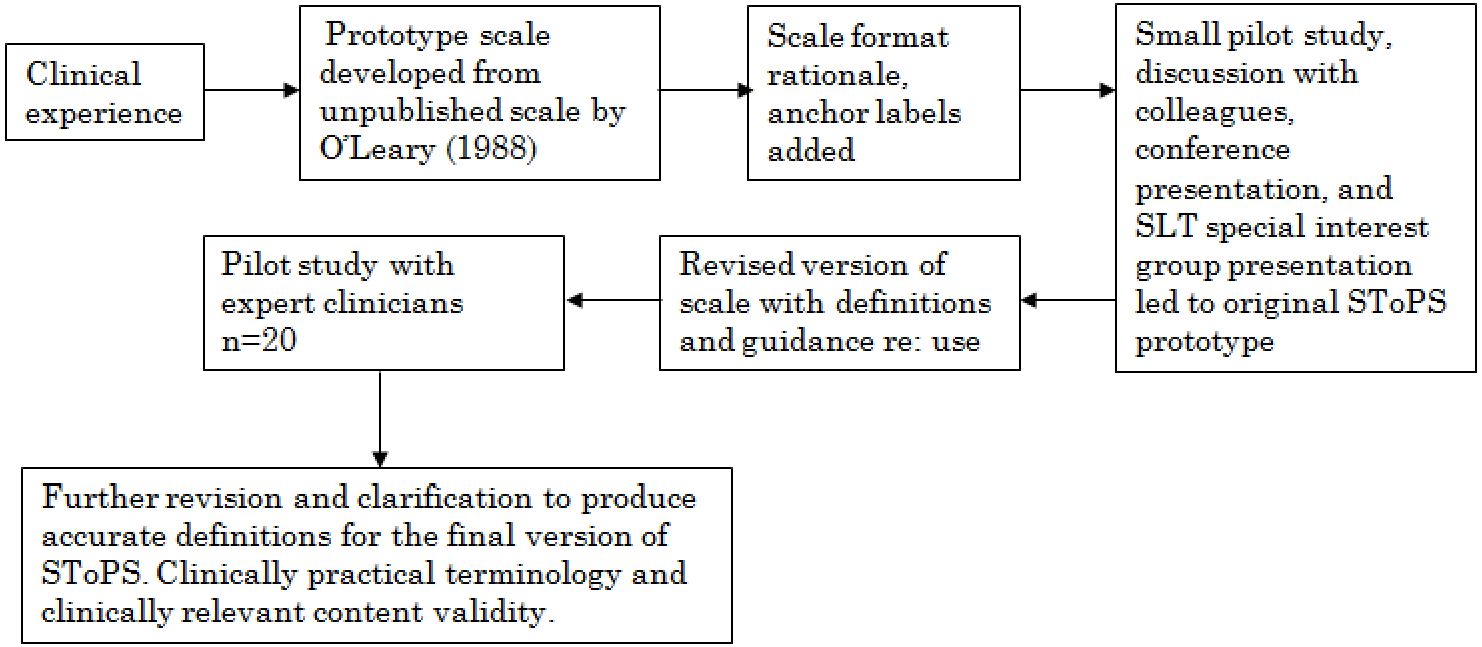
These features are intended to enable reliability and validity of the scale, issues that are addressed more fully in Chapter 6.

3.1.1 First version of the SToPS

The scale design process for the Sunderland Tracheoesophageal Perceptual Scale (SToPS) is illustrated in Figure 9. The first version of the SToPS was an extended and modified version of an unpublished scale by O'Leary (1988) which consists of four parameters (Quality, Acceptability, Fluency and Intelligibility) in a 1-5 equally appearing interval (EAI) format. Scale points are totalled to give a potential score of 20, which relates to the optimal outcome attainable. The rationale for the modification of O'Leary's scale relates to it being insufficiently sensitive to differentiate between the inter and intra-patient variations in tracheoesophageal speakers when applied to patients in the author's clinical practice. Additional reservations included the absence of specified baselines and guidance notes, the inability to distinguish between tonicity types and the absence of other key clinical features of tracheoesophageal voice such as “Strain”, “Wetness” and “Stoma noise”. The evidence base for the inclusion of these aspects of tracheoesophageal voice was detailed in section 3.2. A further concern about O'Leary's scale relates to its format of totalling scores across potentially independent parameters. When the scale was used in clinical practice it was observed that the majority of speakers achieved scores of over 16 but the parameters included did not allow differentiation of the subjective impression of key perceptual differences between speakers. Furthermore the scale had not undergone validity or reliability investigation.

The design and development of the SToPS will be discussed in relation to the essential criteria for scale design outlined above.

Figure 9. Scale design process for the Sunderland Tracheoesophageal Rating Scale (SToPS).



Clinical experience

Prototype scale developed from unpublished scale by O'Leary (1988)

Scale format rationale, anchor labels added

Small pilot study, discussion with colleagues, conference presentation, and SLT special interest group presentation led to original SToPS prototype

Revised version of scale with definitions and guidance re: use

Pilot study with expert clinicians n=20

Further revision and clarification to produce accurate definitions for the final version of SToPS. Clinically practical terminology and clinically relevant content validity.

Parameter selection

The selection of parameters for the first version of the SToPs was in keeping with Streiner and Norman's (1995 Chapter 3) seminal examination of optimal scale item selection. Further aspects of this work were included for subsequent stages of the SToPS' development. The recommendations from Streiner and Norman included in this draft relate to: a) a consideration of previous tools as a basis for new scale development (p15) as outlined above regarding O'Leary (1988), b) undertaking a literature review of research findings from the area (p19) and c) clinical observation from the author of this thesis (p17). A full rationale for each parameter selected with reference to the literature is outlined in 3.3.2 but a brief synopsis is included below. The parameters selected for the first version of the SToPS are listed in Table 4.

Table 4. Parameters selected for the first draft of the SToPS.

Global Parameters	Uni-dimensional Parameters
Overall Grade	Neoglottal Tonicity
Impairment of Social Acceptability	Stenosis
Impairment of Intelligibility	Wetness
	Strain
	Stoma Noise
	Impairment of Volume
	Impairment of Fluency

The importance of measuring both global and uni-dimensional aspects of tracheoesophageal voice was summarised in 2.3.2 and consequently both sub-types were included.

Global parameters

Global parameters have a key role in a comprehensive, optimal tracheoesophageal scale. Furthermore the aim of this thesis is to examine voice perception from multiple rater type perspectives and it is essential to include some commonality of scale items with such over-arching parameters that have been demonstrated to have good reliability with non-professional raters (Eadie and Doyle 2002a; Eadie and Doyle 2004; Eadie and Doyle 2005a). Studies to date have included Intelligibility, Overall Grade, Acceptability, Naturalness and Pleasantness. The latter two parameters were not selected for the SToPS on the basis that they appear to be more descriptive terms that may have very different meanings to different people. Consequently they are not likely to be stable concepts across individual listeners. Furthermore they may be easily affected/influenced by a listener's personal preference, especially if the speech sample also contains other paralinguistic features of accent or if the speaker has learned English as a second language (Mackey et al 1997).

Overall Grade has been demonstrated to have good reliability for both laryngeal and alaryngeal voice as discussed in 2.2 and 2.3 above. Acceptability has been described as a key parameter because listener discomfort will "override measures of intelligibility" (Eadie and Doyle 2002a). The concept of "acceptability" is obviously not relevant to laryngeal voice but is clinically crucial in tracheoesophageal voice. Alaryngeal voice invariably does not sound "normal" and with the exception of optimal speakers has features very different to dysphonic laryngeal speakers and hence may not be "acceptable" in some form to listeners. Consequently the parameter Acceptability has the potential to capture different factors to Overall Grade i.e. an intelligible, functional voice with atypical features that can be uncomfortable for the listener e.g. vibrating secretions on phonation, stoma noise. Furthermore this parameter could measure the social penalty of being a female tracheoesophageal speaker. This may relate to male and female tracheoesophageal speakers being indistinguishable from voice stimuli as low fundamental frequency patterns occur equally across both genders (van As 2001 p70).

Intelligibility has been included in many tracheoesophageal voice scales although obviously it is not solely a measure of voice quality. For this reason it has not been included in any published laryngeal voice scale. This is presumably because most voice clinicians do not feel that disordered laryngeal voice (dysphonia) renders the patient unintelligible (at least in most circumstances). Interestingly this may not necessarily be true. For example, there is some evidence that school children have more difficulty in retaining spoken information from dysphonic speakers in relation those with normal voice quality (Rogerson and Dodd 2005). This may relate more to increased listener effort for dysphonic speech rather than intelligibility problems per se. There is however little other published literature on this topic and it warrants further investigation. As the ultimate aim of communication is to convey meaning, "intelligibility" has been considered a key outcome of tracheoesophageal voice in many studies (Finizia et al 1999). There is considerable evidence that laryngectomy patients have compromised intelligibility. Several authors have attributed this reduced intelligibility to non-optimal neoglottic tonicity (Nieboer et al 1988; van As et 2001; Jongmans et al 2003; Jongmans et al 2010). Assessing intelligibility in connected speech is problematic; no established criterion validity exists in relation to tracheoesophageal voice and in laryngeal voice it has been demonstrated that ratings vary depending on environment and background noise (Cox and McDaniel 1984). Although a number of studies have confirmed single word and sentence level problems post laryngectomy (Bridges 1991a; O'Leary et al 1994; Miralles and Cervera 1995; Schuster et al 2006) restricting assessment at this level does not predict functional levels in connected speech (Cox and McDaniel 1984). It appears important to include this parameter in the SToPS as sentence level functional intelligibility assessment revealed significant issues attributed to voice quality, extraneous (stoma) noise, low sound intensity and high effort (McAuliffe et al 2000). Furthermore low rankings of Overall Voice Quality and Acceptability did not correspond to low scores of Intelligibility (Finizia et al 1998). This suggests that Intelligibility is independent of the other two

Global parameters and provides further evidence for its inclusion in the SToPS.

Uni-dimensional parameters

The definition and importance of this parameter sub-type was outlined in 2.3.2. Uni-dimensional parameters have been defined as forming composites that in turn make up the global impression of voice (Doyle and Eadie 2005a). Their key role is in permitting analysis of more individual aspects of voice perception e.g. strain, wetness. The crucial uni-dimensional parameters that constitute tracheoesophageal voice quality are outlined below.

Tonicity describes the perceptual impression of neoglottal tone. This concept was previously outlined as the key determinant of alaryngeal voice quality (1.4). There is some evidence that other uni-dimensional perceptual parameters are also linked to Tonicity. Van As-Brooks et al (2005) reported their bipolar scale parameter corresponding to the perception of vibrating secretions (bubbly-not bubbly) correlated with hypotonicity; this is because non closure of the neoglottis permits oesophageal secretions to be regurgitated with the tracheoesophageal airstream. The same study also observed the perceptual impression of strain (for the parameter "tense-not tense") was linked to hypertonicity where the walls of the neoglottis are tightly closed. Whisper quality has also been noted to be attributable to severe hypotonicity and strain (McAuliffe et al 2000).

Several studies have suggested that Tonicity is the major parameter that influences Overall Grade ratings (Singer et al 1986; Perry 1989; van As-Brooks et al 2005; Hurren et al 2009). However Stenosis (a key part of the tonicity spectrum) has been omitted from scale design to date with the exception of one study (Hurren et al 2009). This investigation reported

Stenosis was equally related to “poor” Overall Grade scores as were moderate to severe hypo and hypertonicity. As the perception of Tonicity requires an understanding of neoglottal physiology and its potential correspondence to other parameters it could be hypothesised that such assessment requires more advanced psychoacoustic skills.

Strain relates to the psychoacoustic impression of phonating against resistance. Neoglottal resistance and the air flow required to produce phonation have been observed to be greater in tracheoesophageal than laryngeal voicing (Singer 1983). This is associated with a) the high neoglottal closure pressure in hypertonicity (Perry 1989 p100), b) severe hypotonicity that occurs in jejunum grafts (McAuliffe et al 2000), c) excessive stoma closure pressure manually compressing the neoglottis (Perry 1989 p100) and d) a tightly stenosed neopharynx (Singer et al 1986).

Wetness is defined as the psychoacoustic impression of secretions or bolus residue vibrating within the neopharynx. It has been demonstrated to relate to hypotonicity because oesophageal secretions are transported superiorly on phonation through the open neoglottis (van As-Brooks et al 2005). Wetness has also been observed to occur due to mucus pooling above the neoglottis (Omori and Kojima 1999). It may also relate to liquid bolus residue vibrating on the egressive airstream. Wetness has been suggested to be associated with poor acceptability scores (Blom et al 1986).

Reduction of vocal volume in comparison to the level that would be expected in laryngeal speakers has been observed to occur in some but not all tracheoesophageal speakers (Clark 1985; McColl 2006). This would appear to be most noticeable with the whispery quality that occurs in severe hypotonicity (Perry 1989 p100) and stenotic voice quality (Singer et al 1986). Normal laryngeal speakers have been demonstrated to use a volume of 60-65db in quiet one to one conversational surroundings (Davis 1981).

However there is some evidence that the inter patient variability of loudness extends to the opposite end of the spectrum with some tracheoesophageal speakers having higher volume levels than their laryngeal counterparts as the range of volume in SVR speakers has been reported as 56 - 77 dB (Delsupehe et al 1998). It can be postulated that volume relates to neoglottal tonicity with hypotonic and stenotic speakers whispery voice quality at the lower end of the spectrum and the hypertonic speakers' tightly occluded larger neoglottal mass (van As-Brooks et al 2005) producing louder more strained voice quality. It would seem optimal for investigators to consider the range of tonicity types of the patient cohort when reporting the mean and range of voice stimuli volume. It is also important to consider the methodology of sampling voice intensity. One study just measured volume from a sustained vowel at maximum intensity and reported SVR speakers had a mean volume of 70.7 dB (Max et al 1996). However there was no assessment of whether such high levels would be maintained during everyday speaking situations. The knowledge base of tracheoesophageal speaker volume is currently in its infancy.

Whisperiness is defined as the psychoacoustic impression of air passing through a neopharynx where the neoglottis is either absent or fails to fully close on phonation. This physiological aspect of neoglottal function was previously outlined (section 1.4). This relates to a) hypotonicity (van As-Brooks et al 2005, b) pharyngeal reconstruction grafts which have a "breathy almost aphonic quality" (Deschler and Gray 2004) due to an absent neoglottis (van-As Brooks et al 2005) and c) stenosis where rigidity of the tissue produces a "coarse whisper quality" (Singer et al 1986).

Stoma Noise occurs due to an inadequately occluded stoma during tracheoesophageal voicing. This has been reported as the most undesirable feature of voice from the patients' personal perspective (Silverman and Black 1994); it can be so severe as to mask a whispery hypotonic voice signal

(Perry 1989 p84). Stoma noise has also been linked to low levels of Acceptability in oesophageal speakers (Shipp 1967).

Fluency can be within normal limits in comparison to normal laryngeal speakers (Pindzola and Cain 1988). However this parameter can decrease: a) as hypertonicity increases due to restriction of the egressive airstream (Perry 1989 p23), b) in relation to stoma noise as an inadequately closed stoma causes air wastage with increased breath pauses and reduced rate of speech (Finizia et al 1999) and c) as an artefact of the voice recording process in speakers with poor ability in reading aloud or producing a spontaneous speech sample to demand.

Prosody: tracheoesophageal speakers can achieve prosody control that is not significantly different to that of normal laryngeal speakers (Pindzola and Cain 1988; Bridges 1991b). However there is marked variation between speakers (Bridges 1991b). Sentence level prosody in reading aloud has been suggested to be better conserved in hypertonic voices (van As-Brooks et al 2005). The between speaker variation in prosodic skill has been attributed to differences in the length, mass and passive compliance of the neoglottis (Moon and Weinberg 1987).

Scale format

A 0-3 equally appearing interval (EAI) scale format was selected for the SToPS with the exception of "Neoglottal Tonicity" which is an 11-point bipolar semantic scale. The format aspect of scale development was summarised in sections 2.2.3 and 2.3.3 with respect to both laryngeal and tracheoesophageal scales. The rationale for the 0-3 scale relates to its established and effective use in the internationally accepted measurement tool for laryngeal voice quality i.e. the GRBAS rating scale (Hirano 1981). The inherent differences between laryngeal and tracheoesophageal voice and the corresponding issues pertaining to voice rating were outlined in 2.2 and 2.4. However sufficient commonality of themes was identified to justify the application of key findings from laryngeal voice quality measurement to tracheoesophageal voice. The GRBAS EAI format has been demonstrated to have good reliability (Dejonckere et al 1993; de Bodt et al 1997; Millet and Dejonckere 1998; Webb et al 2004). Its superior reliability in relation to longer more complex scales has been attributed to its conciseness in parameter number and scale length (Webb 2005 p136). This issue of balancing scale sensitivity with maximum clinical utility was discussed in section 2.3.3. This concluded that other scale formats may offer more sensitivity but would be difficult to implement in clinical situations. The exception to this 4 point EAI scale is the longer, bipolar scale format for "Neoglottal Tonicity" (Figure 10).

Figure 10. "Neoglottal Tonicity" scale.

5 4 3 2 1 0 1 2 3 4 5

Hypo ← **Tonic** → **Hyper**

The rationale for this different scale for this one parameter was to aim to facilitate: a) its sensitivity and b) ease of use for the rater. The extra sensitivity offered by a zero to five scale was important because tonicity has been attributed as the major indicator of tracheoesophageal voice quality (van As et al 2003; Hurren et al 2009). The surgical and other management options can produce subtle but functionally crucial variations in tone and hence in voice quality as detailed in section 2.1. The rationale for including a bipolar format design was to enhance ease and consistency of scale use. Hypertonicity and hypotonicity are opposite end points of one continuum with a shared zero baseline of neutral tonicity. This format represents this spectrum and allows raters to see the whole tonicity spectrum on a single scale at a glance facilitating ease of use for judges. The alternative format of two separate equally appearing interval scales was felt to increase the chances of judges abstaining from committing to a tonicity rating if there was uncertainty by rating both the hyper and hypotonic scales as zero.

The final aspect of scale design is to label each scale point with an adjective. If only some scale points are marked: a) there is a rater tendency to select those with an adjective in preference to those unmarked and b) restricting labelling to scale endpoints can pull rater responses to the ends of a scale (Streiner and Norman 1995 p 37).

Parameter baselines and definitions

No investigations to date have investigated whether professional judges can rate against an internal representation of the most optimal outcome in tracheoesophageal voice quality. Other rater types would not be expected to possess such psycho-acoustic ability as they are not experienced in tracheoesophageal voice. Some parameters must obviously relate to that of optimal tracheoesophageal voice as they measure features that do not occur in either normal or dysphonic laryngeal voice i.e. Wetness and Stoma Noise. The importance of allocating specific baselines was highlighted in section 2.3.1. This is especially important for the global parameters where it is difficult to ascertain a baseline zero score even for non-dysphonic laryngeal speakers. The allocated baselines for each of the SToPS' parameters will be clearly defined with the rationale for selection in section 3.2 but a brief overview is provided in Table 5.

Table 5. Parameter baselines for the first draft of the SToPS.

Parameter	Scale Format	Baseline
Overall Grade	**0-3 EAI**	Optimal tracheoesophageal voice
Impairment of Social Acceptability	**0-3 EAI**	Optimal tracheoesophageal voice
Impairment of Intelligibility	**0-3 EAI**	Normal intelligibility for a laryngeal speaker at normal conversational volume
Strain	**0-3 EAI**	No perceived effort
Wetness	**0-3 EAI**	No audible sound of vibrating secretions in the neopharynx during phonation
Impairment of Volume	**0-3 EAI**	Normal conversational volume for a laryngeal speaker
Whisper	**0-3 EAI**	Absence of whisper quality
Stoma noise	**0-3 EAI**	Absence of stoma noise
Neoglottal Tonicity	**11 point bipolar**	Mid point of zero is neutral tone, neither lax nor tight quality
Stenosis	**0-3 EAI**	Absence of tense, strained whisper quality, resonance in a rigid narrow neopharynx

3.1.2 Preliminary clinical application of version 1 of the SToPS

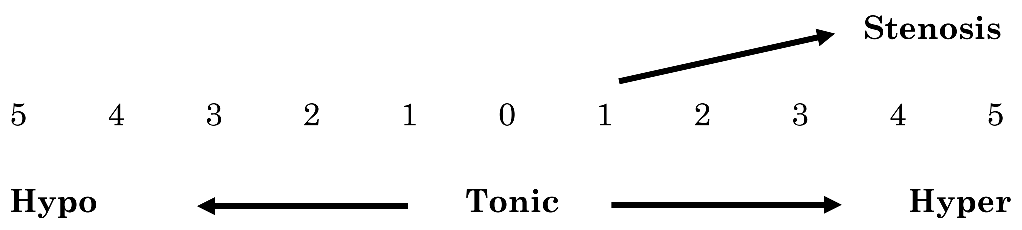
The original prototype of the scale was tested in collaboration with three experienced SLT raters. Voice stimuli were selected from audio recordings of tracheoesophageal speakers made routinely during clinical work. The audio samples consequently reflected typical clinical measurement needs and included patients who had undergone:

- standard laryngectomy with both primary muscle and non-muscle closure techniques with or without myotomy pharyngolaryngectomy surgery with jejunal free flap reconstruction;
- botulinum toxin injection/ secondary myotomy;
- change(s) of voice prosthesis type.

The three SLT‘s judged each voice stimulus independently and only then discussed their scale point allocation and the rationale for its selection. All raters described the scale as easy to use, clinically applicable and to enable the differentiation between patients and treatment effects. There appeared to be good concurrence in scale point allocation. The only exception was for “Stenosis”; all three SLT’s reported they were unable to perceptually detect mild or moderate degrees of stenosis that may co-occur with other tonicities. However they reported greater confidence in identifying the parameter in its most extreme form of an aphonic whisper as described by Blom et al (1986) and outlined in 3.1.1 (Whisperiness page 94). Raters reported it was challenging to differentiate severe hypotonicity from stenosis; both cause aphonia but hypotonicity is perceptually qualitatively distinctive. Severely hypotonic voices are associated with wet voice quality (van As-Brooks et al 2005) and the resonance is distinctive as it occurs in a dilated, voluminous resonating chamber e.g. jejunum or gastric graft. In contrast the aphonia of stenosis occurs in a rigid, narrowed neopharynx which gives a “coarse whisper quality” (Singer et al 1986) more in keeping with whispered voice quality in laryngeal speakers.

Only one change was made to the SToPS following this first clinical application i.e. “Stenosis” was removed as a separate parameter and integrated into the “Tonicity” parameter. The “Tonicity” scale thus became a bipolar scale with a third branch of Stenosis. Consequently raters are required to select one of four options; a) neutral tonicity (centre point of zero), b) the right branch Hypertonicity (1-5), c) the left branch, Hypotonicity (1-5) or d) the Stenosis branch (this is an all or nothing judgement) as shown in Figure 11. Further evidence for the change to this parameter format is the seminal work of Perry and co-workers (Cheesman et al 1986; Perry 1989; McIvor et al 1990) that was outlined in section 1.4 Figure 6. This scale is in keeping with Perry and co-workers’ seminal theory of tonicity.

Figure 11. Revised "Neoglottal Tonicity" scale to include stenosis.



There are limited studies to inform the evidence base of neoglottal tonicity and minimal research to date has investigated the sub-category of stenosis. Perry considered tonicity but not stenosis to be on a continuum but further research is required to investigate whether a stenotic continuum can be identified and even if this is established whether it can be accurately and reliably assessed.

Only one further recommendation was identified by the 3 SLT's involved in the collaborative clinical application. They reported sustained vowel voice stimuli caused more severe ratings for parameters whereas connected speech samples from the same patient were rated as less impaired. Consequently only connected speech samples were selected for the subsequent pilot study. Detailed guidance notes (Appendix A) for version 2 of the SToPS were developed with the aid of notes and comments from the collaborative session.

3.1.3 The pilot study

A pilot study was held as a 3 hour afternoon session organised by the south of England Head and Neck Oncology SLT Special Interest Group. Twenty experienced SLT's specialising in this clinical area trialled version 2 of the SToPS. None of these participants subsequently took part in the investigation of the reliability of the scale. The group simultaneously listened to twenty tracheoesophageal audio and audio visual voice stimuli selected to represent the range of parameters and scale points reflected in the SToPS. Group members rated each voice individually then discussed their judgements as a group before moving onto the next voice sample. A written summary of the group's discussion and comments was completed at the time of discussion of each voice stimulus. The pilot group reported that the SToPS and the guidance notes were easy to use and would meet the needs of clinical practice in SVR rehabilitation and outcome measurement. However the pilot group felt several key aspects were not included in the SToPS and these omissions prevented them from adequately summarising the speakers' tracheoesophageal voice outcome. This led to several parameters being added after this pilot. Details of these four additional parameters with a summary of the rationale for their inclusion are now considered.

Impairment of articulatory precision

The pilot SLT's sought guidance regarding how to score articulatory factors as they perceived these to be influencing scores for Social Acceptability and Intelligibility. Factors that were considered by the pilot group relate to the following aspects observed in clinical practice:

a) surgery i.e. unilateral hypoglossal nerve (XIIth cranial nerve) paresis or partial base of tongue resection to ensure tumour clearance,

b) habitually decreased articulatory pressure and precision and/or reduced lip and tongue range of movement patterns that occur due to articulatory style in the absence of pathology.

The pilot SLT group discussion concluded that there is a clinical and research requirement to measure whether articulatory issues are affecting

Social Acceptability and Intelligibility parameters as this would permit differentiation from factors that occur at the phonatory source i.e. neoglottal level. Discussion concerned how articulatory precision exists on a continuum from habitually decreased articulation precision and range of movement to severe dysarthria. The baseline and scale points subsequently allocated to this parameter will be detailed further in section 3.2.

Positive paralinguistic features, reading ability and accent

The pilot SLT's generally agreed that ratings of global parameters could be positively influenced by certain speaker attributes that were not included in the SToPS and did not relate to decreased articulation skills as outlined above. The group discussed two voice stimuli characterised by hypotonicity and wet voice quality. One speaker had exceptionally precise articulation, prosody, a RP accent and advanced reading aloud skills, including use of pause to enhance meaning, whereas the other had a marked local accent and none of the positive features. The group felt SLT raters may be inadvertently measuring parameters that do not relate to SVR outcome. Furthermore it was postulated that members of the local community (naïve raters) would be more likely to be thus influenced as they are not trained to differentiate the relevant aspects of the speech and voice signal. Panel members were aware naïve raters were to be recruited as part of this thesis. Discussion focussed around accent and how voice stimuli obtained from subjects reading aloud vary according to the skill of the reader. On the basis of this discussion three extra parameters were subsequently added to the SToPS i.e. Positive Paralinguistic Features, Reading Ability and Accent with the rationale of aiming to encapsulate pause, intonation, precise articulation i.e. diction, advanced skill in reading aloud and accent.

3.2 An overview of the final version of the SToPs following the pilot study

This section will provide a comprehensive overview of the parameters selected for the final version of the SToPS prior to the initiation of tests regarding its reliability and validity. The final rationale for each parameter will be outlined with reference to the literature. This is to provide more evidence regarding content validity in addition to the clinical consensus achieved from the previous pilot. The baseline and definitions for each parameter will also be summarised. All parameters have a baseline of zero as the most optimal score with scale points increasing as the attribute being measured becomes more severely impaired. The sole exception is Positive Paralinguistic Features which has a zero point to indicate the absence of these positive features. The parameters were sub-divided into Section A for those that measure voice quality and Section B for those that assess aspects other than voice quality. The final version of the SToPS is included as Appendix B. The definitions for scoring each scale are included in Appendix A "Guidance notes for using the Sunderland Tracheoesophageal Perceptual Scale".

3.2.1 Section A. Voice quality parameters

Parameter 1 Overall voice rating scale

Definition: This parameter is defined as the overall impression of alaryngeal tracheoesophageal voice; specific parameters that make up the overall score are not specified.

Scale Description: The scale baseline at zero relates to optimal tracheoesophageal voice. The 0-3 equally appearing interval scale includes the adjectival scale markers Excellent (0), Good (1), Adequate (2) and Poor (3).

Rationale: This scale is significant to SVR rating as it represents a global parameter rating and is a key outcome measure of the overall impression of the voice. As this thesis also aims to include other rater types it is crucial to include this most basic type of overarching parameter. The baseline was selected on the basis that tracheoesophageal speakers cannot achieve a baseline of normal laryngeal voice quality.

Parameter 2 Tonicity

Definition: The psychoacoustic impression of neoglottal tone.

Scale Description: The format is a bipolar, semantic 11 point equally appearing interval scale. The scale mid-point of zero corresponds to neutral tonicity i.e. the psychoacoustic impression of a neoglottis that is neither hyper/hypotonic nor stenosed. Hypotonicity and Hypertonicity are rated 1-5 on either side of a mid-point zero score.

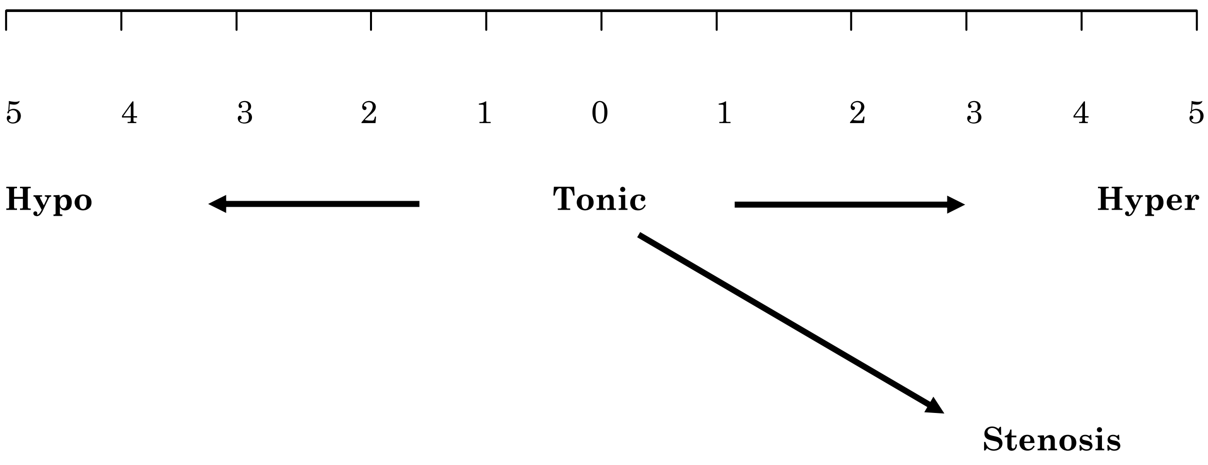
Rationale for selection: Neoglottal tonicity is a major determinant of alaryngeal voice as summarised in section 3.1.1. This parameter uses a 0-5 format in contrast to all other parameters which use 0-3. This is to enhance scale sensitivity as this is a key outcome measure for surgical and other interventions. The bipolar scale format was chosen for ease of clinical use as and to facilitate raters committing to a psychoacoustic assessment of Tonicity as outlined in section 3.1.1.

Parameter 2a Stenosis

Definition: The psychoacoustic impression of no audible neoglottal vibration within a rigid, fibrosed neopharynx.

Scale Description: Scale format is a binary rating i.e. a present or absent rating as opposed to 0-5 for the other tone types (see Figure 11). It is only judged to be present in its most extreme form where no other tonicities are perceived to co-occur. The scale format was altered following the preliminary clinical application of the SToPs as detailed in 3.1.2. The scale is a separate branch from neutral tonicity.

Figure 12. SToPS tonicity rating scale.



Rationale: Stenosis is one of the five tonicity sub-types indentified in the literature (Cheesman et al 1986; Perry 1989; McIvor et al 1990). This scale format is in keeping with Perry's theory of tonicity (see section 1.4 Figure 6). There is very limited evidence regarding the effect of stenosis on voice quality.

Parameter 3 Strain

Definition: The psychoacoustic impression of effort required to produce voice.

Scale Description: The zero baseline refers to an absence of the perception of strain. Scale points 1-3 correspond to mild (1), moderate (2) and severe (3) strain respectively.

Rationale for selection: The reasons for including this aspect of tracheoesophageal voice were detailed in section 3.1.1. Strain occurs due to hypertonicity but also due to the severe hypotonicity that occurs in grafts reconstructions following pharyngolaryngectomy (McAuliffe et al 2000). It can also occur if the neopharynx is tightly stenosed and due to poor speaker technique if the stoma is occluded with excessively digital pressure that manually compresses the neoglottis (Perry 1989 p100). The baseline of no perceptual strain would be expected to be achieved in neutral and mild hypotonicity.

Parameter 4 "Wetness" (gurgliness) of voice quality

Definition: The psychoacoustic impression of secretions vibrating in the neopharynx during voicing.

Scale Description: The baseline of zero represents no audible sound of vibrating secretions in the neopharynx during phonation. Scale points 1-3 correspond to mild (1), moderate (2) and severe (3) wetness respectively.

Rationale for selection: This perceptual feature has been noted in many previous studies and confirmed on videofluoroscopy (van As-Brooks et al 2005) and stroboscopy (Dworkin et al 1999) and high speed digital imaging (van As et al 1999). There is evidence wetness relates to the absence of a neoglottic bar which occurs in hypotonicity and stenosis which permits oesophageal secretions to be regurgitated on voicing (van As-Brooks et al 2005).

Parameter 5 Impairment of volume

Definition: The psychoacoustic impression of reduced volume of the voice.

Scale Description: The baseline of zero refers to conversational volume of voice that is judged to fall within the same limits as would be expected for normal laryngeal speakers. Ratings of 1-3 refer to mild, moderate and severely impaired volume respectively. A rating of 3 is reserved for voice that is whisper only. The scale was not designed to measure excessive volume.

Rationale for selection: This parameter aims to investigate whether habitual volume is sufficiently loud in a one to one quiet setting and not the range of volume that can be achieved. Evidence from the literature regarding this parameter was summarised in section 3.1.1. Studies have demonstrated that some patients are clearly able to achieve the baseline of normal laryngeal volume but some have impaired volume (Delsupehe et al 1998).

Parameter 6 Impairment of social acceptability

Definition: The rater's impression of how socially acceptable they perceive the speaker to be.

Scale Description: The baseline of zero relates to the optimal level that can be achieved for a tracheoesophageal speaker. The equally appearing interval scale points 1-3 correspond to mild (1), moderate (2) and severe (3) impairment of acceptability respectively.

Rationale for selection: There is some evidence that tracheoesophageal speakers have impaired Social Acceptability and that some voice qualities are more acceptable than others as discussed in section 2.4.4. The pilot study showed that some tracheoesophageal speakers are judged to have this level of acceptability.

Parameter 7 Whisper

Definition: The psychoacoustic impression of whisperiness in the voice quality.

Scale Description: The baseline of zero refers to the absence of whisper quality. The equally appearing interval scale points 1-3 correspond to mild (1), moderate (2) and severe (3) whisperiness respectively. A score of 3 is reserved for total aphonia.

Rationale for selection: There is evidence that whispery voice is a common feature of alaryngeal voice quality (Blom et al 1986; Perry 1989 p100; van As-Brooks et al 2005). Some tracheoesophageal speakers do not have this parameter as part of their voice quality e.g. those with a hypertonic neoglottis where there is closure of the neoglottis (van As-Brooks et al 2005).

3.2.2 Section B. Parameters not related to voice quality

Parameter 8 Impairment of intelligibility

Definition: The rater's perception of difficulty in understanding the subject's speech in relation to what would be expected from a normal laryngeal speaker, in a one to one speaking situation with no background noise.

Scale Description: The baseline of zero refers to the intelligibility of a normal laryngeal speaker in a one to one speaking situation with no background noise. Scale points 1-3 are assigned to mild, moderate and severe impairment of intelligibility respectively.

Rationale for selection: The baseline in relation to normal laryngeal speakers was selected because there is evidence this level of intelligibility can be achieved by tracheoesophageal speakers (Finizia et al 1998) but not all speakers attain this level (McAuliffe et al 2000). Current theories of the aetiology for impaired intelligibility were outlined in 3.1.1 and concern non-optimal neoglottal tone (Nieboer et al 1988; Jongmans et al 2003; Jongmans et al 2010) poor dentition/articulation (O'Leary et al 1994), low volume (Clark 1985) and stoma noise (Perry 1989).There is some evidence that SVR speakers' intelligibility can be reliably measured with an equally appearing interval scale (Cullinan et al 1986; van As et al 2003; Vlantis et al 2003) as highlighted in section 2.3.

Parameter 9 Stoma blast

Definition: The psychoacoustic impression of air escaping from the stoma during tracheoesophageal voicing.

Scale Description: The baseline of zero refers to an absence of any stoma noise. The equally appearing interval scale points 1-3 correspond to mild (1), moderate (2) and severe (3) stoma blast respectively. Even a relatively mild and occasional occurrence during the sample is scored as 1 (mild).

Rationale for selection: This is a well documented feature of alaryngeal speakers (Bridges 1991a; Deschler et al 1994; Delsupehe et al 1998; Dworkin et al 1999; Finizia et al 1999; McAuliffe et al 2000). More importantly patients reported this to be the most negative feature of their tracheoesophageal voice (Silverman and Black 1994). However not all patients exhibit extraneous noise from the stoma and consequently some will achieve the baseline of absence of this feature. It occurs solely in patients with a stoma and unlike other laryngeal dysphonic qualities will be unfamiliar to non-professional listeners and draws attention to the stoma. Consequently it has the potential to relate to poor acceptability judgements especially for naïve listeners. For this reason the SToPS rater guidance specifies that even mild intermittent stoma blast should be rated as a score of 1 (mild).

Parameter 10 Impairment of fluency

Definition: The psychoacoustic impression of slow rate or irregular flow of speech. More precisely, the number of words per minute or per single cycle of egressive airstream known as a breath group or a problem with the periodicity of the flow of speech, where rate of speech is irregular and may be interspersed by pauses or blocks. This scale should only be used in the absence of any concomitant speech and language pathology such as stammering or dysphasia which should be treated as separate issues.

Scale Description: The baseline refers to fluency that would be expected of a normal laryngeal speaker. The scale points increase in relation to increasing impairment. A score of 1 represents a mild reduction in fluency compared to a normal laryngeal speaker, with 2 to moderate at 5-10 syllables per breath group and 3 to severe at 5 or less syllables per breath group.

Rationale for selection: The baseline zero scale point of normal laryngeal fluency can be attained by some tracheoesophageal speakers (Pindzola and Cain 1988). However fluency problems are a well documented feature of tracheoesophageal speech (Perry 1989 p23). This was discussed in section

3.1.1. Perry (1989 p23) reported a strong relationship between reduced fluency of speech and increased neoglottal hypertonicity. Other factors such as poor stoma occlusion (Finizia et al 1999), air wastage, strained voice quality and poor skills in reading aloud may also have a detrimental impact on Fluency; for this reason a "Poor Reader" parameter was added to the scale post pilot as discussed above and in Parameter 14 below.

Parameter 11 Impairment of articulatory precision

Definition: The psychoacoustic impression of a reduction in the pressure and precision of articulation.

Scale Description: The baseline of zero refers to articulation precision that is within the normal range for non-laryngectomised speakers. This scale aims to encompass the full range of articulatory issues starting at the scale point of 1 that refers to patients who have poor precision pressure and range of articulation patterns either as their habitual articulatory setting and/or because they are adentulous patients. The scale point of 2 relates to markedly reduced articulation range and precision or mild dysarthria where intelligibility starts to be affected; at this level articulatory issues would be expected to be apparent to naïve listeners. Scale point 3 is reserved for moderate to severe dysarthria that would cause marked intelligibility issues even if the subject were still a laryngeal speaker.

Rationale for selection: This parameter was added after the pilot study as detailed in 3.1.3 because participating expert SLT's reported that reduced articulatory precision appeared to have a negative effect on their global parameter ratings. This issue was also highlighted by O'Leary et al (1994) when raters reported that clarity of articulation influenced acceptability judgements; the authors of this study concluded further studies were necessary to investigate the relative contributions of articulation and voice.

Such parameters are measured in laryngeal voice in the Vocal Profile Analysis (VPA) (Laver et al 1991) outlined in 2.2.2. Scale format selected for the SToPS covers the spectrum of severity in articulatory precision as found in the VPA but with the key difference of range and precision for all articulators being assessed in one parameter. This is in contrast to the VPA which assesses lip, tongue and jaw articulation separately. Only one study has investigated the reliability of the VPA articulation parameters (in laryngeal speakers) and reported limited agreement (Webb 2005 p125). The SToPS parameter aims to investigate whether agreement can be improved with this simpler format.

Parameter 12 Positive features (paralinguistics/diction)

Definition: The psychoacoustic impression of features of diction, intonation, stress and pause that are perceived as superior to the speaker's peer group.

Scale Description: The baseline of zero score relates to a neutral level where diction, intonation, stress or pause features are judged to be typical of the local population. The scale point of 1 relates to positive features that are above average in comparison to laryngectomy peers, with prosody judged to be present. Scale point 2 is assigned if there is excellent phrasing, diction and intonation. The end point of 3 is reserved for outstanding features that would be present in professional media presenters with normal or almost normal intonation present. The scale is structured so ascending points relate to more positive features with scale points 1-3 corresponding to good, excellent and outstanding adjectival markers. This is in contrast to the other parameters that the least positive scores as 3.

Rationale for selection: This parameter was added after the panel pilot study as detailed in section 3.1.3. The rationale for inclusion was due to the panel suggesting a tendency for these features to have a positive influence on global parameter scores. Alaryngeal voice is significantly different to laryngeal and has more potential to be socially unacceptable. Furthermore a key aim of this thesis is to compare professional and naïve rater judgements. The pilot SLT panel postulated Naïve raters were more likely to be influenced by such factors. The inclusion of this parameter was considered to permit this factor to be investigated further should the results of this thesis show poor agreement for naïve and professional judges.

Several aspects were included in just one parameter i.e. diction, intonation and prosody to reduce the already considerable number of parameters in the SToPS. The articulatory aspect i.e. referred to as “diction” in this parameter differs from articulatory skills covered by Parameter 11 (Impairment of Articulatory Precision). Parameter 11 aims to assess the degree to which articulation differs negatively from a normal baseline whereas in contrast to Parameter 12 which is designed to measure a speaker’s positive articulatory attributes in relation to this normal baseline. Such articulatory aspects are routinely considered by SLT’s working in the field of laryngectomy. It is crucial to assess both positive and negative variations in articulation prior to surgery as it enables the SLT; a) to predict a patient’s intelligibility in the immediate post-operative recovery period when verbal communication is limited to silent articulation and b) to determine therapy to ensure the patient acquires the articulatory precision and pressure required for optimal intelligibility with an electronic larynx. As this parameter has the potential to be more subjective more specific guidance was devised to attempt to define this category as outlined in the SToPS Guidance Notes (Appendix A). This scale’s format differs from the other 0-3 parameters in that scale points increase the more positive the attribute becomes. This was judged to be an easier format for raters.

Parameter 13 Accent

Definition: The psychoacoustic impression of the speaker having no regional accent.

Scale Description: The zero baseline definition is for speech that is perceived to be Received Pronunciation (RP) English. The equally appearing interval scale points 1-3 relate to mild, moderate and marked presence of an accent respectively.

Rationale for selection: This parameter was added to the SToPS following the SLT panel pilot study as outlined in 3.1.3. This was because the panel described how the presence of an RP accent appeared to positively influence ratings of global parameters. The potential social penalty or prestige of accents has been reported in the field of sociolinguistics (Giles and Sassoon 1983). There is some evidence that accent can be rated with a 9 point equally appearing interval scale by student SLT's (Mackey et al 1997) but judges only achieved good intra but not inter rater agreement. This thesis is the first investigation of accent in relation to tracheoesophageal voice. There is some limited observation that expert SLT's experienced more difficulty in rating laryngeal speakers if they had an accent rather than RP (Webb 2005 p132).

Parameter 14 Reading ability

Definition: The ability to read aloud correctly, fluently and without hesitation.

Scale Description: The zero baseline refers to no difficulty in reading aloud. Scale points 1-3 represent mild, moderate and severe problems with this parameter.

Rationale for selection: This parameter was added in response to the pilot study panel's suggestion that good ability in reading aloud was potentially increasing global parameter scores in conjunction with Accent and Positive Paralinguistic Features. A further rationale relates to the aim of permitting further investigation of any influence of reading ability on Overall Grade if considerable variation was found between Naïve and SLT/ENT raters. It was felt that poor reading ability may have more potential to influence naïve raters' judgements as they are unfamiliar with the differentiation of voice and speech in overall type scales. Reduced speed of reading due to lack of literacy or confidence in reading aloud also has the potential to have a negative effect on rating of the Impairment of Fluency parameter.

3.3 The development and design of a rating scale for naïve listeners

3.3.1 Scale design

A key aim of this thesis is to compare and analyse correlations in alaryngeal voice assessment from five different rater perspectives: SLT, ENT, Naïve, Patient self-rating and Carer rating of their friend/relative. There is an essential requirement for some commonality between the scales and scale design must be tailored to facilitate non-professional judges' perspectives. Consequently the rating scale for the naïve rater group was designed to mirror the professional scale wherever possible, but in a simpler format. The adaptation to scale format includes utilising an adjectival rather than equally appearing interval scales. Parameter nomenclature was also simplified. Tables 6 and 7 summarise the rating scales for the four different rater groups.

The first draft of the naïve rating scale was designed according to the rationale specified above. The pilot scale with details of scale format and rationale for selection is outlined below.

Table 6. A comparison of the rating scales.

Expert Scale	Naïve Scale	Patient scale	Carer Scale	Justification for selection and design	Anatomical Structure
Overall Rating 0-3 with 4 point adjectival anchors	Overall Rating 4 point adjectival	Overall Rating 4 point adjectival	Overall Rating 4 point adjectival	Simple common scale required to compare rater groups. Overall Grade most reliable measure in laryngeal voice assessment.	Neoglottis, Stoma Occlusion, Articulation
Social Acceptability 0-3 with 4 point adjectival anchors	Social Acceptability 4 point adjectival	Social Acceptability 4 point adjectival	Social Acceptability 4 point adjectival	Simple common scale required to compare rater groups. Commonly used parameter in SVR.	Neoglottis, Stoma Occlusion, Articulation
Impairment of Intelligibility 0-3 with 4 point adjectival anchors	Hard to understand Yes/No	Intelligibility 4 point adjectival	Intelligibility 4 point adjectival	Noted in literature as a feature of SVR voice	Neoglottis, Stoma Occlusion, Articulation
Impairment of volume 0-3 with 4 point adjectival anchors	Not loud enough	Volume of voice in comparison to need	Volume of voice in comparison to need	Noted in literature as feature of some SVR speakers. Aphonia occurs due to stenosis and severe hypotonicity.	Neoglottis
Impairment of Articulatory Precision 0-3 with 4 point adjectival anchors	Hard to understand Yes/No	Intelligibility 4 point adjectival	Intelligibility 4 point adjectival	Noted in literature, recommended at the expert pilot study.	Articulation.
Tonicity, Stoma Noise, Fluency, Poor Reader, Positive Paralinguistic Features, Accent Tonicity 11 point Bipolar Scale or Stenosis Present/Absent All other parameters 0-3 with 4 point adjectival anchors	No equivalent	No equivalent	No equivalent	For professional raters only as complex constructs. Evidence for Tonicity, Stoma noise and Fluency from the literature. Remaining parameters recommended at the expert pilot study.	Neoglottis, Stoma Occlusion, Articulation.
Whisper 0-3 with 4 point adjectival anchors	Whispery Yes/No	Whispery Yes/No	Whispery Yes/No	Evidence from the literature that an absence of neoglottic vibration is a feature of SVR voice and causes aphonia.	Absent neoglottis due to a stenosed and fibrosed neopharynx or due to neoglottal hypotonicity (+/- graft repair)
Strain 0-3 with 4 point adjectival anchors	Strained Yes/No	Strained Yes/No	Strained Yes/No	Noted in literature as a feature of SVR voice	Hypertonicity, Hyperocclusion of the stoma due to poor technique
Wetness 0-3 with 4 point adjectival anchors	Gurgly Yes/No	Gurgly Yes/No	Gurgly Yes/No	Noted in literature as a feature of SVR voice	Hypotonicity, Stenosis, Extensive surgery with graft repair.

Table 7. Common scale items to all rater groups - format of scale design in relation to parameter and rater type.

	Expert Raters	Naïve Raters	Patient/Carer Raters
Overall Rating	0-3 with 4 point adjectival anchor markers	4 point adjectival	4 point adjectival
Social Acceptability	0-3 with 4 point adjectival anchor markers	4 point adjectival scale	4 point adjectival scale
Intelligibility	0-3 with 4 point adjectival anchor markers	Modified terminology Yes/no response	4 point adjectival scale
Volume	0-3 with 4 point adjectival anchor markers	Modified terminology Yes/No response	Modified terminology 4 point adjectival rating scale
Whispery	0-3 with 4 point adjectival anchor markers	Yes/ no response	Yes/ No response
Wetness	0-3 with 4 point adjectival anchor markers	Modifed terminology Yes/no response	Modifed terminology Yes/No response
Strained	0-3 with 4 point adjectival anchor markers	Yes/no response	Yes/No response

Parameter 1 Overall voice rating

Definition: No formal definition was provided to the Naïve Raters. However a written information sheet provided for this group (Appendix C) requested raters imagine how they would feel about having a voice like the speaker or if the voice was that of their partner or close relative.

Scale format: An adjectival scale with continuous responses in four steps; Excellent, Good, Adequate, Poor.

Rationale for selection: This is a common scale item for all five rater types included in this thesis. This global parameter of overall impression was chosen as the simplest type of parameter for non professional raters to assess voice stimuli. The rationale of omitting a specific baseline or definition was to avoid influencing the judgements from the naïve perspective. The basis for asking naïve judges to consider the voice being that of a close relative or their own voice was to encourage them to focus on how the person may be perceived in real life settings rather than this being a perfunctory task of assessing serial voice stimuli. It is hypothesised that the Naive baseline will differ from that of professional raters as this group will not have the professionals' internal reference point of optimal tracheoesophageal voice.   There is also the potential for these untrained raters to be unable to differentiate voice from the whole speech signal and hence to be influenced by factors unrelated to the tracheoesophageal voice outcome. These factors were detailed in section 3.1.1 above (Parameters 11-14). The Naïve Rater Scale format differs from the professional version (See Table 7) although the SToPS scale format included identical adjectival markers under the 0-3 scale points. The rationale for the format change relates to findings from the most comprehensive Naïve rater tracheoesophageal voice perception investigation to date (van As et al 2003) as previously discussed in section 2.4. Van As et al reported naïve listeners' scores clustered at the severe end of bipolar scales. Consequently the Naïve

Rater Scale was designed with the hypothesis that naïve judges may find an adjectival scale simpler to use as clinical observation appeared to indicate that untrained people commonly use adjectives not scale point references to comment on voice quality in day to day situations.

Parameter 2 Social acceptability

Definition: No formal definition was provided to the Naïve raters. However written guidance outlined that this parameter may or may not be the same as the Overall voice rating (Parameter 1) and asked raters to reflect on how they would see others reacting to this voice quality and whether it would be pleasing or unpleasant to listen to.

Scale format: An adjectival scale with continuous responses in four steps; Excellent, Good, Adequate, Poor.

Rationale: This parameter is a scale item common to all three rating scales developed in this thesis. This is a key aspect of tracheoesophageal voice outcome. It has been suggested that only this group can judge the social penalty of being a tracheoesophageal speaker (Eadie and Doyle 2005a).Variations of this parameter have been included in numerous studies as summarised in 2.3. The rationale for scale format is identical to the issues summarised for Parameter 1 outlined above. The baseline reference point is unclear for this rater group but it is hypothesised to differ from that of professional raters who judged acceptability in relation to optimal tracheoesophageal voice. Again Naïve listeners would not have this internalised reference point. The guidance note requested listeners to reflect on how the voice may be perceived by others and aimed to encourage rater judgements to represent their perspective of how the general community would respond to the voice.

Parameters 3-7 Uni-dimensional parameters (Gurgly, Strained, Not loud enough, Whispery, Hard to Understand)

Scale definitions: No written or verbal definitions were provided.

Scale format: Binary scale format requiring only a yes/no response to indicate presence or absence of the quality.

Rationale: These parameters were selected on the basis that they correspond to some of the uni-dimensional parameters included in the professional version of the SToPS i.e. Wetness, Strain, Whispery, Impairment of Volume and Impairment of Intelligibility respectively. This is again to provide the commonality of scale that will enable rater type comparisons to be carried out. Nomenclature was simplified to facilitate Naïve raters' understanding of these constructs as outlined in Table 8.

Table 8. Professional and Naive rater parameter terminology.

Expert rater term	**Naïve rater term**
Wetness	Gurgly
Strain	Strained
Impairment of volume	Not loud enough
Whisper	Whispery
Impairment of intelligibility	Hard to understand

The remaining seven parameters included in the professional version of the SToPS i.e. Tonicity, Stoma Noise, Impairment of Fluency, Impairment of Articulatory Precision, Positive Paralinguistic Features, Accent and Reading Ability were not included in the Naïve rater scale. The rationale for their exclusion relates to: a) the consideration that they were too complex constructs for this group to be able to identify from the tracheoesophageal stimuli and b) the need to decrease the number of parameters for this group.

These again relate to the work of van As et al (2003) as they observed that Naïve judges could not use the scales to differentiate between voice qualities i.e. they rated all uni-dimensional parameters as severe and were not able to focus on specific parameters once they heard a “deviant” voice (van As et al 2003). This finding was highlighted in section 2.4. Van As and co-authors did not consider their findings may relate to the large number and complexity of the bipolar scales raters were required to judge per stimulus. The naïve modified scale from the original SToPS consequently was designed with a simplified scale format requiring a binary yes/no response for only five parameters with modified terminology. A key aim of the naïve version of the SToPS was to ascertain whether this group are able to detect qualities with the simpler format. No written or verbal definitions were provided for the parameters to establish whether they could assess these parameters from an internal point of reference.

3.3.2 Pilot of naïve listeners

Three acquaintances of the author were recruited to pilot the naïve listener scale. None of the raters had any experience of laryngectomy or speech, language or voice impairment. Twenty SVR speaker voice stimuli were played to the panel from a minidisc stereo system. All listeners heard the voice samples simultaneously and rated the voices using the Naïve Rating Scale outlined above. The author noted the pilot volunteers allocated the same score on the overall severity scale for a speaker with neutral tonicity, reduced precision and pressure of articulation and a marked accent as for another patient with a hypotonic, gurgly voice due to jejunum graft. The latter speaker had exceptional precision and pressure of articulation, no accent and good use of phrasing and pause i.e. seemingly reflecting the skills included in the parameter Positive Paralinguistic Features. The naïve raters did not demonstrate overt awareness of this factor. However it appeared to indicate some tendency to base their judgements on variables between patients that do not relate to SVR outcome. This discrepancy in skill and focus could have the potential to affect the outcome of the overall scale, with either marked mismatch in rating between professional and naïve raters or both rater types allocating “Good” rating to markedly hypotonic voices. This provided more evidence to aim to investigate the potential effect of such factors by including the parameters Accent, Reading Ability, Positive Paralinguistic Features, Articulatory Impairment in the professional version of the SToPS.

Another key issue from the pilot concerned one rater who reported discomfort when marking the Overall rating as “poor”. She expressed a desire to give patients a higher grade due to sympathy and reported it felt “unkind” when they were clearly struggling to communicate.

3.3.3 The post pilot naïve rater scale

As the naïve listeners reported no specific concerns or difficulty in using the scale, the version outlined in section 3.2.1 was deemed suitable to investigate inter and intra rater agreement with a new panel of naïve judges. The Naïve Rater Scale is included in Appendix D. The only post pilot change was to issue more specific guidance for raters to aim to counteract any potential bias towards rating patient voice stimuli more superiorly due to sympathy (see Appendix C Naïve rater guidance). This aspect is discussed in more detail in section 4.3.2.

3.4 The development and design of a rating scale for a) tracheoesophageal speakers and b) carers

This rating scale is designed to investigate both the patient’s self rating of their voice and the carer’s perspective of their friend/relative’s voice in a face-to-face interview with the researcher. Table 7 outlined different formats for the different rater type scales used in this thesis. The rationale for interviewing patients and carers rather than allowing them to hear a pre-recording of their voice relates to clinical experience consistently showing that patients feel their voice sounds worse from an audio recording than they anticipated. Furthermore they have ongoing perceptions of their own voice (or their relative's voice) and do not need to hear a stimulus to comment upon it.

The parameters and scale format selected are outlined below. The patient and carer scales are identical with the exception of the pronouns used. The asterisk marks the pronoun that was changed to "your relative/friend's voice" for the carer scale.

Parameter 1 Overall rating

Definition: No definition was provided for the raters.

Scale format: A statement followed by an adjectival scale with continuous responses in four steps i.e. Overall I would say my* voice is: Excellent, Good, Adequate, Poor.

Rationale: This parameter was selected to ensure some commonality of scale items between all five types of judges to fulfil the research aim of comparing rater type perspective. No definition was provided to prevent patients and carers being influenced as to which perspective they are expected to rate the voice e.g. normal laryngeal voice, the voice quality immediately prior to surgery, the voice outcome in relation to expectation, the alternative of mortality due to the cancer. This issue of perspective was discussed in section 2.4.1 and 2.4.2. The four point adjectival scale is identical to that of the Naïve Rater Scale. The rationale relates to the premise that this format is more in keeping with the clinical experience of the author; patients and carers have been observed to spontaneously give semantic descriptors of the tracheoesophageal voice quality in clinical settings but never numerical values.

Parameter 2 Social acceptability

Definition: A verbal definition was provided after the statement in the scale format below i.e. "this means how you think other people feel about *your voice*, especially people you meet who are not close members of your family". The carer version was identical except for the italicised words which were amended to "your relative's voice".

Scale format: A statement followed by an adjectival scale with continuous responses in four steps i.e. The social acceptability of my* voice is:
Excellent, Good, Adequate, Poor.

Rationale: This parameter was selected as a common scale item with the SToPS and Naïve Rater Scale. It is a key item as it reflects how the patient or carer feels the tracheoesophageal voice is perceived by others outside the family. The rationale for scale format is identical to that described in Parameter 1 above.

Parameter 3 Volume

Definition: The volume of voice in comparison to the patient's needs.

Scale format: A statement followed by an adjectival scale with continuous responses in four steps i.e. The volume of my* voice compared to my needs is:
Excellent, Good, Adequate, Poor

Rationale: This parameter measures the patient and carer's perception of the loudness of voice in comparison to need with the same four-point adjectival continuous scale. This aim of this scale point is to encapsulate how the patient functions on a day to day level in his/her environment. This is assumed to be related to lifestyle and consequently may vary from patient to patient. A self-rating assessment is extremely important as the subject is providing feedback about functional voice in everyday speaking situations.

Parameter 4 Intelligibility

Definition: No definition was provided.

Scale format: A statement followed by an adjectival scale with continuous responses in four steps i.e. How would you rate the intelligibility of your* voice? Excellent, Good, Adequate, Poor.

Rationale: This was included as a common scale item for all three scale types developed in this thesis. It is crucial as it reflects the perception of the patient and carer's view of intelligibility in the wider world outside of the research setting. This parameter aims to measure a different factor to volume; it is possible for a voice to be loud yet difficult to understand. However there is likely to be a relationship between the ability to increase volume and intelligibility in louder environments. Scale format is again different to the professional version with the same rationale outlined for Parameter 1. The rationale for including a four point scale for Patients and Carers but not for Naïve raters reflects the task difference. Patients and Carers are required to rate only their own or their friend/relative's voice whereas Naïve judges rated the whole cohort of patient voice stimuli in succession from an audio recording. Patient and Carers were consequently deemed not to be at risk of parameter overload and able to provide information on this key aspect of outcome.

Parameters 5 – 7 Uni-dimensional Parameters (Gurgly, Strained, Whispery)

Definition: No definitions were provided.

Scale format: A series of questions followed by yes/no i.e. Would you say your* voice is: Whispery, Gurgly, Strained?

Rationale: These were chosen to reflect the remaining parameters that were included in the Naïve rater scale (see Table 7). This is to determine the perspective of the Patient/Carer and compare these to the Naïve and SLT/ENT judgements. No definitions were provided to enable patients to judge these from their own internal reference.

3.5 Summary

This chapter has detailed the preliminary stages of the first aim of the thesis outlined in section 2.6 i.e. to devise a valid and reliable tracheoesophageal perceptual rating scale (the SToPS) for SLT and ENT judges. The scale development was summarised including a consideration of the key aspects of content validity; a comprehensive discussion of validity will be undertaken in section 6.1. A rationale was also provided for format and scale design elements that aim to optimise rater reliability. The SToPS was then modified to produce two further scales to assess Naïve and Patient/Carer evaluations of tracheoesophageal voice quality respectively. The development of the SToPS and the Naïve Rater Scale will be used in investigations to fulfil the research aims of investigating the inter and intra rater reliability of both SLT/ENT and Naïve judges. The final research aim of investigating the relationship between SLT, ENT, Naïve, Patient and Carer perspectives of SVR voice will be undertaken by comparing the perspectives of all five rater types using the three rating scales that have been developed.

Chapter 4. An investigation of the reliability of professional and naïve raters in the perceptual assessment of tracheoesophageal voice

4.1 Overview

This chapter covers research Aims 3 and 4 as outlined in Chapter 2 (2.10). The first sub-section (4.2) describes the investigation of professional raters' intra and inter rater reliability for the SToPS; the development of this scale was detailed in 3.1 and 3.2. The second sub-section (4.3), is an investigation of the use of the modified version of the SToPS for naïve raters (as outlined in 3.3.) to assess this group's inter and intra rater agreement with tracheoesophageal speakers.

In order to undertake the studies a number of patient samples were required. The method of recruitment for the patient subjects, the patient demographics and the recording of the voice samples will be described in 4.2.1. The naïve raters used the same stimuli as the professional judges.

4.2 An investigation of inter and intra rater reliability for professional raters using the Sunderland Tracheoesophageal Perceptual Scale (SToPS)

4.2.1 Introduction

This study is designed to investigate agreement of professional (SLT and ENT) raters using the scale developed in Chapter 3. The first section (4.2.2) will outline methodology and this is followed by the results and data analysis (4.2.3).

4.2.2 Methodology

Patient Recruitment

All the SVR patients within the sub-regional area covered by one cancer unit (n= 73) were sent a standard letter requesting that they consider volunteering as subjects for the study. The only exclusion criteria were a) inability to read aloud and b) presence or suspicion of persistent or recurrent cancer. Only one patient was not approached because of these criteria. Ethical approval was obtained from the Local Strategic Health Authority ethics committee (Appendix E) and research governance approval was obtained from the NHS trust prior to contacting patients. Fifty-seven patients (78%) returned a slip in a stamped addressed envelope to register interest in participating in the research. Fifty-five patients recruited subsequently attended the voice recording session. The standardised procedure for the recording is detailed in the protocol specified below. Appointments were allocated to one of six recording sessions by a Speech and Language Therapy Assistant on a “first come, first served basis”.

Patient Demographics

Patient demographics are summarised in Table 9. Fifty-five laryngectomy patients who underwent surgery by one of six surgeons at one ENT unit attended for voice recording. Fifty-one underwent total laryngectomy and four underwent pharyngolaryngectomy with free jejunal graft. Six were female and forty-nine were male. The age range was 48 to 80 years with a median of 64 years 11 months. Length of time from surgery to the audio-recording ranged from 3 months to 15 years 3 months with a median of 3 years 5 months. All patients spoke English as their first language and none had concomitant speech or language impairment prior to their laryngectomy.

Table 9. Patient demographics.

Pt No	Sex	Age	Post-op Period Yr:mth	Surgery	Patient self rating	Carer rating
1	M	62	4,5	TL	Yes	Yes
2	M	71	7,11	TPL + J	Yes	Yes
3	F	76	1,2	TL	No	No
4	M	56	3,5	TL	Yes	Yes
5	M	60	2,4	TL	Yes	Yes
6	M	56	7,7	TL	Yes	Yes
7	M	61	2,5	TL	Yes	Yes
8	M	70	4,9	TL	Yes	Yes
9	M	64	2,6	TL	Yes	Yes
10	F	52	2,6	TL	Yes	Yes
11	M	48	3,4	TPL +J	Yes	Yes
12	M	78	6,0	TL	Yes	No
13	M	70	10,9	TPL+J	Yes	Yes
14	M	66	14,3	TL	Yes	Yes
15	M	72	2,7	TL	No	No
16	M	58	0,6	TL	No	No
17	M	64	12,7	TL	Yes	Yes
18	M	58	1,1	TL	No	No
19	M	71	5,5	TL	Yes	Yes
20	M	77	7,9	TL	Yes	No
21	M	76	2,2	TL	Yes	Yes
22	M	56	2,2	TL	Yes	Yes
23	M	75	2,2	TL	Yes	Yes
24	M	71	4,1	TL	Yes	Yes
25	M	77	17,4	TL	No	No
26	M	76	6,1	TL	Yes	Yes
27	M	62	1,8	TL	Yes	Yes
28	M	56	1,3	TL	Yes	Yes
29	F	62	4,2	TL	Yes	Yes
30	M	66	0,9	TL	No	No
31	M	64	2,5	TPL+J	No	No
32	M	70	6,7	TL	Yes	Yes
33	F	57	9,8	TL	Yes	Yes
34	M	77	12,3	TL	No	No
35	M	60	6,8	TL	Yes	Yes
36	F	67	15,8	TL	Yes	Yes
37	M	63	6,6	TL	No	No
38	M	78	3,4	TL	No	No
39	M	54	3,6	TL	Yes	Yes
40	F	66	13,1	TL	No	No
41	M	66	6,4	TL	Yes	Yes
42	M	56	7,5	TL	Yes	Yes
43	M	58	7,5	TL	Yes	Yes
44	M	60	0,9	TL	Yes	Yes
45	M	51	3,4	TL	Yes	Yes
46	M	80	4,4	TL	No	No
47	M	77	0,3	TL	No	No
48	M	57	2,1	TL	Yes	Yes
49	M	67	7,9	TL	Yes	No
50	M	67	2,6	TL	Yes	Yes
51	M	61	1,0	TL	Yes	Yes
52	M	76	1,6	TL	Yes	Yes
53	M	58	0,11	TL	No	No
54	M	77	1,0	TL	Yes	Yes
55	M	67	0,9	TL	No	No

TL Total Laryngectomy, TPL = Total Pharyngolaryngectomy, J= Jejunum graft

The study required a large number of patient speech samples as a small number of voice stimuli rated by a large number of raters has been demonstrated to give erroneous statistical findings (Kreiman and Gerratt 1996). The 55 volunteers were 75% of the unit's caseload; consequently the cohort recruited was considered as likely to represent the wide range of voice qualities and degrees of severity for each parameter of the SToPS. The tonicity parameter is the longest and most complex equally appearing interval scale in the study with the potential for 11 different scale points. The inclusion of a sufficient range of speakers to cover all points on the scale was confirmed (Hurren et al 2009). The majority of the patients were born in the area covered by the sub-regional ENT unit (the maximum distance from the unit is 30 miles) and consequently speak with a local accent typical of the region. Six patients did not have a local accent; one had a Scottish accent; two had London accents, two had Received Pronunciation (RP) and one a Manchester accent.

Recording of patient voice samples

The consent sheet was read aloud by the SLT Assistant alongside each patient and signed, informed voluntary consent was obtained. The consent form specified that patients could withdraw from the study at any time in the future without this affecting their future treatment and the recordings would be destroyed. No patients requested this option. Prior to the recording four checks were carried out to make sure a representative voice sample was obtained: a) The voice prosthesis was visualised to ensure it was free of mucus secretions in the lumen and cleaned if required, b) A new adhesive stoma cover was refitted if the seal had broken, c) Patients using hands-free tracheostoma valve were asked to replace them with a heat moisture exchange filter as tracheostoma valves cause additional stoma noise (Hamade et al 2006), d) The patient was asked to confirm that the current voice was typical of their usual tracheoesophageal voice.

The above checks were to ensure a typical voice quality sample was recorded to prevent any inadvertent discrepancy between professional/ naïve and patient/carer rater scores due to these factors. This is because professional and naïve raters assessed the voice from this short audio recording whereas the patient and carers rated from an internalised judgement of their everyday speech with no audio stimulus being played. Only four patients used a tracheostoma valve and all utilised them on a part-time basis only.

Recording protocol

All recordings were carried out in a soundproof room in the Audiology Department. The room is designed to meet the standards of NHS Estates Document Health Building Note 12(3). This allows (a1) audiometric testing down to 10 dB n HL in the soundfield as defined in BS EN ISO 8253-1 (ISO, 1989). The soundfield is defined as quasi-free field and is calibrated on at least an annual basis. . A Consultant Clinical Scientist (Audiology) set up the equipment and checked its function (including recommended Stage A checks) prior to each session. A Sony Electret Condenser microphone and a Quest 2700 Type 2 Sound Level Meter (as defined in IEC 651) were attached to a microphone stand. This was positioned to a standard distance of exactly one metre between the microphone and the patient's mouth when the patient sat with his/her back against the chair. The microphone was attached to a Sony MD Walkman programmed so all samples were recorded. All recordings followed exactly the same procedure at maximum volume.

The mouth to microphone distance is greater than that specified in the majority of both laryngeal and alaryngeal recording protocols. There are no standardised recommendations for alaryngeal (SVR) speakers. As detailed in section 1.3.1 stoma noise is a feature of SVR speech and a requirement of the study is to accurately represent this sound on the audio recording in as close a manner to real life speaking situations as possible. The rationale for the microphone placement at one metre from the mouth was to allow the recording of stoma sounds (emitted from the base of the neck), simultaneously to speech, (emitted from the mouth) without one being unrepresentatively dominant. The raters are thus evaluating the recordings as if they were one metre away from the subject.

Prior to the recording session the patients were randomly allocated a voice sample number from 1-55 and this number was recorded onto a master list. This allowed patient anonymity to be preserved at all times. A separate minidisc was used to record each voice; the voice sample number was written onto an adhesive label and attached to each minidisc. Before the patient audio recording commenced the SLT Assistant spoke aloud the allocated sample number to record this onto the minidisc. A 70-decibel SPL warble tone (Frequency Modulated (FM) 1 kHz) was then recorded onto each disc as a calibration tone. FM (Warble) tones were selected in preference to steady tones to prevent problems with standing waves, where there would be variation in the levels if you move in the room.

Reading aloud a standard passage

A standard recording protocol was followed for each patient as detailed below.

"The Rainbow" passage (Fairbanks 1960) (Appendix F) was selected as the sole voice stimulus for this study. The patient was requested to read it aloud once to practice before recording took place. Any unfamiliar words were explained and practiced with the research assistant. This is a standard, phonetically balanced passage used in numerous perceptual voice rating studies to date (Webb et al 2004). Prolonged vowel voice stimuli alone have been used and recommended in the laryngeal voice disorders literature (de Krom 1994). As intelligibility and articulatory parameters are included in the SToPS a connected speech sample was essential. A previous study

(Cullinan et al 1986) demonstrated no difference between reading aloud and discourse in the rating patterns of naïve and expert judges for the intelligibility of tracheoesophageal and oesophageal voice. A further rationale for the use of a reading passage was that spontaneous speech stimuli do not control for the variation between speakers in eloquence when asked to speak into a microphone for research purposes; those with more fluency could potentially influence naïve listeners' perception rather than allowing them to focus solely on voice quality.

Editing and preparation of the voice stimuli

Once all 55 recordings were completed, the spoken voice sample number and the Rainbow passage sample for each subject were edited onto a master minidisc. The order was then randomised using the shuffle facility on the minidisc system for session 1 for raters. The order of voice samples was then shuffled once more to produce a different order for a second master disc for use during the re-test session. The master list of voice samples, mapped to patient names was only available to the research assistant who carried out the recording.

Professional rater panel recruitment

SLT's and ENT surgeons working in SVR within the northern region of England and Northern Ireland were invited to participate in the project. The criteria for selection specified that each rater must have worked in a head and neck multidisciplinary team and have experience of at least 40 tracheoesophageal speakers. This level of experience should allow raters to have encountered enough SVR speakers to establish an internal anchor reference point for alaryngeal voice quality. Twenty-two professional raters were recruited; twelve were speech and language therapists (SLT's) and ten were ENT surgeons. Rater type and experience is summarised in Table 10. SVR experience ranged from 2-25 years. Raters were further categorised within their profession as expert or non-expert according to their type of experience rather than solely on the number of years post-qualification. This study defined the Expert SLT group as raters who have additional expertise in laryngeal voice rehabilitation and assessment rather than in terms of years working within the profession or in head and neck cancer. The surgeons were referred to as expert if they were Consultant Head and Neck Surgeons who had worked in a joint clinic with SLT's. Only one surgeon from the ENT Expert group works in a Joint Voice Clinic for laryngeal problems as well as in head and neck cancer. This reflects the sub-division within the ENT profession whereby head and neck cancer surgery is a separate specialism to laryngeal disorders. For this reason the same criteria did not apply to defining the ENT Expert group. Consequently the main differential between Expert and Non-Expert ENT relates to the former group's higher exposure to alaryngeal speakers than the Non –Expert ENT group; however neither group have any form of perceptual assessment training or use such assessment in their clinical work.

Table 10. Professional rater experience.

Rater	Rater Type	Years of SVR experience
SLT1	Expert SLT	13
SLT2	Expert SLT	10
SLT3	Expert SLT	8
SLT4	Expert SLT	16
SLT5	Non-expert SLT	15
SLT6	Non-expert SLT	2
SLT7	Non-expert SLT	2
SLT8	Non-expert SLT	2
SLT9	Non-expert SLT	2
SLT10	Expert SLT	16
SLT11	Non-expert SLT	10
SLT12	Non-expert SLT	7
ENT1	Non-expert ENT	10
ENT2	Non-expert ENT	2
ENT3	Non-expert ENT	2
ENT4	Non-expert ENT	4
ENT5	Non-expert ENT	11
ENT6	Expert ENT	17
ENT7	Expert ENT	11
ENT8	Expert ENT	13
ENT9	Expert ENT	9
ENT10	Expert ENT	25

Panel Training

Each panel recruit underwent three hours training prior to rating the voices. The training programme familiarised each rater with the scale and the written guidance notes that specify definitions and grades for each parameter. (Appendix A). Training included the raters using the scale with recorded tracheoesophageal voice stimuli. Opportunity for discussion was an integral part of the training. The SLT version of the scale consisted of 14 parameters whereas the ENT version was restricted to 12 parameters. SLT raters were asked to rate two extra parameters: Accent and Reading Ability as specified in 3.2.2. The aim of including these additional parameters was to investigate whether naïve raters showed positive bias to rate voices less favourably if speakers had a marked accent or poorer ability to be fluent, skilled readers. ENT raters were not trained in these parameters as they are not trained to assess voice/speech boundaries and such additional load was considered to have the potential to negatively affect their rating of the more core parameters for the perceptual assessment of tracheoesophageal voice quality.

The panel rating sessions

Each panel member evaluated all 55 subjects' voice samples from the master disc within 2 weeks of attending the training session. Each rater received fifty-five rating forms collated into the order in which the voices would be heard from the master disc. The only information marked on each sheet was the number of the voice sample and the gender of each subject. The panel were allowed to refer to the written guidance notes for rating during the voice evaluation.

Eight SLT raters rated the voices simultaneously. The research assistant played the master discs on a Sony stereo minidisc deck; model TC –TX 373, with Dolby BNR. Raters heard each sample twice and could request a third repeat if desired. Thirty voice samples were rated in the morning in two sessions with a 30 minute break. Twenty-five samples were rated in the afternoon with a 20-minute break after the first ten samples. No discussion about the task was permitted during the rating sessions or during breaks. For practical reasons 4 SLT and all 10 ENT surgeon raters listened to the master disc individually with a standard set of Sony headphones (model MDR-XD200). Whilst this represents a subtle difference in method, there is no reason to suggest that this would have any relevant effect on the task results.

In order to measure test- retest (intra-rater reliability) the procedure was repeated one month later with the voices presented in a different random order. The format was identical to the first session. Again the 8 SLT raters judged voices simultaneously with a research assistant and the remaining

13 raters judged the voices in individual sessions. One ENT surgeon did not repeat the re-test task due to workload pressure. Raters were asked to read the rating guidance information sheet again prior to starting the re-test ratings.

4.2.3 Results and data analysis

The raw scores assigned to each voice sample 1-55 by individual professional raters in sessions 1 and 2 were entered into Cytel Studio 8. A separate database was created for each parameter. Each database was analysed with the StatXact package to calculate weighted kappa co-efficients for both intra and inter rater reliability.

The range and mean of the raw kappa scores were calculated for all twenty-two raters from both professional groups. Further more detailed analysis involved categorising the professional raters into seven groups i.e. all raters, all SLT, all ENT, expert SLT, expert ENT, non-expert SLT, non-expert ENT. The range and mean kappa co-efficients were calculated for each of the seven. Landis and Koch (1977) values were used to ascertain the strength of agreement from the kappa co-efficients; these classifications are summarised in Table 11.

The mean scores for inter and intra-rater reliability of the seven sub-groups for each parameter are summarised in Table 12 and Table 13. Similarly the range and mean scores are displayed in Forest Plots (Appendix G (intra rater and Appendix H inter rater). The comprehensive results were formatted into tables by parameter for both intra and inter rater reliability outcomes are included in Appendix I.

Table 11. Landis and Koch's (1977) values of kappa co-efficients in relation to strength of agreement.

Value of K	Strength of agreement
<0.20	"poor"
0.21-0.40	"fair"
0.41-0.60	"moderate"
0.61-0.80	"good"
0.81-1.0	Very "good"

There is no accepted level of strength of agreement that should be attained for a perceptual voice rating scales to be considered clinically useful and/ or robust enough to be employed in research. Previous studies investigating the reliability of laryngeal voice scale have suggested Landis and Koch "good" level of agreement with the boundary of 0.61 or above as being acceptably reliable to use in research and clinical situations (Hirano 1989 ; Webb et al 2004). Table 11 outlines these values of kappa co-efficients in relation to strength of agreement. However such a rigid cut off point would cause scores that fall just 0.01 below this boundary to be discounted. A very wide range of co-efficient scores were noted for the majority of the parameters in this study (Forest plots Appendices G and H). These account for marked discrepancies in rater agreement; such variations in the rater perceptions of the voice parameters can fail to be acknowledged if only mean scores are considered. With the acknowledgment of the disadvantages listed above, this thesis will mark the boundary point of a mean weighted kappa co-efficient of 0.61 and above as a "good" level of agreement. This will be fully discussed in the discussion chapter.

Intra-Rater Professional Results

Table 12. Intra rater mean weighted kappa co-efficients for professional raters.

Intra- Rater	All Professionals	All SLT	All ENT	Expert SLT	Expert ENT	Non-Expert SLT	Non-Expert 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*
Articulatory Precision	0.47	0.45	0.50	0.51	0.33	0.40	0.63*
Paralinguistic Features	0.64*	0.68*	0.58	0.69*	0.49	0.68*	0.66*
Accent	N/A	0.48	N/A	0.54	N/A	0.43	N/A
Reading Ability	N/A	0.37	N/A	0.36	N/A	0.37	N/A

* "good" or above level of agreement (Landis and Koch 1977)

Intra rater was superior to inter rater agreement. The professional rater group achieved a "good" level of agreement for eleven of the twelve parameters assessed, the exception being Articulatory Precision. This parameter does not rate tracheoesophageal voice quality and the rationale for its inclusion was specified in for reasons outlined in 3.1.3.

SLT's attained higher mean scores than ENT raters for all parameters except for Articulatory Precision. The whole cohort of SLT raters attained "good" agreement for eleven of the fourteen parameters they evaluated with "moderate" agreement for Articulatory Precision and "fair" for Accent and Reading Ability. Again these parameters are not directly related to tracheoesophageal voice outcome. The All ENT sub-group achieved "good" intra rater agreement for eight of the twelve parameters with the remaining four parameters being classified as “moderate” agreement. However the mean score for Wetness (mean 0.59) fell only 0.02 below the aforementioned arbitrary boundary mark. Only the ENT raters failed to reach “good” agreement for Tonicity, Wetness and Paralinguistic Features.

The Expert SLT sub-group attained higher kappa means than the non-expert SLT’s for eleven parameters and one parameter had an identical mean for both groups. This contrasts with the Expert ENT sub-group who had higher scores than their non-expert colleagues for only two out of twelve parameters. Non-Expert ENT surgeons scored higher than Expert ENT for nine parameters and equivalent scores to the experts for one parameter.

Inter Rater Professional Results

When professional raters are considered as a whole group a “good” level of agreement was attained for only three of twelve parameters i.e. two of the three global parameters: Overall Grade, Social Acceptability and the uni-dimensional parameter Strain. The All SLT group attained “good” agreement for six parameters in comparison to the All ENT group who attained this level for three parameters. The Expert SLTs had nine classifications of “good” agreement in contrast to five for their less expert SLT colleagues. The Expert ENT group reached “good” agreement for three parameters with the Non-Expert ENT surgeons attaining this level for two.. The All ENT group achieved “good” levels of agreement for the same three parameters as the All Rater group results. The SLT sub-group attained “good” agreement for the same three listed above plus Volume, Whisper and Fluency.

Table 13. Inter rater mean weighted kappa co-efficients for professional raters.

Inter-Rater	All Professionals	All SLT	All ENT	Expert SLT	Expert ENT	Non-Expert SLT	Non-Expert 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
Articulatory Precision	0.43	0.33	0.37	0.46	0.50	0.27	0.47
Paralinguistic Features	0.43	0.53	0.37	0.53	0.27	0.53	0.45
Accent	N/A	0.28	N/A	0.35	N/A	0.26	N/A
Reading Ability	N/A	0.22	N/A	0.23	N/A	0.19	N/A

* “good” or above level of agreement (Landis and Koch 1977)

Expert SLT’s also attained higher or equivalent mean kappa co-efficients for all fourteen parameters when compared to the Non-Expert SLT group. Expert ENT’s attained superior mean co-efficients for five parameters in relation to the Non-Expert ENT group who achieved higher mean scores for seven parameters.

Summary

Table 14. Parameters with "good" inter and intra rater agreement for professional raters according to profession and expertise.

	Expert SLT	Expert ENT	Non-Expert SLT	Non-Expert ENT
Overall Grade	✓	✓	✓	✓
Tonicity	✓			
Strain	✓	✓		
Wetness	✓			
Volume	✓	✓	✓	
Social Acceptability	✓		✓	✓
Whisper	✓		✓	
Intelligibility	✓			
Stoma Noise				
Fluency	✓		✓	
Articulatory Precision				
Paralinguistic Features				
Accent				
Reading Ability				

Table 14 summarises the parameters that met Landis and Koch's definition of "good" or above agreement for both intra and inter rater agreement. Mean kappa co-efficients indicated higher levels of agreement for intra rater than inter rater reliability. Expert SLT's attained more "good" agreement classifications for both inter and intra rater reliability than their Expert ENT colleagues. Expert ENT surgeons attained this level for more parameters than their less Expert ENT colleagues. . When professional raters were analysed as one group (All Raters) they attained "good" inter and intra rater reliability for just three parameters (Overall Grade, Strain and Social Acceptability). However the ENT group did not achieve "good" reliability for inter rater Social Acceptability when considered as a sub-group. The SLT sub-group achieved "good" intra and inter rater reliability for six parameters (Overall Grade, Strain, Volume, Social Acceptability, Whisper and Fluency) whereas the Expert SLT group achieved this boundary level plus for three additional parameters (Tonicity, Wetness and Intelligibility).

The non-voice parameters consistently had the lowest levels of agreement for all groups for both inter and intra rater reliability i.e. Articulatory Precision, Paralinguistic Features, Accent and Reading Ability. These were the parameters added after the pilot study as outlined in 3.1.3. Even the Expert SLT group failed to achieve “good” reliability for these parameters with the exception of intra rater agreement for Paralinguistic Features (0.69).

The reliability results have implications for future use in clinical practice and research in relation to which parameters may be selected for inclusion in a finalised tool. These key issues are discussed in Chapter 6.

4.3 An investigation of the inter and intra-rater reliability for naïve raters with a modified version of the SToPS

4.3.1 Introduction

This study aims to investigate levels of intra and inter-rater agreement for naïve judges to fulfil the fourth research aim of this thesis. The recruitment of the naïve rater panel is detailed in 4.4.2 followed by an outline of the training session and the voice evaluation session. The voice stimuli were the identical to those utilised in the professional rater study.

4.3.2 Methodology

Recruitment of the naïve listener panel

Previous studies have recruited undergraduate speech and language therapy students as naïve listeners (Watson and Williams 1987; Williams and Watson 1987; Nieboer et al 1988; Pindzola and Cain 1988; Eadie and Doyle 2002a; Eadie and Doyle 2004; Moerman et al 2004; Eadie and Doyle 2005a). For obvious reasons such recruits may not be defined as naïve listeners. The issue of naïve listeners was summarised in 2.4.3. Naïve listeners were recruited for the study by informal invitation from friends and family of clerical staff within the SLT department. To meet the selection criteria naïve listeners were required to have:

- resided within the area covered by the ENT unit, in order to be familiar with the local accent;
- hearing within normal limits;
- no previous involvement with laryngectomy patients or with people with any type of disability that affects speech, language or voice;
- no voluntary or paid employment in healthcare/social services as this may cause increased sensitivity for people with reduced communication ability.

Details of the naïve listeners are summarised in Table 15.

Table 15. Naïve rater information.

Rater	Gender	Age	Occupation
Naïve 1	female	49	Secretary
Naïve 2	female	26	Teacher
Naïve 3	female	26	Planning clerk
Naïve 4	female	53	Housing officer
Naïve 5	female	57	Teacher
Naïve 6	male	56	Joiner
Naïve 7	male	56	Retired electrician
Naïve 8	female	72	Retired waitress
Naïve 9	male	74	Retired miner
Naïve 10	female	41	Translator for technical manuals

The age range of naïve listeners recruited was 26-71 years with a mean age of 50 years.

Naïve rater preparation for the rating sessions

As the naïve raters were selected to represent the general population, it was not appropriate for them to undergo training to become familiarised with tracheoesophageal voice quality. A 15-minute orientation session was carried out immediately prior to rating the voices. An information sheet about the study and their role was distributed to each panel member (Appendix C). Participants discussed all pertinent issues as a group with a SLT supervising the session. Raters were informed that the voice samples would be of patients who had undergone throat surgery. The aim of the research study was explained to be an investigation of how different people judge these voices. It was emphasised that their views were important as they reflect those of the general public who may meet the subjects in everyday life. Raters were encouraged to be truthful about their reaction to the voices and were reassured that subjects would not be informed of the results. The only additional guideline provided was for raters to imagine that each voice sample was “their own voice”, or “the voice of a close family member”. This specific instruction was aimed at making the rater more engaged with the speaker and how they are perceived in daily life as opposed to completing an abstract task of listening with no reflection on the person behind each sample.

The panel were permitted to hear only one voice sample of a laryngectomy prior to the rating session. The sample was a speaker with a tonic voice (neutral tonicity as defined in 1.4), with neither stoma noise nor any articulation issues. This speaker was selected as a bench mark. The raters were informed that this subject would be regarded as a “good”/excellent SVR speaker. The aim of this prior exposure to such a speaker was to enable raters to have an anchor reference point. The rationale was to try to prevent the panel hearing a “good” voice early in the fifty-five subject series but rating it as “poor” due to lack of awareness of the range. It is acknowledged that exposure to more SVR voices may have provided a better awareness of the likely range of voice quality. However playing more than one voice was judged to have the potential to start a training effect and hence prevent assessing a truly naïve listener viewpoint

Structure of the naïve listener rating session

The raters listened to the same master CD as described in the professional rater study (4.2.2). Each rater sat at a separate table and discussion was not allowed until all fifty-five ratings were completed. The group heard the samples simultaneously, in randomised order, from a Sony stereo cassette deck model TC –TX 373 with Dolby BNR. A thirty minute coffee break was arranged after the first thirty voice samples. Each sample was played twice, with a break of one minute between each play back. The only information marked on each of the fifty-five rating sheets was the voice sample number and the gender of the subject.

4.3.3 Results and data analysis

The raw scores assigned to each voice sample 1-55 by individual naive raters for sessions 1 and 2 were entered into Cytel Studio 8. A separate database was created for each parameter. Each database was analysed with the StatXact package to calculate weighted kappa co-efficients for both intra and inter rater reliability. The mean weighted kappa co-efficients for intra and inter-rater reliability for each parameter are summarised in Table 16 and Table 17. The comprehensive co-efficients for this investigation were formatted into tables by parameter for both intra and inter rater outcomes (Appendix J).

Intra rater Naïve results

Table 16. Naïve rater intra rater mean weighted kappa co-efficients

Parameter	Weighted kappa co-efficient
Overall Grade	0.72*
Tonicity	N/A
Strain	0.20
Wetness	0.31
Volume	0.09
Social Acceptability	0.69*
Whisper	0.53
Intelligibility	0.52
Stoma Noise	N/A
Fluency	N/A
Articulatory Precision	N/A
Paralinguistic Features	N/A
Accent	N/A
Reading Ability	N/A

* “good” or above level of agreement (Landis and Koch 1977)

Naïve raters attained a “good” level of agreement for two of the seven parameters assessed: Overall Grade (0.72) and Social Acceptability (0.69). These are global parameters. There was a large range of mean scores for Volume and Overall Grade. Three parameters were classed as “moderate” agreement and two as “poor”.

Inter rater Naïve results

The naïve group achieved “good” agreement for the identical parameters as for the intra rater investigation (the global parameters Overall Grade and Social Acceptability). No parameters indicated “poor” reliability but three were “fair” and two “moderate”. The third global parameter, Intelligibility mean co-efficient was 0.59 i.e. fell just 0.02 under the 0.61 boundary required for “good” agreement.

Table 17. Naïve rater inter rater mean weighted kappa co-efficients.

Parameter	Weighted kappa coefficient
Overall Grade	0.62*
Tonicity	N/A
Strain	0.24
Wetness	0.31
Volume	0.32
Social Acceptability	0.63*
Whisper	0.41
Intelligibility	0.59
Stoma Noise	N/A
Fluency	N/A
Articulatory Precision	N/A
Paralinguistic Features	N/A
Accent	N/A
Reading Ability	N/A

Summary of Naïve results

Naïve raters demonstrated "good" levels of agreement for both intra an inter rater reliability for the two global parameters Overall Grade and Social Acceptability. The third global parameter, Intelligibility, was linked to "moderate" levels of agreement for both intra and inter rater reliability. Intra rater agreement was also superior to inter rater for Overall Grade, Social Acceptability and Whisper.

Chapter 5. An investigation of the relationship between professional, naive, patient and carer raters in the perceptual assessment of tracheoesophageal voice

5.1 Overview

This section will investigate the variation in the perception of tracheoesophageal voice from the view point of SLT, ENT, Naïve, Carer and Patient in the same investigation. This is the fifth and final aim of this thesis. This is the first time this range of rater types has been undertaken in one study. The intra and inter rater agreement of professional (SLT/ENT) and naïve raters has been examined in the previous chapter and these will be compared to the judgements of patients and carers in an investigation of inter rater agreement from five rater perspectives. The importance of patient self rating and carer rating of outcomes was outlined in 2.4.1 and 2.4.2 respectively. The scope of this thesis does not include an investigation of patient and carer intra-rater agreement. Consequently the patient and carer rating scores are not examined in a separate section as undertaken for the professional and naïve judges but are integrated into 5.3.3 data to compare the five rater types inter judge reliability investigation that forms the body of this section. This section begins with the methodology of the patient data collection required for this investigation (5.2) followed by that of each patient's carer (5.3). The subsequent section (5.4) then reports the findings for the comparison of all five rater groups.

5.2 Methodology for the investigation of patient's self perception of tracheoesophageal voice quality

5.2.1 Recruitment of SVR patients

Patients who had volunteered and attended the voice recording session (4.2.2) were invited by letter to attend a further appointment in the Speech and Language Therapy Department. Forty patients attended; the remainder failed to respond to the invitation letter. The patient demographic table in Table 4.1 (4.2.2) details the patients who attended.

5.2.2 Patient rater interviews

Interviews were carried out on a one to one basis without a carer being present. The questions and potential responses from the patient and carer scale (Appendix K) were read aloud by a research assistant alongside the patient. Each item could simultaneously be viewed by the interviewer and patient. Each subject was allowed as much time as needed to respond. This method is different to the professional (4.2) and the naïve study (4.3) where voices are rated from an audio recording; patients did not hear the audio recording of their own voice at any time during the study. The aims of the study were explained and it was stressed that patients should give their true opinion. They were informed they should not feel they were expressing dissatisfaction with their clinical care if they rated their voice as being less than satisfactory.

5.2.3 Data entry

The raw scores from each patient for each item on the scale were entered into a database using the statistical package Cytel Studio 8.

5.3 Methodology for the investigation of carer perception of tracheoesophageal voice quality

5.3.1 Recruitment of carers of SVR patients

The aim of interviewing carers was to investigate the perceptions of tracheoesophageal voice by those who live alongside SVR speakers on a daily basis. For the purpose of this study a carer is defined as someone the patient spends time with on a regular basis. This could be a friend or relative. The carer recruitment procedure was identical to that outlined for patients 5.2.1. All were carers for a patient who had previously attended for voice recording. The letter sent to invite patients for interview also requested if the patient would agree to ask a carer to attend for a separate interview immediately after their own interview. All carers spoke English as a first language; to the author's knowledge, none had language, cognitive or psychiatric impairments that would affect their comprehension or ability to participate in the rating.

5.3.2 Carer interviews

The interview format was identical to that for the patient interviews as described in 5.2.2 and used the patient and carer scale (Appendix K). Carers were interviewed alone, without the patient subject present, and were reassured that their relative/ friend would not be informed about any opinions they expressed. Carers were not asked to listen to the recording of the patient during the interview format. Thirty-seven carers attended as detailed in Table 9 (4.2.2). One patient reported he had no family or social contacts and attended alone, one patient preferred his carer was not included in the research and one patient reported his carer was chronically ill and unable to attend. The scale for carers differed from the patient scale only by including a question concerning carers' auditory acuity. Carers were asked if their hearing ability was within normal limits and all participants self rated their hearing as normal for day to day purposes.

5.3.3 Data entry

The raw scores from each carer for each item on the scale were entered into a database using the statistical package Cytel Studio 8.

5.4 The relationship between professional, naïve, patient and carer ratings

5.4.1 Overview

This section details and reports the results of the comparisons of the raw scores from the five different types of rater. The methodology is outlined (5.4.2) followed by comparison of: patient and carer raters (5.4.3), patient and naïve raters (5.4.4), patient and professional raters (5.4.5), carer and naïve raters (5.4.6), carer and professional raters (5.4.7) and naïve and professional raters (5.4.8). A summary of the overall findings of this investigation are detailed in 5.4.9.

5.4.2 Methodology

The raw data from each parameter database for each rater group was cut and pasted into new databases in Cytel Studio 8; thus each parameter had a separate database which included the raw scores for all rater types. Naïve and professional raters had test and re-test data but only session one data was selected for this investigation. Two sub-groups from the professional rater study were included i.e. Expert SLT and Expert ENT. The rationale for selecting only these sub-groups relates to: a) Expert SLTs demonstrating the best agreement (4.2.2) and b) Expert ENT surgeons being uniquely responsible for selecting and undertaking the surgical procedures that have the potential to affect voice outcome and carrying out long term follow up and review of tracheoesophageal speakers.

Table 18. Weighted kappa co-efficients for rater type versus parameter.

	Patient Vs Carer	Patient Vs Naive	Patient Vs Expert SLT	Patient Vs Expert ENT	Carer Vs Naive	Carer Vs Expert SLT	Carer Vs Expert ENT	Naïve Vs Expert SLT	Naïve Vs Expert ENT
Overall Grade	0.57	0.33	0.34	0.30	0.41	0.48	0.36	0.67*	0.63*
Social Acceptability	0.58	0.23	0.33	0.28	0.38	0.51	0.44	0.60	0.53
Intelligibility	0.74*	0.19	0.39	0.43	0.17	0.40	0.46	0.36	0.19
Volume	0.63*	0.15	0.30	0.33	0.03	0.45	0.49	0.20	0.07
Whisper	0.29	0.28	0.21	0.14	0.15	0.33	0.16	0.32	0.22
Wetness	0.62*	0.18	0.14	0.13	0.20	0.08	0.16	0.28	0.25
Strain	0.40	0.14	0.21	0.07	0.11	0.17	0.05	0.22	0.16

N.B. Mean weighted kappa coefficients are listed for all group comparisons except Patient versus Carer. This group did not have this type of calculation due to methodological difference whereby they only rated one voice each.

* "good" or above level of agreement (Landis and Koch 1977)

Certain raw scores had to be modified as the naïve group only rated some parameters with a yes/no response rather than a four point scale. This meant that an extra line was added to each database for the other four rater types whereby scores of two and three on the four point scales were converted to one to allow a corresponding comparison with naïve raters. Each database was analysed with the Cytel StatXact package to calculate weighted kappa co-efficients for inter-rater reliability between each of the rater groups. The weighted kappa co-efficients are detailed in Table 18 to compare rater comparator pairs with scale parameter. The comprehensive inter rater reliability results tables are listed in Appendix L.

5.4.3 Comparison of Patient and Carer Raters

The highest levels of agreement of all comparator pairs were observed between patient and carers. "Good" levels of agreement were attained for three of the seven parameters; one was a Global parameter (Intelligibility) and the other two were uni-dimensional (Volume and Wetness). The co-efficients for the other two global parameters Overall Grade (0.57) and Social Acceptability (0.58) were marginally below Landis and Koch's cut off point for "good" agreement. Strain was classed as "moderate" agreement and Whisper "poor" agreement.

5.4.4 Comparison of Patient and Naive Raters

Patient and naïve raters achieved the lowest reliability of all comparator pairs in this investigation. Only one co-efficient can be classed as indicative of "fair" agreement (Overall Grade 0.33) with the remaining six parameters falling into the "poor" agreement classification.

5.4.5 Comparison of Patient and Professional Raters

The professional raters are divided into 2 sub-groups Expert SLT and Expert ENT. However both groups compared identically in relation to the patients' self-rating in respect of the classification of agreement. Only "fair" agreement was achieved for the three global parameters and "poor" agreement for the four uni-dimensional parameters ranging from 0.07 – 0.30 mean kappa co-efficients.

5.4.6 Comparison of Carer and Naïve Raters

This comparator pair attained only "moderate" agreement for the global parameters Overall Grade and Social Acceptability but the remaining five parameters can be classed as reaching only "poor" levels of agreement.

5.4.7 Comparison of Carer and Professional Raters

The Expert ENT and Expert SLT groups again performed almost identically when compared to carers. The cut off points allocated by Landis and Koch mean some differences in category of agreement are defined but the actual raw scores are close to each other. The Expert SLT group reached "moderate" agreement with Carers for two global parameters (Overall Grade and Social Acceptability) one uni-dimensional (Volume), "fair" agreement for Intelligibility and Whisper and "poor" for Wetness and Strain. Expert ENT agreement with Carers was "moderate" for Social Acceptability, Intelligibility and Volume, "fair" for Overall Grade and "poor" for Whisper, Wetness and Strain.

5.4.8 Comparison of Naïve and Professional Raters

Naïve and Expert ENT/ SLT raters reached higher agreement than with the Patient and Carer comparator groups. Expert ENT and SLT groups both achieved "good" inter rater agreement with Naive for one parameter alone i.e. Overall Grade. Social Acceptability was the next closest agreement classed as "moderate" (0.60 for SLT and 0.53 for ENT). However only two other parameters, the third global parameter (Intelligibility) and Volume achieved even "moderate" agreement but only for SLT versus Naïve not the

ENT versus Naïve co-efficient. All other parameters and groups within this section were classed as “poor” agreement.

5.5 Summary

Only 8% of the total sixty-three weighted kappa rater group comparison scores within this investigation can be classed as “good” agreement (Landis and Koch 1977). There is generally limited agreement between rater types. Patients and carers achieved the highest level of agreement with three parameters classed as “good”. Naïve versus ENT and Naïve versus SLT groups achieved “good” agreement for Overall Grade only. The implications for clinical practice and research regarding rater type and parameter will be considered fully in the following discussion chapter.

Chapter 6. Discussion

6.1 Introduction

The previous chapters have detailed the aims of developing a valid and reliable rating scale for professional and other raters to undertake tracheoesophageal voice perceptual analysis. This is the first set of studies to investigate SVR voice outcomes from SLT, ENT, Naïve, Patient and Carer perspectives and the relationships between their different judgements.

This chapter will discuss the key findings from the five research aims (listed in 2.6) in three separate sections. The strengths and limitations of each of the three studies, along with the areas for further research, are summarised at the end of each individual section.

Section 1 (6.2) examines the reliability and validity of the professional scale (SToPS).

Section 2 (6.3) examines the reliability and validity of the modified version of the SToPS for naïve raters.

Section 3 (6.4) analyses the relationships between the five key rater types (Expert SLT, Expert ENT, Naïve, Patient and Carer).

6.2 Is the SToPS a valid and reliable scale for the perceptual assessment of tracheoesophageal voice by SLT and ENT raters?

The summary of the literature review (2.5) outlined the requirement for a new perceptual tool to assess tracheoesophageal voice as no scales to date have demonstrated robust validity and reliability. Two research aims of this thesis are to develop a valid assessment tool for professional raters that met these criteria and to examine the reliability of the scale according to rater type and expertise. It has been suggested that the evidence of a scale's reliability should first be considered prior to obtaining evidence of its validity (Streiner and Norman 1995 p6), based on the rationale that a scale that lacks stability and reproducibility is inherently unreliable. However it would seem optimal when designing a new scale to consider elements of validity at the same time as reliability. The reliability of the SToPS will be examined first (6.2.1) prior to its validity (6.2.2).

6.2.1 The reliability of the SToPS

This section will begin with a discussion of the inter and intra rater kappa co-efficient results for reliability. This will be followed by an outline of the other aspects of reliability theory which require consideration before it can be established whether the SToPS may have clinical applicability beyond the remit of this research investigation. These complex inter-relational issues were outlined in 2.2 and 2.3.

Intra rater issues

Twenty one of the professional raters (12 SLT and 9 ENT) completed the rating tasks. When analysed together they showed "good" (Landis and Koch 1977) reliability for eleven of the twelve parameters. This suggests there is a relatively stable internalised baseline for these parameters against which a psychoacoustic evaluation can be made (Kreiman et al 1993). Test-retest weighted kappa co-efficients demonstrated superior intra rater reliability compared to inter rater reliability for all parameters. This replicates findings in the perceptual analysis of laryngeal voice (Dejonckere et al 1993; de Bodt, Wuyts et al 1997; Webb et al 2004; Zraick et al 2011). Kreiman et al (1993) suggested a theoretical explanation for this phenomenon. The superior reliability of test-retest is due to each judge having a different individual baseline and scale point references because of individual perceptual habits, biases and sensitivity to each parameter being measured. Consequently comparing a judge's individual perceptions to that of other raters will always have lower reliability. However no rater sub-groups in this thesis attained the level classified as "very good" agreement (co-efficients over 0.81). This indicates the raters' internal standards are not entirely stable. Potential influencing factors for intra rater agreement variation were discussed in 2.3. These will be explored more fully in the next section as they are equally relevant to inter rater reliability.

In this study only three parameters failed to meet the definition of acceptable reliability that was outlined in 4.2.2 with any professional rater group. These were a) Articulatory Precision, b) Accent and c) Reading Ability. These three parameters have a common denominator; they do not measure voice quality and were selected solely on the recommendation of the pilot study SLTs (3.1.3) to enable analysis as to whether these factors influenced professional or naïve rater scores for Overall Grade and Social Acceptability (3.2.2). They are not normally assessed by SLTs and the low agreement potentially relates to the lack of an internal representation, especially for issues around judging articulation in tracheoesophageal voice. The inclusion of these items and the lack of rater agreement even for Expert SLTs warrant more discussion which will be undertaken in the section that considers the limitations of the SToPS.

Although the All Professionals group reached "good" agreement for most parameters the raw mean co-efficient scores showed some variation according to rater type (Table 12 section 4.2.3). The SLTs achieved more of the arbitrary “good” classifications of agreement than ENT but analysing whether this is statistically significant is outside the scope of this thesis. Again these key findings will be discussed in detail in the following section as they are equally pertinent to inter rater factors.

This thesis’ finding of the superiority of intra rater scores in relation to inter rater differs from previous studies in tracheoesophageal voice perception which reported the opposite observation (Cullinan et al 1986; Lundstrom et al 2008). Similarly van As et al (2003) found superior inter rater agreement for two thirds of their 21 parameters. These studies do not discuss the apparent anomaly of the superior inter rater scores in relation to the evidence base from laryngeal perceptual studies which are in keeping with the findings of this thesis. Previous studies have also reported higher intra rater agreement than this thesis (Cullinan et al 1986; van As et al 2003; Kazi et al 2006a, 2006b, 2009; Lundstrom et al 2008; Ward et al 2011). Both of these apparent differences may be explained by how agreement was calculated. The issue of statistical selection was outlined in 2.3.4.

No previous studies have investigated the effect of rater type and expertise upon intra rater reliability of tracheoesophageal voice but a few have examined the inter rater effect which is discussed below.

Inter rater issues

A number of key factors require consideration regarding the reliability of the SToPS. These relate to parameter selection, rater profession and expertise, task, training, scale format and voice stimuli related issues. These aspects apply equally to both inter and intra rater reliability.

Parameters

In this investigation only three SToPS' parameters (Overall Grade, Social Acceptability and Strain) were classed as "good" reliability when analysing the All Rater group as a whole. However further analysis by sub-group (Table 13 section 4.2.3) demonstrated only Overall Grade could be classified as "good" agreement by all five of the individual rater sub-groups. There was a profession effect for attaining this arbitrary level of agreement for Social Acceptability, Whisper and Fluency (only SLTs achieved "good" agreement) and an expertise effect for Strain (only Expert SLTs and Expert ENT achieved "good" agreement). However these are arbitrary cut off points of agreement of 0.61 as suggested by Landis and Koch (1977)

It is important to consider why raters agree about some parameters more than others. It would appear from the current scales that the highest co-efficients are for global parameters rather than the uni-dimensional ones that require more complex discrimination from the stimuli. The superior reliability of these over-arching parameters is in keeping with studies in laryngeal perceptual assessment which have found Overall Grade to be the most reliably assessed (Dejonckere et al 1993; Webb et al 2005; Millet and Dejonckere 1998). However comparing the inter rater results of the SToPS with previously reported tracheoesophageal voice perceptual studies presents many difficulties due to variation in scale format and parameter nomenclature. Only four studies to date have investigated inter rater agreement in tracheoesophageal voice perceptual assessment with a validated perceptual scale for alaryngeal voice (van As et al 2003; Kazi et al 2006a; Kazi et al 2006b; Kazi et al 2009). The Kazi et al series adopted one global parameter, "Overall Voice Judgement", from van As et al's scale (2003), as well as the unvalidated (for this patient group) GRBAS. All three studies by Kazi et al reported high reliability for van As et al's parameter and the Overall Grade from GRBAS. In contrast van As et al did not observe clearly superior reliability of global parameters in comparison to the more specific ones. Three other studies used non-validated scales (Vlantis et al

2003; Moerman et al 2006; Ward et al 2011). These generally agreed that "Overall Grade" ratings appear the most reliable.

The most noteworthy factor from the above studies is that reliability coefficients for global parameters are much higher than those reported: a) for laryngeal voice and b) in this thesis. If uni-dimensional parameters measured by an EAI scale are considered, only four studies can be directly compared to this thesis (van As et al, Vlantis et al, Moerman et al and Lundstrom et al). These studies are summarised in Table 1 (2.2) along with the statistical method used to measure reliability. This thesis represents the first study to use weighted kappa co-efficients to measure tracheoesophageal voice quality reliability. The lower rates of reliability for the SToPS may reflect: a) the more robust choice of statistics for this study that account for chance agreement and b) the inclusion of a much greater number of raters and voice stimuli. The exceptions to this are the four parameters that were added post pilot: Positive Paralinguistic Features, Impairment of Articulatory Precision, Accent and Reading Ability that achieved the lowest agreement. These appear to have a poor internal representation even for the Expert SLT sub-group and issues around the definition of these parameters and scale design will be fully discussed when considering the limitations of this investigation.

Rater type and expertise

The reliability of individual parameters cannot be considered in isolation . Expert SLTs achieved more parameters that can be classified as "good" agreement than the less expert SLTs and both expert and non expert ENT surgeon groups .. This is the first investigation to examine these factors in tracheoesophageal voice perceptual assessment. Previous studies listed in Table 2.1 (2.2) have recruited SLTs familiar with SVR speakers but it is not possible to extrapolate their expertise in voice perceptual analysis. This thesis has included a definition of expertise for SLT and ENT raters (4.2.1).

The raw mean co-efficient scores (4.2.2) were consistently higher for SLT raters than ENT judges with the exception of Articulatory Precision. Similarly the Expert SLTs mean scores were superior to less expert SLTs for all parameters except two which had identical means for both groups. This pattern was not observed for ENT surgeons. Analysis of the statistical significance of these raw co-efficients is outside the remit of this thesis. Furthermore these are mean scores and the range of rater kappa co-efficients varies considerably (Appendix G).

Expert SLTs failed to reach "good" agreement for five parameters. As discussed in the previous section, four of these less reliable parameters are not routinely assessed and do not relate to voice quality. Their internal representations may be expected to be lower for these parameters in contrast to those that are more routinely used in clinical practice. The fifth less reliable parameter for this group, Stoma Noise warrants discussion. Its assessment is an essential part of clinical practice as it detracts from and competes with the speech signal (Perry 1989 p84); consequently the reduction of stoma noise can be a valid therapy goal. It is possible that SLTs are perhaps not accustomed to measuring its severity especially if it is intermittent. Further training and consensus agreement may enable this to be more reliably assessed in future studies and as a key outcome measure in clinical practice.

As Expert SLTs achieved more categories of "good" agreement it suggests they may have more stable internalised representations of key tracheoesophageal voice parameters Similarly the same pattern for inter rater agreement indicates this group may share some commonality of internal representation of the concepts. Such inter-rater agreement does not however equate to them achieving the "correct" scores; this important consideration will be discussed further in the validity section (6.2.2). There has been no research to date that has examined the potential relationship between the acquisition of perceptual skills for dysphonic laryngeal voice and the subsequent transfer to expertise in alaryngeal voice assessment. Eadie et al (2010b) found naïve raters who were musicians had significantly better intra-rater reliability than non-musicians for rating dysphonic voice stimuli. This could not be accounted for in terms of skill in pure tone pitch discrimination. This preliminary study suggests some cross over of ability of auditory perception. Similarly, it would appear reasonable to suggest that enhanced laryngeal perceptual skills enable the discrimination of parameters within the more complex tracheoesophageal voice stimuli. This area of voice perceptual theory requires further investigation.

The limited research into the effect of expertise on the reliability of perceptual assessment has involved the rating of only laryngeal not tracheoesophageal voice quality. Both Bele (2005) and Bassich and Ludlow (1986) reported that student SLTs did not achieve the same reliability as expert SLTs (both studies included training and Bassich and Ludlow provided written definitions and some anchors). This is in keeping with this thesis whereby unified pre-task training, practice rating and textual anchors and definitions did not facilitate all raters performing equally. One study (de Bodt et al 1997) specifically examined both the effect of profession and degree of expertise with the GRBAS scale. However expertise was defined purely in terms of number of cases encountered rather than in voice rating skills or experience. They reported there was no significant difference between ENT and SLT, but mean scores suggested professional background was more important than years of experience in dysphonia as there was a trend towards SLTs showing better inter rater agreement. Direct comparison to the SToPS is hindered by the variation in the definition of expertise but this suggests that SLTs' clinical use of scales may facilitate better agreement than their ENT colleagues who encounter many patients but are not required to perceptually categorise their voice quality.

The effect of experience, purely in terms of the number of dysphonic laryngeal voices encountered, has been linked to rater ability (Kreiman 1997) in terms of less experienced raters (who have encountered fewer pathological voices) showing a higher likelihood of increased variations in internal standards. However in this thesis some of the Expert ENT raters had encountered large numbers of speakers yet still did not achieve “good” intra and inter rater agreement. An alternative view, that perceptual training causes more inter rater variance for laryngeal voice rating has also been suggested (Kreiman et al 1994). Kreiman et al attributed expertise to better psychoacoustic awareness of a wider range of parameters but raters then focus on different aspects of the auditory signal, paradoxically creating less inter rater agreement. This effect of expertise causing reduced agreement was not observed in this investigation of the SToPS. Expertise in rating voices and experience of voice stimuli may be two different phenomenon and not interchangeable categories. Differentiating and contrasting these two factors is an area for future investigation.

Scale format

This is the first tracheoesophageal perceptual scale to use an equally appearing interval (EAI) scale with clear textual definitions and a baseline of optimal alaryngeal voice for the appropriate parameters. This baseline was not necessary for all parameters as some can be based on a normal laryngeal baseline when it is feasible for tracheoesophageal speakers to achieve this level of function. These were specified in 3.2.1.The inclusion of specific textual definitions would be expected to be associated with superior reliability but previous studies have reported higher co-efficients. As observed above, this finding could be due to the methods of statistical analysis. However another explanation could relate to studies importing laryngeal scales to assess tracheoesophageal voice e.g. Kazi et al (2006a; 2006b; 2009) reported high levels of reliability with the GRBAS. However tracheoesophageal voices are markedly different to laryngeal qualities and consequently maintaining the same scale format will force judges to polarise all the voice samples to the severe end of the scale thus producing artificially high reliability.

Van As et al (2003) provided specific textual definition anchors explaining that raters should compare their 3 point adjectival overall judgement scale to laryngeal voice. However they failed to specify the baseline for their 21 bipolar scales and whether it relates to tracheoesophageal voice or normal laryngeal voice. Furthermore some parameters were transferred across from laryngeal voice assessments and as outlined in relation to Kazi et al's studies this potentially polarised ratings and increased reliability.

A different issue relates to the study by Ward et al (2011) where scale format was similarity/difference of each parameter for two paired voice stimuli at a time. This scale format does not rely on raters' internal baselines and eliminates the problem of unstable internal representations and the effect of previous stimuli influencing rater judgements. However this is not a feasible format to use routinely in clinical settings and a key aim of this thesis was to develop a scale that can be used quickly without compromising validity and reliability. The EAI scale used in the SToPS appears to be sufficiently reliable for the key parameters if the raters are sufficiently skilled and trained (according to specifications of adequate reliability suggested by Landis and Koch (1977) and Streiner and Norman (1995 p7).

Only one previous study has included the auditory perception of tonicity (van As et al 2003). The difficulties with scale construction for this complex parameter were discussed in 2.3 along with the rationale for the chosen format selected for the SToPS. The design of the SToPS enabled it to become the first study to include a measure the perceptual impression of Stenosis i.e. the absence of tone due to fibrotic tissue in the neopharynx (Cheesman et al 1986; Singer et al 1986; McIvor et al 1990). Further indication that the

Tonicity parameter is a complex parameter to rate relates to the finding that only SLTs achieved a "good" intra rater agreement classification and only Expert SLTs attained this level for inter rater judgements. Again further research is required to determine whether rater group types show statistically significant differences. The STOPS' tonicity parameter requires raters to assign a stimulus to one of the three branches of the scale or alternatively to the mid-point of zero of neutral tonicity. In contrast to this thesis , van As et al reported very high reliability co-efficients for their two Tonicity scales i.e. Hypertonic-Not Hypertonic (intra 0.93, Inter 0.89) and Hypotonic-Not Hypotonic (intra 0.80, inter 0.90). Such division of Tonicity into two scales whilst ignoring Stenosis may be expected to increase reliability. Van As' scale permits selection of the mid-point of both scales if the rater is uncertain about tonicity and this may also falsely increase reliability scores.   The previously mentioned statistical choice issues may also be responsible for the superior reliability.

Rating task and training

This thesis used anchors in pre-rating training for Overall Grade and Tonicity and practice use of the scale for all parameters with feedback and discussion. Textual anchor definitions were also available during training and rating. No previous investigations have examined the effect of task or training on the reliability of tracheoesophageal perceptual assessment but laryngeal voice perceptual research has some limited evidence regarding these factors. There has only been one study to examine such influence on SLTs (Eadie and Kapsner- Smith 2011). This showed that auditory anchors before each stimulus improved inter but not intra rater agreement. There is some evidence that naïve raters are helped with auditory anchors (Kreiman and Gerratt 1998; Chan and Yiu 2002). It would be impractical to offer anchors for all parameters and scale points for the SToPS for clinical use. Some research raises issues about the routine use of anchors i.e. the anchor tasks per se may cause raters to behave differently than in day to day clinical situations (Kreiman et al 1990) and without training anchors can cause high listener variability (Chan and Yiu 2002). Further research with qualified SLT raters is clearly required before their role in clinical practice can be established.

Training programmes would intuitively appear to be a relatively easy way to improve agreement. Related areas of research would appear to agree (Chan and Yiu 2006; Eadie and Baylor 2006; Lee et al 2009; Iwarson and Peterson 2011). However there is no conclusive evidence that training can bring about improved, enduring reliability for SLT raters in SVR. This research area is in its infancy and has yet to be translated into clinical practice improvements.

Voice stimuli related issues

There is some evidence (again from laryngeal voice studies) that the type, range and order of presentation of voice samples can affect rater behaviour. More extreme cases of voice stimuli have been found to be more reliably rated (Kearns and Simmons 1988; Gerratt et al 1993; Kreiman et al 1993). Similarly, the severity grade allocated to a stimulus has been shown to shift if the ones preceding it are very severe or very mild (Gerratt et al 1993). The study design for this thesis involved the presentation of the voice stimuli in the same random order for all raters i.e. professional and naïve. The re-test session was again the same for all judges but presented in a different random order. The scope of this thesis did not include investigation regarding whether raters show similar shift in relation to the severity of preceding stimuli. However the range of voices included in this investigation with the SToPS was analysed in more detail (Hurren et al 2009) and demonstrated a wide range of tonicity types and severities were included. As a clear link between Tonicity and Overall Grade was also established it is reasonable to suggest that a sufficiently wide range of tracheoesophageal qualities were covered. This would preclude polarised stimuli artificially inflating reliability coefficients as suggested by Kreiman et al (1993).

Summary of the reliability of the SToPS

These results can only be related to this panel of judges with the 55 stimuli selected for this study. Further investigations would be required to see if these results were replicable with different raters and different tracheoesophageal speakers. This would need to include similar training and auditory and textual anchors as evidence from laryngeal perceptual research suggests these can have a positive effect on reliability. This is the first study to investigate classifications of the reliability of tracheoesophageal voice perception by expertise and profession. SLTs attained more "good" categories than ENT raters. Furthermore expertise in voice perceptual rating skills may be associated with the findings of higher raw mean co-efficient scores and the superior number of parameters that could be classified as "good" agreement. . This level of agreement may be linked to experience in rating voices in general rather than the number of tracheoesophageal speakers encountered. Training does not appear to afford "fast track" expertise in rating voices as all raters underwent a three hour training and scale familiarisation session. Overall Grade was the most reliably rated parameter and intra rater agreement was superior to that of inter judge, reflecting the evidence base from laryngeal voice perception. This parameter would seem to offer a useful outcome measure for tracheoesophageal speakers for all levels of rater. This study included the optimal statistics to account for chance agreement. This together with more complex, but more clinically pertinent, scale design may explain why higher coefficients were reported in previous tracheoesophageal perceptual studies.

Reliability is an essential part of validity as a scale cannot be valid if it is not reliable but this is only one aspect and other testing is required to ascertain whether a scale is sufficiently valid. This provides the basis for the following section.

6.2.2 The validity of the SToPS

Reservations about the validity of pre-existing tracheoesophageal scales were analysed in detail in 2.2 and summarised the rationale for a new tracheoesophageal perceptual scale for clinical and research practice. The essential aspects of validity theory were outlined in 1.6.2 and will be considered in relation to the SToPS.

Content validity

The SToPS intends to measure the key aspects of the voice quality of tracheoesophageal speakers. With this aim in mind its development involved the selection of a comprehensive range of parameters as detailed in 3.1 and 3.2. This included: a) explicit rationale for selection including clinical relevance and with reference to the literature, b) clear definitions, c) baseline reference point (for the zero score) where the parameter is judged to be optimal and d) textual reference descriptions to act as anchors for scale points 1-3 for all the parameters which were made available to raters during the training and rating.

A further aspect involved using SLTs experienced in SVR to pilot and provide feedback on the scale (3.1.1 and 3.1.3). Expert panel review is considered to be the minimum standard for content validity (Guildford 1954). This study included three SLTs for version 1 of the SToPS and twenty for the formal pilot (3.1.2). This meets the criterion of 3-10 panel members which has been suggested as sufficient for this purpose (Streiner and Norman 1995 p5).

Criterion validity

Criterion validity is difficult to establish for the SToPS as it involves comparing a scale to a pre-existing tool or to another “gold standard” assessment. Reservations about the validity of the only pre-existing scale (van As et al 2003) were outlined in 2.2. Consequently comparing the SToPS to this scale is unlikely to be of theoretical or clinical value. Furthermore there is no clinical consensus in the literature regarding the gold standard for the assessment of tracheoesophageal voice. Alternative methods of assessment, especially of tonicity, were listed in 1.4 but the evidence base to date suggests there is no suitable alternative definitive outcome in lieu of perceptual voice assessment. All other assessments have issues of reliability and interpretation and cannot be considered as an established criterion against which to compare the SToPS.

Construct validity

In the absence of criterion validity and where clear justification for a tool’s development is established, as in the need for a scale such as the SToPS, construct validity must be considered (Streiner and Norman 1995 p9). The laryngeal perceptual assessment literature has examined whether a scale measures or correlates with the theoretical construct i.e. are we measuring what we think we are measuring and are the constructs the prime variables in the phenomenon? This aspect of validity testing was outlined in 1.6. Gerratt and Kreiman (2000) postulated the key difficulty in establishing this type of validity in perceptual voice analysis stems from there being no correct quality judgement for a given voice stimulus. Even experienced raters show variation in allocating a scale point to a stimulus; this is not random rater error, rather raters a) use different perceptual strategies and b) disagree about the perceptually important parameters in dysphonic voices (Kreiman and Gerratt 1996). We cannot attribute voice quality to the stimulus; instead we should view it as an interaction between the voice and the rater’s psychoacoustic perception (Kreiman and Gerratt 1996).

As the ultimate aim of validity testing is to establish what can be inferred from a scale (Streiner and Norman 1995 p147) the fundamental issue is to consider whether the SToPS has measured what it intended to measure. It is likely that the imperfect agreement observed with the twenty-two professional raters included in this thesis occurred due to the reasons outlined by Kreiman and Gerratt that have been summarised throughout this thesis. Although their research relates to laryngeal voice it is reasonable to assume their key theoretical hypotheses apply equally to tracheoesophageal voice. The first parameter in the SToPS, "Overall Grade" was measured against an explicit baseline of optimal tracheoesophageal voice. This is the first time this has been undertaken in a study. As this parameter was associated with "good" agreement for all raters it suggests raters have a relatively stable, internalised standard against which they judged each stimulus. This psychoacoustic interaction between the stimulus and the rater is referred to as a "hypothetical construct". This construct of optimal tracheoesophageal voice consequently appears to have meaning for the full range of raters included in this investigation.

It is possible that Expert SLTs have stronger internalised representations of the voice parameter constructs included in the SToPS as observed by them achieving more classifications of "good" agreement. However some parameters achieved low reliability even for experienced, expert raters and the constructs behind these parameters must be questioned. This may relate to the relative unfamiliarity of these parameters which are not typically assessed by SLTs or may relate to scale format and design. This important aspect will be addressed more fully when considering the limitations of the SToPS.

Although a key aspect of validity is the reliability of a measure, it is crucial to emphasise that robust reliability co-efficients do not prove scale validity per se (Kreiman and Gerratt 1996; Kreiman and Gerratt 1998; Kreiman and

Gerratt 2000). Reservations about the validity of previous studies' scales with high reliability co-efficients were discussed in the previous section due to scale format, statistical choice and baseline issues. This study has aimed to address the issues identified in previous studies. Construct validation is "part science and part art form" (McDowell and Newell 1996 p36) and researchers can only aim to understand such constructs by repeat testing with different predictions. Greater weight would be given to the scale's construct validity if similar or increased reliability is achieved with different professional raters and stimuli on a number of occasions.

The final consideration concerns whether some parameters within the SToPS could be classed as hypothetical constructs whereas others may be able to undergo specific construct validity testing. These are areas for further research now that a reasonable tool for perceptual tracheoesophageal voice assessment has been developed. Parameters in the SToPS that relate to hypothetical constructs are Overall Grade, Social Acceptability, Strain, Stoma Noise, Wetness, Whisper, Articulatory Precision, Paralinguistic Features, Accent and Reading Ability. In contrast, parameters such as Tonicity, Volume and Fluency may be able to be tested in relationship to other assessments. Tonicity has the potential to relate to intra-oesophageal manometry measurements (Perry 1989; Chone et al 2008) or tracheal manometry measurements (Allan et al 2005), Volume to decibel level recording (Sedory et al 1989; Deschler et al 1999; Delsupehe et al 1998), Intelligibility to a formal, standardised test of intelligibility (DeMaddalena and Zenner 1995; McHenry 2011) and Fluency to speaking and articulation rate using software designed for this purpose (the SLAAP utilised by Kendall (2009) to calculate speech rates and pause times).

6.2.3 Summary of the investigation into the reliability and validity of the SToPS with professional raters including its limitations and areas for further research

The first aim of this study was to design a tracheoesophageal perceptual scale for professional raters in the absence of a suitable pre-existing tool. This was achieved with the development of the SToPS. The second aim was to assess its reliability and validity. The SToPS was found to meet arbitrary levels of rater agreement determined by Landis and Koch (1977). Pre-existing expertise in voice perceptual analysis may be associated with the number of parameters that achieved the arbitrary "good" classification but further statistical analysis would be required to determine whether the differences are significant. . Further discussion about this aspect will follow in relation to the application of the SToPS into clinical practice where raters are likely to have a wide range in skill in voice perceptual analysis. Some aspects of validity testing were possible, especially in relation to content validity which was addressed via a literature review and pilot studies (3.1.3). Criterion validity was more complex as there is no gold standard assessment in SVR against which to measure tracheoesophageal voice perception. Consequently there is still a need for future work to determine how this scale relates to other assessments of tracheoesophageal phonation and to address the issue of construct validity. These aspects are outlined below in relation to the limitations of this study and potential future directions.

The discussion of the limitations of this study will be discussed in two categories: a) those that relate to study design and how this could have been improved with hindsight and reflection, b) those that could not be addressed due to the inherent limitations in the scope of a thesis.

Limitations of design

The first design aspect concerns the four parameters that failed to reach acceptable levels of agreement for even the Expert SLT group (Impairment

of Articulatory Precision, Positive Paralinguistic Features, Accent, Reading Ability). Their common factors concern:
i) their addition post SLT pilot, ii) they entail factors that do not pertain transparently to tracheoesophageal voice quality and iii) their lack of inclusion in routine clinical assessment and practice in the fields of either voice or laryngectomy clinical practice. The lack of intra rater agreement for all but Positive Paralinguistic Features suggests these hypothetical constructs do not have psychoacoustic internal representation and hence they lack validity. This may relate to scale design, parameter definition or that these factors cannot readily be partitioned and ranked with an equally appearing interval scale, not to mention the likelihood that they probably depend highly on socioacoustic, socioarticulatory judgement and not psychoacoustic variables.

Reading Ability may have influenced performance and rating as the stimulus type was reading aloud. A future study could factor in this variable by e.g. selecting a reading passage with known reading age threshold, permitting prior familiarisation with words before speaking and/or recording the education level of patients to include in statistical analyses.

Impairment of Articulatory Precision poses a more complex issue. This scale attempted to assess this factor by including the full range of articulatory concerns in one scale i.e. lack of dentures, habitual articulatory setting and severe dysarthria. Assessment tools for dysarthria are relatively well researched and developed and could be utilised if tracheoesophageal speakers have this type of speech pathology. SLTs working in laryngectomy have a clinical need to observe and informally assess articulation patterns and intelligibility issues that are not caused by oromotor or anatomical abnormalities but relate to issues such as accent, lack of dentures or habitual articulatory settings. However the poor reliability reported in assessing non-dysarthric dysphonic and normal subjects on these aspects with the Vocal Profile Analysis (Webb 2005) suggests scalar tools may not be the optimal form of assessment, with a categorical division more appropriate. Thus the design of this thesis could have included simply noting adentulous or dysarthric patients.

Degree of regional Accentedness of speech did not undergo further piloting once added to the SToPS, nor was it included in the 3 hour training session. This may have had a bearing on its limited levels of reliability. Although Expert SLTs attained only "moderate" intra rater and "fair" inter rater agreement for this parameter it is important to discuss clinical implications of discounting it from further use. Clinical observation has shown that a marked non native English accent combined with a non-optimal tracheoesophageal voice can cause intelligibility issues. Even laryngeal speakers with an accent and no speech or voice pathology can require increased listener effort to understand their speech. Again simply noting the type of accent rather than attempting to rank it would help address this issue especially as listener familiarity rather than the degree of accent may be the key issue. This thesis was undertaken in an area with few non native English speakers. In more multi-ethnic areas it is potentially more crucial to consider this non voice feature, especially when considering Intelligibility (the most basic level of laryngectomy outcome measure).

The final problematic parameter "Positive Paralinguistic Features" is a global parameter that included prosody, diction and intonation. Again it appears to lack construct validity in view of its limited inter rater agreement. However SLTs had "good" intra rater agreement and this suggests they may have some internalised representation of this parameter although the poor agreement between SLTs indicates they are using the parameter idiosyncratically.

One more clinically relevant and reliable solution may have been to have replaced this with the parameter “intonation”, and a clear definition of what listeners should be focusing on in this respect. This was considered but discounted as laryngeal literature had demonstrated it to be unreliably rated with the Vocal Profile Analysis (Webb et al 2004). However tracheoesophageal voice is very different to laryngeal dysphonia and it may be possible to arrive at a valid and reliable characterisation in alaryngeal perceptual assessment. This is a subject for further work.

A final reflection on the four parameters above concerns pilot studies in general involving ‘expert’ raters. Even a large panel who agree can show bias (Streiner and Norman 1995 p5). Although experienced SLTs were in agreement that factors such as articulatory precision may impinge on perceptual alaryngeal voice assessment, none had any experience in research and scale design nor sufficient familiarity with the literature to provide an expert opinion on the feasibility of their recommendations. With hindsight the inclusion of these factors was in excess of the scope of one thesis. Furthermore it would have been prudent to have tested out these suppositions with a further pilot which may have demonstrated the difficulty in assessing these parameters with the current definitions and scale design.

Some further design limitations of this investigation warrant discussion. The Rater Guidance notes (Appendix A) were intended to provide assistance to raters. With hindsight and in the light of studies in laryngeal voice appearing after this section of the thesis was completed that highlight enhanced reliability from such textual anchors (Awan and Lawson 2009), then tighter definitions would have been more optimal. Furthermore, Fluency was not fully encompassed by the guidance notes as it failed to clearly specify the dimensions along which to rate blocks and pauses due to neoglottal spasm and their relationship to speech fluency and rate. This thesis defined this parameter purely in terms of speaking rate in relation to the number of syllables per breath group. Fluency and speech rate (and articulation rate) are in fact independent variables – one can have dysfluent speech that is either fast or slow, one can have fast or slow speech that is either fluent or dysfluent. A future study will need to take this into consideration if insights into effects of fluency and rate on perception of speech are to be investigated. Additionally it may be necessary to increase specificity of what is understood by an impairment of Fluency i.e. whether it should include all or some of these aspects: a) reduced rate of speech, b) presence or not of pauses of a given duration, c) repetitions of sounds, syllables and words, d) filled (e.g. um, er) versus unfilled pauses, e) reduced number of syllables per breath group. However additional considerations in fluency have been highlighted in a recent thesis that indicates this is a highly complex parameter (Kendall 2009). Kendall demonstrated the key factor perceived by listeners as a change in rate is primarily due to changes in pause. Furthermore the mean number of syllables per breath group is determined by the type of speech sample, speech rate is affected by the length of utterance and pause and speech rate are stylistic effects when reading aloud. Future studies should consider investigating any such effect of stimulus type and material in addition to clarifying definitions of Fluency.

Areas for further research in view of design limitations of this thesis

This section highlighted several limitations to the current study from the point of view of design and scope. Key areas of design that would benefit from further investigation will now be considered:

a) It is important to investigate alternative methods of noting lack of dentures, impaired articulation, reading ability and accent as highlighted above.

b) The guidance notes for raters require clarification and amendment especially in relation to Fluency as previously detailed above. The SToPS is currently being included as an outcome measure in a PhD thesis (Coffey work in progress). Findings and rater feedback have been requested and can be considered for incorporation into future revision of the guidance notes. One hurdle facing any assessment of features that contribute to perceived severity or key components in speech and voice output concerns the intermittent versus constant presence of perceptual features. The SToPS only specified Stoma Noise should be rated as present even if occurring intermittently on the grounds that it is such an atypical quality and can mask the speech/voice signal. However further investigation and clarification of the effect of constant versus intermittent features as other parameters may also warrant such criteria for categorical rating e.g. Strain, Wetness.

c) Longitudinal and pre and post treatment use of the SToPS as an outcome measure would confirm whether the scale is sensitive to change. It would also be important to assess reliability across raters regarding the change and to relate this to patients ‘self perception. Such findings would provide further evidence regarding validity of the scale especially in relation to key management variables (e.g. voice prosthesis type or surgical interventions).

Limitations of scope

Further limitations in this thesis concern aspects that cannot be considered within the confines of one study. Firstly the selection of quadratic weighted kappa co-efficients allowed individual rater behaviour to be analysed in detail to obtain a range of scores but this did not permit investigation of whether there is a statistically significant difference according to profession or expertise. Further analysis with intra class correlation would allow this issue to be addressed.

There was no consideration of the influence of voice stimuli i.e. the effect of order of presentation of voice stimuli as severity of a voice quality affects the rating of subsequent voice samples. This was discussed previously (2.2.3) in relation to evidence from the laryngeal voice perception literature (Kearns and Simmons 1988; Gerratt et al 1993; Kreiman et al 1993; Kreiman et al 1998) but has never been investigated in relation to tracheoesophageal voice quality. Similarly this thesis did not analyse whether some tracheoesophageal voice stimuli are more reliably rated than others. Again this has been demonstrated in laryngeal voice where this type of investigation has been suggested to improve understanding of the factors underlying voice quality perception (Kreiman and Gerratt 2000). Some tracheoesophageal voices are more complex, in the sense that moderate to severe hypotonic and hypertonic voice qualities would seem likely to be judged to include a larger number of uni-dimensional features e.g. strain, wetness, whisperiness, reduced volume, reduced intelligibility. This contrasts with speakers who have neutral tonicity voice quality which would be more likely to be associated with the absence of such qualities. However, even if some stimuli are found to have more rater agreement it is still essential to assess the full range of voices that can occur post-laryngectomy and scales should include parameters that represent the full breadth of potential voice qualities. Research could then aim to focus on methods of improving agreement for more complex stimuli. This aim is discussed more fully below in future investigation directions.

This thesis did not include any statistical investigation of systematic rater variation to account for reliability problems but informal analysis of the data did not suggest this had occurred. There is no evidence from the laryngeal literature that systematic bias is a feature of dysphonic voice perception (Kreiman and Gerratt 2011).

The investigation of the SToPS used a cross-sectional cohort of patient voice stimuli who varied in their interval since surgery from 3 months to 17 years. Further credence to the reliability and validity of the SToPS would be gained if the SToPS could be demonstrated to be sensitive to the measurement of perceived change in voice quality in a longitudinal study of tracheoesophageal speakers.

In this study hearing acuity of listeners was based on self report of acuity levels. It is unclear whether precise acuity levels within the normal range have a bearing on outcomes in studies such as this. A future study might introduce pre-recruitment audiological screening. As this was a pragmatic study involving 'typical' listeners, it was deemed that a broad self and clinician report would be adequate.

Areas for further research in view of limitations of scope in this thesis

As regards limitations of scope, future work could fruitfully develop the SToPS through examination of the following:

a) Investigate whether rater type and expertise are linked to statistically significant differences in agreement.

b) Investigations to determine if there are any aspects of systematic bias in raters (including in relation to profession).

c) Improving agreement appears particularly important if data collection is to be undertaken by ENT raters or SLT raters with less experience in voice perceptual assessment. Investigations in laryngeal voice perception suggest that using auditory and textual anchors and perceptual skills training can increase reliability by assisting with this complex cognitive process. It would be important to include both the training and textual references for anchor point protocols used in this thesis if replica studies were undertaken. However further research could include longer training and the use of perceptual auditory anchors (during training and rating) to determine their potential to improve reliability further. Stoma noise is a key parameter but the inter rater reliability limitations even for Expert SLTs indicate increasing agreement would be a key area as it needs to be included used in routine clinical goal setting and outcome measurement. Also, voice stimuli from this thesis could be used with moduli to ascertain if professional raters may have greater agreement when employing direct magnitude estimation methods (Eadie and Doyle 2002b; Eadie and Doyle 2002c).

d) Research to determine how ratings of uni-dimensional parameters may combine to i) determine how raters allocate scores for the global parameters and ii) contribute to how raters judge the auditory impression of tone also appear a fruitful areas for further research. This would help select anchor stimuli for training or during rating tasks which in turn may increase agreement of raters which is an area for further study as outlined above (b) for the key parameters. As tone is a key determinant of the Overall Grade parameter any such patterns of uni-dimensional parameters would provide an evidence base for interventions that aim to improve these aspects of voice quality e.g. Strain would be expected to improve with botulinum toxin injection; Wetness may be reduced with surgical interventions that decrease hypotonicity or pooling of boluses/saliva. The new rank and sort software (NeAR) (Gould et al 2012) may also be useful in allowing more subtle differences in voice quality to be investigated in relation to improvements in the uni-dimensional parameters as it permits stimuli to be listed in order of severity rather than the zero to three equally appearing interval scale in the SToPs where judges must commit to distinct categories.

e) Studies to investigate whether certain voice stimuli are more reliably rated than others and to determine the underlying perceptual parameters that may account for such variation. This could then lead to investigations of improving agreement as outlined in b).

f) Investigation of the order of presentation of stimuli and any effect of how this influences the severity of judgement of subsequent stimuli.

g) The knowledge base of the inter-relationships of all tracheoesophageal voice assessments in relation to tracheoesophageal voice perceptual analysis can now be furthered by investigations utilising the SToPS. This will provide evidence of criterion validity of all the investigations. A most promising avenue it now opens is to investigate the perception of the parameter Tonicity i.e. comparing SToPS perceptual ratings to measures from videofluoroscopy, intra-oesophageal manometry or tracheal manometry. However these instrumental measures are not entirely objective and comparing instrumental and perceptual measures is not without its challenges. Videofluoroscopy and the other techniques do not represent gold standard assessments. There is no evidence that videofluoroscopic assessment has a sufficient level of inter and intra rater reliability in measuring post-laryngectomy structures. Furthermore, researchers have claimed it is unable to assess the total closure pattern of the lateral walls, the duration of closure, tenseness of the musculature or viscosity of the mucosa (Lundstrom et al 2008). Intra-oesophageal manometric measurement on even the standard anatomy of non-laryngectomised patients is problematic i.e. correct placement depends on concomitant videofluoroscopy (Ergun et al 1993) and marked intra patient variations relate to stress causing pressures to rise and catheter placement itself having the potential to produce a physiological reaction in the oesophageal and hypopharyngeal muscles (Wilson 1997). Such difficulties mirror the situation in laryngeal voice rating where acoustically, physiologically or kinematically measured dimensions do not necessarily correlate with perceptual impressions, Nevertheless, establishing the relationship of tracheoesophageal voice perception to instrumental measures has not been possible to date as a sufficiently developed perceptual tool has not been available. The SToPS now permits such studies to proceed.

The focus of the last sections was on expert judges. Attention is now turned to the issues around naïve judge performance.

6.3 Is the modified SToPS a reliable and valid perceptual assessment of tracheoesophageal voice by naïve raters?

The third research aim of this thesis was to design a modified scale for naïve listeners based on the SToPS, and to investigate its inter and intra rater reliability with these judges (4.3).

This section will look at how far the aims have been achieved in terms of design and development, and how these relate to validity and reliability. From this, possible strengths and limitations are identified with directions for future progress outlined.

6.3.1 Reliability

The intra and inter rater weighted kappa co-efficients are detailed in Tables 4.7 and 4.8 (4.3.3). The parameters will be referred to with the same nomenclature used in the professional version of the SToPS to facilitate ease of discussion. However simpler parameter terminology was used in the investigation to assist raters' understanding of the voice qualities (Table 3.4). Parameter numbers were also reduced - scale format was altered for all parameters except Overall Grade and Social Acceptability to a simpler present/absent rating scale.

Two key findings were identified in this study. Firstly, only two parameters (Overall Grade and Social Acceptability) achieved "good" intra and inter rater reliability. Secondly, these two parameters were the only ones that demonstrated superior intra compared to inter rater agreement; the reverse was found for the less reliable parameters. Such a pattern of inferior intra rater agreement suggests there is an unstable internal representation for these parameters in keeping with the poor inter rater agreement as intra rater reliability is generally expected to be superior. Both findings will be discussed and related to both the professional rater outcomes and to the previous literature.

Parameters that achieved "good" agreement

The naïve judges' superior agreement with two global parameters was in keeping with the findings for professional raters both in the SToPS and in previous laryngeal perceptual studies (Dejonckere et al 1993; de Bodt et al 1997; Dejonckere et al 1998; Millet and Dejonckere 1998; Munoz et al 2002; Webb et al 2004; Zraick et al 2011). Relating the findings of this thesis to the previous literature is difficult. Section 2.2 Table 2 detailed the published studies. There is limited research and only two studies used more robust methodology and truly naïve judges (van As et al 2003; Nagle and Eadie 2010). The majority recruited SLT students and the problems and issues around using such judges were discussed in 2.2. Van As et al (2003) pre-empted naïve raters' ability and removed the Overall Judgement parameter from their naïve scale deeming this group as incapable of making this type of judgement (on the grounds that their internal standard is laryngeal voice). Nagle and Eadie (2010) reported the same findings as this thesis with their global parameter, "Speech Acceptability" attaining both "good" agreement and higher intra than inter rater agreement. However statistical analyses in Nagle and Eadie did not account for chance agreement and the paired comparison scale was an easier rater task than the equally appearing interval scale in this thesis. Consequently this study is the first to demonstrate that naïve raters can achieve "good" agreement for simple global parameters with an equally appearing interval scale.

The findings of the Naïve rater scale in this thesis suggest this group have a degree of stability of judgement against some type of internalised psychoacoustic system for these concepts. The discussion of the professional rater study above postulated: a) clinicians had an internalised reference of optimal tracheoesophageal voices against which to rank voice stimuli and b) experience of ranking dysphonic voices may facilitate Expert SLTs to achieve a greater number of "good" agreement classifications. The naïve raters' co-efficients for these parameters were very similar to Expert ENT raters who had prior exposure to many voices. It is important to consider a potential model for this ability in naïve judges. Firstly in the absence of any pre-existing internalised reference point they may have assessed voices in terms of the difference of the stimulis from "normal voice quality". Secondly they could have merely polarised their ratings; judging all voices as relatively severe would cause high levels of agreement. Formal analysis of such systematic rater bias is outside the scope of this thesis. However, informal examination of the data tables showed naïve raters used the full range of scores to judge voices with the exception of one judge who never used the zero i.e. optimal baseline for Social Acceptability. The only parameter that appeared to show some trend of polarisation was "hard to understand" (Intelligibility) as most judges rated the stimuli to indicate they did not find any issue with intelligibility in the majority of instances. The third possibility to account for this good naïve rater ability is that they may have learned an internal baseline during the orientation session when they heard a sample of a tracheoesophageal voice and were informed this would be classed as a “good/excellent” speaker. This means they are no longer truly naïve listeners and their ratings may then not match real life encounters which would be expected to involve one, isolated tracheoesophageal speaker. Studies from germane areas indicate that naïve judges’ ratings may alter appreciably even on the basis of brief exposure to stimuli or training. This issue of potential naïve rater learning is discussed further in 6.3.3.

When considering the evidence of naïve rater reliability with overall grade type rating scales in laryngeal perceptual studies, comparisons are again restricted as most studies recruited SLT students. One previous study that included truly naïve raters has reported high levels of reliability for Overall Severity (Eadie et al 2010a) although again statistical analysis did not account for chance agreement. Gould et al (2012) reported naïve judges achieved statistically significant similarity to Expert SLTs using a different task i.e. sort and rank software for the GRBAS Overall Grade parameter. This is a key finding as it demonstrates that naïve raters have the same concept of dysphonic voice quality and can differentiate overall grade in a similar manner to trained judges.

This thesis’s second key finding, of superior intra to inter rater agreement for all parameters (with the exception of the two global parameters Overall Voice quality and Social Acceptability) is the opposite pattern to the one observed for professional raters with the SToPS and from previous laryngeal perceptual studies. The fact that the exceptions concern the only two parameters that were classified as “good” agreement may relate to these parameters having a more stable internal baseline as outlined above.

Parameters that failed to reach “good” agreement

Three key factors are associated with these parameters: a) they all required a simpler binary “presence/absence” response which would be expected to be associated with higher reliability, b) with the exception of Intelligibility, they are uni-dimensional and require specific detection from within the voice stimulus rather than an overarching impression and c) they had less intra than inter rater agreement in contrast to the parameters that had "good" agreement. One explanation for the findings is that naïve raters are unable to reliably sub-divide these complex voices into their constituent parts. Whilst Expert ENT performed similarly to Naïve judges for Overall Grade and Social Acceptability there was a trend for their raw mean co-efficient scores to be higher for these parameters, especially for test-retest. Although ENT raters do not normally perceptually assess voice quality they may be more able than naïve judges to reliably identify more complex aspects of the voice signal. This may be because they have a more stable, internalised representation against which to assess these factors. Alternatively the naïve judges may not find the nomenclature of the parameters to represent meaningful constructs against which to map the voice stimuli psychoacoustically. The consistent pattern of low intra rater reliability for these parameters is an additional indicator that raters lack a stable internal representation for these factors. Further research is needed to establish whether this is an inability to isolate certain voice qualities or whether they perceive the signal differently and use different perceptual strategies which require different terminology.

A further possibility is the low intra rater agreement was due to a learning effect whereby they had become more adept at rating by the re-test session causing a contrast to the ratings of the first session where they were less skilled. This aspect of naïve learning was highlighted by van As et al (2003). However the methodology in this thesis included a one year gap before re-test making the learning hypothesis less likely. More detailed analysis of rater behaviour and patterns is outside the scope of this investigation. This aspect should be investigated in further studies.

There is minimal research to support or refute this study's findings that more complex parameters had poor reliability and a pattern of inferior intra to inter rater agreement. Two studies showed "good" inter rater agreement for uni-dimensional parameters (van As et al 2003; Nagle and Eadie 2010). However both studies have statistical concerns as highlighted before and the latter study was a simpler paired comparison task. Van As et al reported much lower intra than inter rater agreement for naïve raters but this pattern was also observed for two-thirds of the parameters judged by SLT raters and may relate to scale design or to choice of statistics. Van As et al carried out further analysis of naïve rater behaviour and demonstrated this group: a) consistently rated parameters lower than the SLTs and b) could not use the scales to differentiate the uni-dimensional parameters that make up the tracheoesophageal voice quality.

They concluded this is because they have an internal reference of laryngeal voice quality. Such clustering of ratings at lower points of the scales would account for van As et al's naïve judges achieving such high inter rater co-efficients i.e. superior to those of the SLTs. There is limited research from the field of laryngeal voice to support van As et al's conclusion that naïve raters' difficulty with more complex constructs relates to their internalised baseline of laryngeal voice. This psychoacoustic ability presents challenges even for expert professional raters analysing laryngeal dysphonic voice. It may be expected that naïve raters would have problems in identifying and grading nineteen parameters simultaneously and it may not relate to the tracheoesophageal voice quality per se. Furthermore studies have demonstrated rater difficulty in partitioning continuous variables in other fields e.g. visual, tactile (Schiavetti 1997). One tracheoesophageal voice study required naïve judges to rate just two parameters, one global and one uni-dimensional Vocal Effort (Eadie al 2010a); high inter and intra rater agreement was reported but again statistical selection did not account for chance and clustering of scores enhancing reliability was not considered.

A further important consideration is that the naïve raters in both van As et al's study and this thesis did not have the training offered to the professional raters. This is necessary to preserve the ability of naïve raters to perceive the alaryngeal voices in the same manner as members of the community encountering tracheoesophageal voice for the first time. However studies in laryngeal perceptual analysis have demonstrated training programmes can facilitate naïve listeners to detect subtle perceptual differences (Chan and Yiu 2006; Chan et al 2012). It is then not surprising that they may not achieve the same levels of agreement as the trained raters. The issue of clinical use of naïve raters and whether training is appropriate will be discussed in the future directions sub-section below.

A final consideration is whether naïve judges are a homogenous group or whether they are sufficiently heterogeneous to consider some as more skilled at perceptual voice analysis. Naïve raters with musical training showed better intra rater reliability for breathiness and roughness perception (Eadie et al 2010b). Similarly, naïve raters have varied ability in understanding the speech of dysarthric speakers (McHenry 2011). McHenry reported there was no effect of age, gender or education on raters' ability and it was concluded there is currently no identifiable reason why some are more skilled at decoding less intelligible speech, but factors such as musical ability, foreign language training or similar could usefully be investigated in future work.

6.3.2 Validity

The naïve version of the SToPS cannot be considered to be an entirely valid scale as the co-efficients for most parameters failed to reach the level specified in section 4.2.2 that characterise "acceptable" agreement according to Landis and Koch (1977). The findings will be related to the key aspects of validity theory below; these were summarised in 1.6.2.

Content validity

The naïve scale was devised as an adaptation of the professional scale. This modification was necessary to achieve the third research aim of comparing the perceptions of tracheoesophageal voice of professional raters and members of the community. An alternative method of addressing content validity could have involved composing a completely new scale. For instance, a naïve panel could have ranked voices in severity then reached consensus regarding terminology they felt best described tracheoesophageal voice quality. This would be a fruitful area for future research. However this more robust method of content validity would be likely to produce different nomenclature to the professional scale and prevent direct rater type comparisons. Similarly, if the initial scale had been developed by a naïve panel then adapted for professionals it may not have been clinically relevant nor encapsulated the main aspects of tracheoesophageal voice quality identified by the literature review. Naïve and professional raters may require separate parameters for the scale to be valid for each group but more research is required.

One naïve rater completed a pilot of the modified naïve version of the SToPS and reported no difficulty in using the scale to rate voice samples. A pilot of more raters may have been beneficial. However, feeling ease at using the scale and being able to reliably rate voices are two different aspects as the rater has no sense of their performance as they use the scale.

One previous study used a modified professional scale for naïve judges but did not simplify scale format (van As et al 2003). As previously outlined, van As et al's naïve raters scored higher for inter rater agreement than SLTs but this was due to naïve scores clustering at the more severe scale points. Furthermore the naïve rating pattern indicated they did not use the Uni-dimensional parameters to differentiate between voices. This highlights how reliability can occur without validity. Evidence from van As et al (2003) and this study suggests that only simpler overall impression type parameters may have content validity for this group.

Criterion validity

The difficulties with this aspect of validity were highlighted in the discussion of the professional scale 6.2.2; there is no gold standard against which we can compare tracheoesophageal perceptual scales. This is the first study to have compared a large number of naïve and professional raters with statistical analyses that control for chance agreement. The naïve raters showed concordance with the professionals for both global parameters. This provides some evidence of the validity as both rater types appear to have some agreement regarding these hypothetical constructs. However analysis did not extend to examining perceptual biases or behaviours. Consequently it is possible that naïve raters scored all voices as more severe as reported by van As et al (2003). Further evidence for the validity of global parameters being used by naïve judges could be obtained if they were found to differentiate severity of voices across the whole cohort of voice stimuli and if further testing with other naïve raters provided similar findings.

Construct validity

Key aspects of construct validity testing were summarised in the professional rater study 6.2.2. It is more difficult to establish for naïve raters: a) whether the scale measured what it intended and b) included the key variables of the phenomenon. The original SToPS was designed to be clinically relevant for professional raters and encapsulate the main aspects of tracheoesophageal voice. However naïve raters cannot reliably differentiate and rank uni-dimensional parameters from voice stimuli (van As et al 2003). This finding was replicated in this thesis using a simpler yes/no format. Such failure in validity relates to either a) the terminology not having meaning for raters to allow for reliable perceptual analysis or b) the problem is a psychoacoustic perception issue because naïve raters do not have the skills to breakdown the voice stimulus into component parts. Only further studies would confirm these hypotheses.

It is also important to consider the parameters that were reliable in the modified SToPS. Raters appear to have some commonality of internal construct for Overall Grade and Social Acceptability to achieve reliability provided it was not due to rating all voices as severe as discussed above. Judges were not given a specified baseline against which to compare the voices but in the absence of tracheoesophageal voice experience it is assumed to be against normal laryngeal voice quality. If other naïve raters were able to reproduce or increase reliability levels reported in this thesis it would provide weight to naive judges having some stability of construct of rating atypical voices.

6.3.3 Summary of the study of naïve inter and intra rater reliability including its limitations and areas for further research

The aim of this study was to develop a modified version of the SToPS to investigate how members of the speakers' community perceive tracheoesophageal voice. This is the first study to include statistical analysis that accounts for chance agreement when investigating the perceptions of this rater group. Satisfactory levels of reliability were achieved for only two global parameters but this is sufficient as an outcome measure for this rater group.

A number of limitations in this study warrant consideration. Again these relate: a) to issues of study design and methodology and b) to the scope of what can be investigated in one thesis.

Limitations of design

Several design issues warrant discussion. The naïve raters did not assess the uni-dimensional parameters or Intelligibility with a four point scale like the professional raters as it was replaced with a binary yes/no response. It may have been more optimal to have investigated these as a four point adjectival scale. The rationale for trying a more simple scale relates to van As et al (2003) finding that naïve judges could not utilise bipolar seven point scales effectively. However the adjectival scale used in this study for Overall Grade and Social Acceptability may have been simpler for naïve judges to use and understand than the numerical scale selected by van As et al.

The written guidance sheet (Appendix C outlined in 3.3.1) provided for this group during the orientation session also warrants discussion. Raters were requested to imagine how they would feel if they or a close friend or relative had voice like the tracheoesophageal speaker. However the scale format (Appendix D) only required judges to tick boxes of adjectives, not to relate the voices to themselves or their family. It may have been more optimal to also request this on the rating form.

The written guidance for Social Acceptability advised judges that the parameter aimed to reflect how they see others reacting to each tracheoesophageal speaker's voice quality. This included subjective adjectives i.e. whether they perceive it would be attractive or pleasing or unpleasant to listen to. Furthermore they were given the analogy that this is similar to how accents may be perceived whereby "some people love to hear certain accents but find some grating and hard to listen to". It would have been more optimal to have avoided such subjective language that does not provide clear definitions for raters and instead just have requested it concerns how others would perceive the voice in terms of social acceptability.

The naïve rater is valued for their representation of a real life everyday encounter with the public but the test-retest design may potentially compromise this viewpoint. This thesis attempted to control for any familiarisation causing desensitisation or learning effect by delaying the re-test for one year. However it is possible some degree of learning occurred even within the first rating session as so many voices (n =55) were included. Blom et al (1986) asked groups of naïve raters to assess only small numbers of voices then analysed the results as a whole. Undertaking a similar design in a future investigation would be cumbersome and time consuming but would control for a learning effect.

Areas for further research in view of limitations of scope in this thesis

The importance of addressing the following gaps in the evidence base will now be considered:

a) The naïve cohort recruited for this study was not matched to either: i) the patients or ii) to the local community in terms of age, socioeconomic demographic, education level or gender. Although the naïve recruits provided a perspective from the local community they cannot be considered equivalent to the patients' peers or representative of the range of people who the patient may encounter in normal life situations. Furthermore naive recruits were selected on self report of normal hearing levels. It later emerged that two raters possibly had some presbyacusis. Audiometric screening as part of recruitment would have been beneficial. However the patient age and employment demographic does mean their peers are likely to have age related or industrial hearing loss. Specific investigation of hearing impairment versus normal hearing in raters and its effect upon judgement of tracheoesophageal voice seems essential. This would provide clinically relevant insights to inform patients of how they may be perceived and how intelligible they may be to different members of their community.

b) Although naïve judges rated Intelligibility and Volume this does not necessarily represent how speakers would be perceived in real life communicative situations. The voice stimuli were recorded and played back in a quiet research setting. However one key study (Clark 1985), reported older naïve raters (mean age 57 years) judged intelligibility of alaryngeal and laryngeal speech recorded over background noise less successfully than younger judges; this was attributed to presbyacusis of the older rater group but could equally have been due to cognitive factors. Investigating real life communicative situations in relation to varying demographics of naïve raters would provide valuable clinical insights as highlighted in a) above.

c) With respect to reliability, it would be important to establish the reproducibility of the "good" inter and intra rater agreement for the two global parameters with a separate cohort of naive raters and different voice stimuli. This thesis did not include investigation of rater bias or systematic polarisation in the use of the scales. Although informal inspection of the data tables did not suggest this had occurred, this has previously been observed in one naïve cohort (van As et al 2003). Further research does seem warranted to confirm or dismiss this previous finding. It is difficult to establish the validity of the Naïve rater scale as the aims of this thesis necessitated a modified professional rater scale format, not a scale specifically designed for this group. As this study confirmed van As et al's (2003) findings that naïve raters have difficulty with rating uni-dimensional aspects of voice quality it is important to consider whether it is fruitful to carry on replicating research requiring this group to use modified versions of professional scales that require advanced psychoacoustic ability. A tool specifically devised for naïve raters may ensure greater validity but would potentially prevent easy, direct comparison to ratings of professional judges. However professional and naïve rater scales can serve different purposes and future research may be more optimally focussed to investigate the unique view of each rater type rather than persistently comparing perspectives. A specifically designed naïve rater scale, as opposed to a modified version of a professional scale, could measure the outcomes found to be most pertinent to representatives of the community. Furthermore this would ensure greater validity as parameter nomenclature with the greatest meaning to this group could be specifically selected. Such parameters may well be different to the scales selected for professionals where there is a requirement the tool allows sensitivity to change and differentiates speakers to measure clinical management options.

d) This thesis' design included informing judges that the one voice stimulus they heard in the pre-rating orientation session would be considered one of the best tracheoesophageal voice outcomes. However the written guidance specified raters should not be influenced by the anchor stimulus and allocate poor or adequate categories to all samples if this reflected their personal judgement. One study (O'Leary et al 1994) selected a mid point speaker as the training example. It would be useful for future studies to investigate the effect of pre-training anchors for naïve raters and how this may help reliability but potentially influence and hence sacrifice their unique perspective of the reaction of hearing this type of voice for the first time.

e) Further studies could include alternative study design. This could include alternative rating techniques e.g. direct magnitude estimation to ascertain whether this enhances naïve judges' reliability to confirm or dismiss the previous findings of (Eadie and Doyle 2002a; Eadie and

> Doyle 2005a). Furthermore the potential learning and fatigue effect of rating large numbers of stimuli may be counteracted if naïve raters were required to assess just one or a few voice stimuli with no test-re-test. This methodology was undertaken by Blom et al (1986) where groups of naïve raters judged just five voice stimuli per group although it would be challenging to recruit such a large, demographically representative naïve cohort.

The focus of this section was on naïve judges; the following and final part of this chapter will now examine the inter rater relationship of all five types of judges included within this thesis.

6.4 An analysis of the relationship between Expert SLT, Expert ENT, Naïve, Patient and Carer raters

This section involves the fifth and final research aim: the investigation of how five different types of rater perceived the patient voice stimuli. This is the first investigation to cover this variety of judges and was undertaken to compare whether patients perceive their own voice in the same way as other important rater groups. The key finding from this study was there is generally very little agreement between the different groups as previously summarised (Table 18 5.4.2). The pattern of agreement was lower than the findings from the professional and naïve rater studies. There are no "correct" scores; they are a reflection of the different perspective of each rater type and such results may be expected given the viewpoint and different expectations of each group. It also has important implications for how we judge SVR success (and from whose perspective) as well as how we determine therapy goals to improve voice quality.

This section will examine each group's comparison to the others (6.4.1, 6.4.2, 6.4.3, 6.4.4) commencing with the patient group. The Expert SLT and Expert ENT raters will not be compared as these findings were previously discussed (6.2). Each sub-section will include a comparison of the findings from this thesis to previous studies reported in the literature. This will include research in laryngeal perceptual analysis and related SLT fields due to the paucity of studies into tracheoesophageal voice perception, especially from a patient and carer perspective. This section will conclude with an outline of the strengths and limitations of this study along with areas and directions for future research (6.5.5).

6.4.1 The Patient group in relation to other rater types

Patient versus Carer

These two groups had the highest inter rater agreement of all the nine comparator groups with three of the seven parameters reaching “good” agreement and two falling only marginally below the cut off point for this category. Such a pattern of agreement may relate to the amount of time they spend together and the intimate nature of the relationship. The carer has a unique perspective of observing the patient communicating in a variety of settings and being privy to the patient disclosing their feelings about their communication. Furthermore both raters share the same personal perspective of judging voice in relation to survival from cancer and in this study rated the voice from a questionnaire rather than an audio recording. The most agreement was for Intelligibility and Volume. This may relate to these parameters being less hypothetical constructs and possibly the easiest to understand; also both patient and carer are likely to receive feedback (overtly or covertly) regarding these factors from other listeners in real life communicative situations. However “Wetness” (the third parameter with “good” agreement) does not fit into this hypothesis. It is a more complex uni-dimensional parameter for which other comparator pairs achieved only “poor” agreement. Without further analysis of this pattern of rating it is not possible to establish whether these groups are better at detecting it or if they equally disregard it as insignificant.

Surprisingly, Overall Grade attained only “moderate” agreement in contrast to its superior reliability for the professional and naïve studies. This appears likely to demonstrate there is variation between how the patient and the carer view voice quality. Further investigation of these differences would be required to determine if this is systematic or random variation between rater types.

Patient versus Naïve

This comparison reflects whether the wider community perceive the voice quality in the same way as the patients themselves. The limited agreement for even the global parameters (mean co-efficients of 0.30-0.43) highlights the considerable difference in how these two groups view voice outcome. In contrast, one previous investigation reported that Naïve and Patients rated similarly for Overall Severity and Vocal Effort for laryngeal voice (Eadie et al 2010a). However tracheoesophageal voice can be markedly different to laryngeal dysphonia and naïve raters may judge this more severely, whereas SVR patients may be less concerned about their voice due to context as previously discussed. Definitive conclusions cannot be drawn without further investigation of rater patterns and bias.

Patient versus Expert SLT

The importance of this comparator relationship relates to the aspirations and expectations of the patient and those of the Expert SLT. The SLT sets therapy goals and assesses the voice in clinical contexts in order to plan management. This management includes the selection of a voice prosthesis or further investigation/treatment to improve voice quality (e.g. referral for videofluoroscopy or botulinum toxin treatment). It is therefore crucial that the SLT is aware of how their views may complement or contrast to patients’ perspective. This study’s findings highlight major differences between patient and SLT perceptions. All perceptual parameters had only “fair” agreement with the exception of “Wetness” which had “poor” agreement. Again there was no analysis as to whether this is systematic bias. One previous study compared SLT and tracheoesophageal patients' ratings of voice (Heaton et al 1996) and found that SLTs generally rated voices more favourably than the patient. This was attributed to a potential SLT bias as voice rating scores reflect the outcome of SLTs' treatment. However only one SLT was included and was not blinded to the patients, so limited conclusions can be drawn. Several investigations of laryngeal voice perception have analysed Patient /SLT perceptions but, again, there is no clear consensus. Studies have reported patients rate higher than SLTs (Sapir et al 1986), lower than SLTs (Lee et al 2005) or not significantly different to SLTs (Liu et al 1998; Eadie et al 2007; Eadie et al 2010a). The two latter studies also reported that patients appear to be using different perceptual strategies than SLTs.

Sapir et al (1986) concluded that it was unsurprising their patient cohort rated their voices more favourably following surgical nerve resection to treat spasmodic dysphonia. They attributed this to perspective as SLTs rate what they hear but patients also include kinaesthetic perception and their recollection of their severely dysphonic voice prior to surgery. Such factors may be equally applicable to tracheoesophageal speakers in addition to other contextual issues postulated by the authors i.e. the voice may be perceived favourably when patients have suffered physical and psychological effects of voice loss and the patient may be unable to recall their pre-morbid voice.

The mixed picture from the literature regarding patient and SLT voice perception may reflect variations in pathology, severity of the condition, whether the condition is life threatening and bias of the patient or SLT. The patient has a complete perspective of their functioning in a variety of environments whereas SLTs normally rate voices in a clinic or a research setting from an audio recording in a soundproofed room. SLTs are not rating in real life situations that reflect the experience of the patient.

A study of Patient/SLT perception of dysarthria (Walshe et al 2008) raised other important considerations: a) Patients hear their speech via bone and air conduction whereas SLTs hear by air conduction alone and b) the large discrepancy between groups due to perspective can make goal setting and progress difficult with patients being perceived as indifferent or over anxious. These equally apply to SVR rehabilitation and further research into this area is likely to yield valuable insights to inform clinical practice.

Patient versus Expert ENT

The unique feature of this comparator pair relates to Expert ENT surgeons being responsible for selecting surgical techniques (e.g. flap or closure type) which is likely, in turn to impact on SVR voice quality (Deschler et al 2004; Alam et al 2008; Yang et al 2011). ENT surgeons do not undertake formal perceptual assessments but informal impressions of voice outcome can potentially influence the surgeon's techniques and whether post-laryngectomy interventions are offered to ameliorate voice quality.

There was a similar lack of agreement for Patient versus Expert ENT as there was Patient versus Expert SLT; the same issues of perspective that were highlighted above are equally applicable to Patient versus ENT. This thesis only included surgeons who had not operated on any of the patient subjects in order to eradicate a possible surgeon bias.

The highest agreement in this sub-section reached only the "fair" classification for the parameter of Intelligibility. It is unclear why this is more reliably rated than Overall Grade and could simply be a chance finding. Further investigation is required, but one potential explanation is patients regard themselves as intelligible but this does not equate to acceptance of their voice quality. Only one previous study has investigated Patient versus ENT perception (Heaton et al 1996) but only included one surgeon who was not blinded to the patients. The lack of studies in the laryngeal literature is surprising given that surgeons may make decisions on surgical intervention and judge the success of their surgery based on how the voice sounds.

6.4.2 The Carer Group in relation to other groups

The carer group spend the most time with the patient in the role of support and (in many cases) confidante. Their unique perception is observing how the patient is perceived by others in real life situations and being privy to perceptions of family members/strangers regarding the patient's speech; any negative perceptions may not be disclosed to the patient to protect them and for this reason carers were interviewed separately to patients. This extensive background information may have influenced their rating whereas naïve and professional raters judged solely from the audio recording.

Carer versus Naïve

Both carer and naïve groups reflect a non-expert perspective of tracheoesophageal voice but the carer is emotionally involved and has contextual information and personal involvement. Again such differences in perspective are likely to explain the "poor" agreement for all parameters except for Overall Grade ("moderate") and Social Acceptability ("fair"). Intelligibility had exceptionally poor agreement and it is difficult to argue that this is because the concept is too difficult for untrained raters to understand as this contrasts to this parameter having the highest agreement in this study for Carer/Patient. Further analysis to determine any systematic bias in the patterns of variation was outside the remit of this study. One systematic difference that may lead itself to analysis would be the effect of naïve raters judging in a research setting whereas carers report on real life perception. Alternatively, carers have possibly adjusted to the new voice and find it intelligible or are not concerned in relation to survival. More research is needed as Carer views have not been represented in the tracheoesophageal or laryngeal perceptual analysis literature to date, though the study here affords an important start.

Carer versus Expert SLT

Understanding perspectives between these rater groups is vital as carers are frequently involved in SLT goal setting and home therapy practice. They also provide feedback to the SLT about how the patient is functioning outside the therapy situation. Agreement patterns were generally higher than for Patient versus SLT although no parameters reached "good" agreement. The differences are once more likely to be due to perspective; the carer may be unconcerned as their relative has survived cancer and voice is not a key issue. Alternatively SLTs may fail to appreciate the severity of the patients' difficulties as they are accustomed to interacting with more severely communication impaired people than tracheoesophageal speakers. A further consideration is SLTs are more able to separate out voice into its uni-dimensional parameters in relation to untrained judges; consequently "poor" reliability would be expected for the uni-dimensional parameters. Again no studies have previously investigated the differences in these group's perceptions.

Carer versus Expert ENT

This inter-relationship reflects what carers may feed back to the surgeon at ENT reviews regarding how the patient is coping and interacting. As discussed in the Patient versus ENT section, discrepancies of perception are especially pertinent if surgical solutions for functional voice issues are possible e.g. botulinum toxin, neopharyngeal surgery. The agreement co-efficients were similar to those of Carer versus SLT but Overall Grade was lower. Further analysis would be needed to assess whether this indicates both professional groups had a similar perspective to each other in relation to carers. The co-efficients for Carer versus ENT were broadly similar to Patient versus ENT except Social Acceptability and Volume were slightly higher. Further analysis is required to assess if there is systematic bias present. This could reflect either patients being less or more concerned about their voice in relation to their surgeon. This is important to investigate due to the ENT surgeons' reliance on their auditory impression to judge patient outcome.

6.4.3 Naïve Raters in comparison to professional raters

Naïve versus Expert SLT

This rater interaction reflects the difference between how the SLT perceives the voice in relation to the perceptions of the patient's local community. SLTs' therapeutic aim is to facilitate voice rehabilitation to a level where patients can interact with unfamiliar listeners and consequently it is crucial SLTs have an understanding of how strangers may judge voice in relation to a professional assessment. There appears to be some commonality of perception as the second highest co-efficient in this study was for Overall Grade which achieved "good" agreement for this comparator pair. Naïve and professional raters also achieved similar inter rater agreement within their own separate studies of reliability (6.2.1 and 6.3.1) this provides further evidence that naïve and speech and language therapist raters have a similar overall impression of voice quality. Social Acceptability only failed to achieve "good" agreement by 0.01, again indicating some commonality of perception of this parameter.

The discussion section regarding naïve rater reliability (6.3.1) postulated that Naïve may agree with each other for Overall Grade and Social Acceptability as they may be polarising ratings to the lower end of the scale as reported by van As et al (2003). These findings of agreement between

Naïve and SLT suggest that this may not be the case as naïve raters agree not only with themselves but also with professionals. Polarisation would be expected to equate to lower agreement in this sub-section unless professionals are also polarising in the same pattern. However an informal review of data tables did not appear to confer with this hypothesis.

Agreement for Intelligibility was surprisingly low in relation to the other global parameters as it is less likely to reflect poor understanding of the construct. It is unclear if this is random or systematic pattern bias. It would be important for further studies to investigate any patterns as therapy goals often target this aspect. A further consideration relates to scale format. Naïve listeners used a binary presence/absence of Impairment of Intelligibility whereas SLT used an equally appearing interval scale 0-4. Statistical analysis required SLT scores of 2 or 3 to be amalgamated to a 1 to reflect naïve scoring. SLTs were potentially detecting subtle differences and reflecting these in a rating of 1 (mild) whereas Naïve may have not indicated presence of impaired intelligibility unless there was a marked effect. Consequently the findings of this study could reflect task design rather than perception per se.

The failure to achieve agreement for the uni-dimensional parameters may reflect the scale format differences outlined above or naïve rater difficulty in analysing voice stimuli into its more complex perceptual components. These parameters were not reliably rated in the naïve version of the SToPS which suggests the latter explanation may be more relevant. Previous studies of Naïve versus SLT perceptions of voice disorders and intelligibility are inconclusive. Perceptual scale assessments for tracheoesophageal voice intelligibility reported naïve raters evaluate higher than SLTs (Finizia et al 1998) and similarly to SLTs (Cullinan et al 1986; Bridges 1991a). The latter two studies have similar statistical issues as discussed in the professional rater study. Naïve raters and SLT have been shown to rank overall severity similarly in laryngeal voice (Gould et al 2012) and Intelligibility in dysarthria with direct magnitude estimation (Walshe et al 2008). Studies of formal intelligibility transcription assessments reported SLTs scored more highly (Doyle et al 1989; Bridges 1991a) or similarly to naïve judges (Finizia et al 1998; Walshe et al 2008).

A seminal study for laryngeal voice (Kreiman et al 1990) asked Naïve and SLT/ENT raters to judge whether pairs of voices were similar-dissimilar; this avoids forcing raters to behave in ways not consistent with perceptual processing. All naïve raters used similar perceptual strategies but professional raters differed in the parameters they considered to be important to assess similarity. Such research could be replicated in tracheoesophageal voice analysis for naïve raters as opposed to further investigations of equally appearing interval scales which would seem likely to replicate rater type dissimilarity.

Naïve versus Expert ENT

This rater pair reflects how ENT surgeons perceive the patient in relation to community members. There was generally lower agreement than observed for Naïve versus SLT but one common parameter, Overall Grade, also reached “good” reliability. This is a key parameter which also attained “good” agreement in the Naïve only study. This suggests that the overarching view of voice will be viewed similarly by both rater types. However all uni-dimensional parameters had “poor” agreement. This may relate to score rationalisation as discussed in the Naïve versus SLT comparator pair or may relate to the difficulties naïve raters have with agreeing on the components they perceive as key features (Kreiman et al 1990) as discussed above. The Intelligibility co-efficient was very low. This may relate to highly different perceptions of this parameter, difficulty in detecting and rating this aspect or reflect differences in the importance of this parameter within the signal (Kreiman et al 1990).

Only one previous study has compared Naïve to ENT raters (Misono et al 2012). They found ENT juniors had some superiority of reliability with the parameters Roughness and Breathiness before and after training. This was tentatively attributed to ENT having more musical training in keeping with Eadie et al (2011) who reported this factor related to superior rater ability in naïve judges. As discussed in the naïve rater discussion (6.3.1), naïve raters have usually been treated as a homogenous group and further investigations are required.

6.4.4 Summary

This is the first study to carry out a comprehensive rater overview; all judge types have key roles in research and clinical practice and differences are likely to reflect variance in perspective.

There was limited agreement between the rater groups for the majority of parameters. The Patient and Carer groups had the most agreement followed by both Expert SLT and Expert ENT in relation to Naïve raters. The scope of this thesis does not include investigation into the pattern of these variations nor whether they are statistically significant. Inter rater differences potentially relate to random variation, some rater types systematically judging parameters more severely than others, due to contextual perspective or differences in perceptual focus on different aspects of the voice stimuli.

There are limited previous investigations with which to compare this study. Previous studies do not reach consensus regarding patterns of rating according to the type of judge. However there is some preliminary evidence from laryngeal voice perception and dysarthria intelligibility perception that Naïve and Patient raters use different perceptual strategies and differ in the parameters they consider to be the key aspects of a voice stimulus. The uni-dimensional parameters showed the least agreement in this thesis and this may relate to these being more complex for untrained raters to assess.

6.4.5 Limitations of the study to compare SLT, ENT, Naïve, Patient and Carer raters and areas for further research

Although the major differences in agreement between rater types may reflect rater perspective, a number of limitations of this study may account for some of the inter rater variance and warrant discussion. Again these relate: a) to issues of study design and methodology and b) to the scope of what can be investigated in one thesis.

Limitations in relation to design

The different rater groups in this thesis used different scale types and scale formats. This required some scores to be amended to enable inter rater comparisons i.e. the professional rater scores for the uni-dimensional parameters Whisper, Strain, Wetness and Volume and the professional and patient/carer scores for Intelligibility were altered where appropriate from scale points 2 and 3 to 1. This was to enable comparison to the naïve raters who could only use scale points 0 (absent) or 1 (present) for these parameters. Although this permitted statistical analysis, it is possible that professional raters with a wider range of scale points partitioned their auditory perceptions differently e.g. scale point one to indicate a mild presence of an attribute whereas Naïve and Patients may have only indicated the presence of the same parameter with scale point one if it was moderate or severe. Furthermore the professionals used a zero to three equally appearing interval scales whereas Naïve and Patients/Carers used a four point adjectival scale. This again may have influenced how parameters were psycho-acoustically partitioned. However the professional raters had identical adjectival markers to aim to counteract any such effect.

The parameter Volume was rated by ENT/SLT and Naive from voice stimuli obtained by reading aloud in a sound proofed room. Professional rater guidance specified to judge against a baseline of normal conversational volume for a laryngeal speaker but Naïve were given no guidance and potentially may have made a judgement baseline in relation to whether they perceived the speaker would be loud enough in all environments. A further potential for varying baselines stems from Patients/Carers rating Volume in relation to their needs. However some patients' daily lives may involve speaking in louder environments and/or with communicative partners with hearing impairment in contrast to others who interact just with partners with normal hearing in one to one quiet settings. It would have been advisable to have counteracted these potential variations: a) by instructing Naïve to judge volume with the same guidance provided for the SLT/ ENT raters and b) by specifying Patients/Carers judge their volume in relation to their needs in a quiet one to one setting with listeners with normal hearing (to reflect the same criteria for the professional and naïve raters). Similarly judgements of the parameter Intelligibility could have been influenced by the variation in Volume baselines outlined above. Intelligibility is linked to vocal volume in loud environments and Patients/Carers who encounter such settings may have allocated lower scores for this parameter in response to those who socialise in quiet locations. This could account for reduced agreement with naïve and professional raters who simply had to rate in a research setting. Furthermore patients with regular interface with hearing impaired peers (or carers observing such interactions) may rate Intelligibility as inferior to professional and naïve judges who assess with normal hearing from a sound proof room recording. Future research of Volume and Intelligibility would need to be carefully designed to account for such discrepancies.

This thesis' methodology is probably not comparable to real life exposure to tracheoesophageal speakers for naïve judges or for SLTs/ENT surgeons in their clinical work i.e. rating fifty-five voice stimuli in rapid succession. The issue of Naïve peers with hearing impairment was previously discussed (section 6.3.3). This can equally apply to Patients and Carers where they are reporting on communication with hearing impaired friends and family. Furthermore research studies of professional judges in laryngeal voice observed such simultaneous rating tasks may cause raters to behave differently than in clinical situations due to fatigue and rating judgements being influenced by the order and severity of preceding stimuli. It would be reasonable to suggest that this effect would be equal (if not exacerbated) in untrained listeners.

Different nomenclature was included for parameters according to rater type and this potentially affected rater perception. Further investigations to ensure raters find parameters meaningful is essential.

This thesis did not undertake a detailed pilot study for the Patient and Carer rating scale. However scale item nomenclature and selection was set by the fifth aim of this thesis i.e. to investigate inter judge agreement of five different rater groups. Qualitative comments from Patients/ Carers were requested and recorded at the end of the rating session. The scope of this thesis does not allow for them to be included but no reservations about the structure or format of the scale were expressed.

Not all patients and carers who were invited to participate opted to be included in the study. Consequently this self selected group may show inclusion bias. A further consideration relates to the audio recordings not being undertaken at the same session as the patient and carer ratings. However this is not considered to have had an impact on the study as no patients had any issues that would have changed their voice quality e.g. any treatment, recurrence of cancer or change of voice prosthesis type and all were in the chronic/stable phase post surgery as seen by the median age at recording of 3.5 years.

Areas for further research in view of limitations of scope in this thesis

Future work could address the following key aspects:

a) The investigation of patients and carers was cross-sectional and consequently their perceptions were sampled at different points post surgery (range 3 months to 17 years; median 4.0 years). Naïve and professional judges were not informed of the time elapsed since each speaker's surgery. This is potentially a crucial factor for patients and carers who may rate the voice more positively as time elapses and adjustment occurs. Conversely, they may initially show indifference to poor voice quality in relation to survival but become more negative about their voice after the euphoria of surviving abates and they are confronted with the altered voice when they resume their typical communication contexts. No test-retest investigation was undertaken in this thesis to assess such factors. It is important to investigate how patient and carer judgements of voice outcome change: i) over time or ii) in the early stages post-operatively on a day to day basis (perhaps in relation to mood or progress perceived). Further studies could ascertain if professional raters could detect subtle differences that patients report or if the patient rating is attributable to mood rather than actual perceptual variation.

b) Further studies are required to analyse any systematic bias or patterns of inter rater variation and the effects of context (patient and carer) or task beyond the analyses carried out here. There was no investigation of systematic versus random rater bias (as highlighted throughout this thesis) in relation to the investigation of inter rater agreement for the five rater types in this section. As highlighted before (6.3.3), naïve raters were not demographically matched and it would be important to ascertain if including such a naïve cohort affected agreement.

c) It would be beneficial to investigate the parameters that best predict patients perceiving they have a successful outcome (using factor analysis or regression analysis). This would allow all surgical options to be evaluated based on a patient perspective. However it must always be considered that patients can be affected by perspective and may not be rating on perception of voice alone. Other rater groups' judgements may demonstrate more agreement to Patient/Carer judgements depending on time elapsed since surgery.

d) Finally it would be important to design further studies with methodology that reflects real life communication rather than continuing to focus on research type settings. This could include studies of patient volume and intelligibility against a variety of types of background environmental noise and varying degrees of hearing ability in the listener.

6.5 Clinical implications

This section will first consider the main points in relation to the SToPS' strengths as a valid and reliable clinical tool. This is followed by an itemisation of the main clinical gains/ implications.

This is the first tracheoesophageal scale with sufficient reliability and evidence of some aspects of validity to use in clinical and research settings. However it is essential to consider the type of rater and the individual parameters in relation to reliability. Although it is not possible to predict how future judges from the various sub-groups may rate other tracheoesophageal voice stimuli, this thesis suggests Expert SLTs achieved the most parameters with the previously defined "acceptable" reliability levels. The sole exception to this is for the parameter Stoma Noise which did not reach sufficient agreement (Landis and Koch (1977). The SToPS cannot be concluded to be a valid and reliable tool for SLTs who are not experienced in perceptual voice rating or for ENT surgeons; this is based on the arbitrary cut off point of 0.61 or above constituting the “acceptable” level of agreement for including voice parameters in clinical scales (Hirano 1989; Webb 2005). However all professional rater groups’ agreement is sufficient for Overall Grade to be considered a reliable and valid outcome measure. Similarly there is evidence that the naïve raters can be considered reliable judges of Overall Grade and Social Acceptability pending further research determining whether any bias/polarisation is preventing the SToPS’ validity. The Patient/Carer ratings having low agreement with the naïve and professional groups does not compromise the validity of the scale as they are likely to bring a different perspective and this lack of reliability should not necessarily be seen as a limitation of the scale.

There are many clinical implications from the evidence in this research:

a) It would be important to consider education programmes to increase the perceptual assessment skills of all SLTs working in SVR so they can readily participate in routine clinical assessment, outcome measurement, audit and research. Alternatively to investigate and implement ways of improving agreement as highlighted in the previous section.

b) Now that a clinically relevant and practical assessment tool has been developed it can be used to engage in studies that map physiology and anatomy to different voice quality components to further our understanding of all these aspects. This would provide further evidence for the criterion validity of the SToPS. This should ideally involve Expert SLTs until it can be established whether those who are less expert achieve the consensus level of agreement deemed to reflect acceptable reliability.

c) The SToPS can facilitate the initiation of research to provide evidence based selection of surgical techniques, voice prostheses and botulinum toxin protocols that may offer the optimal voice quality outcome. Although the Overall Grade parameter is a useful over arching assessment it does not encapsulate the whole of tracheoesophageal voice quality and its key differentiating features. The parameter Tonicity is a key determinant of tracheoesophageal voice quality (Hurren et al 2009) and is crucial to include as an outcome measure for such aspects of research. Again, Expert SLTs would appear the optimal raters for these and other key parameters until further evidence is obtained that others can achieve the classifications of "good" agreement or replica studies determine that raters with less expertise can be reliable judges. The SToPS has not yet been demonstrated being sensitive to measuring the changes that the protocols requiring investigation seek to bring about. Consequently such studies could address both aspects simultaneously.

d) A further application for the SToPS would be to select one or more parameters to use as outcome measure to use in research specifically designed to investigate outcomes in head and neck cancer. This should include surgical and non-surgical treatment (radio +/- chemotherapy) as these can have a marked effect upon outcome of voice quality. Similarly it would be crucial to include investigation of the effect of co-morbidities. Overall Grade would be an ideal parameter for this purpose due to "good" mean reliability for ENT and SLT raters in this study. However rater bias needs to be considered as clinicians may covertly or overtly rate voices more favourably when units are judged on their outcome results by peer review. This would be difficult to control for unless units were requested to ask another cancer unit to rate their patient voices from recordings. This would be time consuming but the only way to ensure any bias was eliminated. Alternative strategies for storing samples such as cloud sourcing may be a way forward so units could access and rate others' data anonymously and blinded. However there would be Information Governance issues to address including informed patient consent.

e) Naïve raters are less likely to have any rater bias towards treatments or patient outcomes and in this thesis demonstrated “good” reliability for two global parameters. However this study found they had “good” agreement with Expert ENT and Expert SLT raters. If future studies confirm there to be no major difference of rating between Naïve and ENT it could be argued it may not be useful to include them in future audit and research. Nevertheless any decision to discount this group would need to be considered in the light of the future recommended studies regarding naïve rater demographic variation especially age/ hearing loss related and real life perceptions in communicative contexts.

f) Patient and Carer perceptions of their own or their relative’s voice quality are essential aspects of surgical outcome and should be routinely collected. They rate voice from an entirely different perspective and their views may never match that of the experts but comparing and discussing such discrepancies is a key point for rehabilitation and goal setting.

g) As this thesis used textual and auditory anchors during training, other studies seeking to replicate the professional rater study or implement the scale into clinical practice should use a similar format. The guidance notes for the SToPS will be freely available to other units seeking to use the scale along with discussion about parameters and suitable perceptual anchors. Perceptual anchors could be made available on the cloud system again if information governance clearance and patient consent was obtained.

h) Further evidence regarding the SToPS' clinical applicability relates to it being selected as an outcome measure in a PhD thesis (Coffey, Imperial College, work in progress) which aims to compare six different types of voice prosthesis. This author specified it was selected in preference to other scales as it had some demonstrated reliability and validity (Hurren et al 2009) and appeared the most clinically relevant and easy to use. The guidance notes for the SToPS were supplied in addition to discussion (by telephone and email) regarding how to implement and rate the parameters and select suitable perceptual anchors. Feedback regarding its use by the recruited Expert SLT panel was all positive with the exception of the Fluency parameter; this sole parameter caused some confusion as to how blocks should be interpreted which is in keeping with the previously outlined requirement for revision of the Fluency definition (6.2.1).

In summary, the SToPS will enable us to monitor outcome and improve patient care. These factors were identified at the beginning of this thesis as the impetus for the development of a new scale. To this end the key aim of this thesis has been fulfilled and it is anticipated that further developments and refinements will be undertaken in the future.

Chapter 7. Conclusions

The aim of this chapter is to provide a final summary of the work. All aspects have already been discussed in detail but this summary will act as a final synopsis to conclude the thesis. The five research aims outlined in section 2.6 will be listed in turn with a brief outline of the key findings and the clinical implications for each and indications for further investigation.

7.1 Aims

7.1.1 Aim 1

To devise a new scale for the perceptual assessment of tracheoesophageal voice by professional raters.

Outcome

This was achieved by constructing the scale (SToPS) based on evidence from the literature and from pilot studies and subsequent revisions (sections 3.1 and 3.2).

Future directions and clinical applications

- Further refinement and development of the scale in relation to parameter definitions and rater guidance notes are an aim of future work.

- The scale and guidance notes will be freely available to SLTs who wish to use the SToPS in clinical or research settings. Feedback of findings and clinicians' experience of the scale will be requested and incorporated into future refinements of the SToPS.

7.1.2 Aim 2

To examine the validity and reliability of the SToPS for professional raters (including examination of inter and intra-rater reliability according to rater type and expertise).

Key Findings

- Reliability was measured in terms of parameters that achieved an arbitrary level of agreement selected for laryngeal voice scales. Reliability was found to vary depending on the parameter and rater from the point of view of both professional group and level of expertise. All professional raters achieved more parameter classifications of "good" agreement for intra rater agreement indicating they had relatively fixed internal yardsticks for given parameters. By contrast this level of agreement was attained for fewer parameters for inter rater judgements suggesting individuals have differing internal yardsticks (section 4.2.3).

- Only one parameter, "Overall Grade" could be classified as achieving "good" reliability for all professional rater sub-groups.

- The SLTs achieved more "good" classifications of agreement for intra and inter rater reliability than ENT raters (with the exception of the Articulatory Precision parameter).

- Expert SLTs achieved more parameters with "good" agreement than their less expert SLT colleagues.

- The Expert SLT group achieved "good" inter and intra agreement for nine of the fourteen parameters; the variables/parameters that demonstrated lower agreement were Stoma Noise, Articulatory Precision, Paralinguistic Features, Reading Ability and Accent.

- Reliability is an inherent aspect of validity and in this respect it can be claimed that the SToPS has achieved some aspects of validity testing, particularly in relation to content validity (section 6.2.1 and 6.2.2). Typically one judges other aspects of validity against some notional 'gold standard' measure. However, there are no correct judgements in voice quality perception and no gold standard measures against which to test construct and criterion validity. Final criterion validity confirmation therefore awaits further studies in relation to manometric, videofluoroscopic and fluency software (pause and rate) assessments (6.2.3) which will give insight into how far objective measures of variables believed to underlie given parameters in the SToPS relate to perceived differences.

Future directions and clinical applications (section 6.2.3)

- To ascertain if new cohorts of raters employing different voice stimuli attain equally "good" reliability with the SToPS as present participants. If this is achieved it would provide added evidence for construct validity.

- Further evidence for the validity of the SToPS could be determined by studies to examine how Expert SLT ratings of Tonicity, Wetness and Strain correlate to videofluoroscopic and oesophageal/tracheal manometric measures. This scale can already facilitate the initiation of research to provide evidence based selection of surgical techniques, voice prostheses and botulinum toxin protocols and other management options that offer the optimal voice quality outcome.

- Investigations to establish how patterns of Uni-dimensional parameters combine to determine Overall Grade. This would enable treatments and management options to more specifically target these parameters that can influence the overall impression of voice.

- Establishing whether the scale is sufficiently sensitive to measure change is essential before more definite conclusions can be made regarding its general application to clinical practice.

- As regards sensitivity it would be helpful to investigate whether there is any systematic bias in individual and group rater judgements in relation to “good” agreement, whether raters tend to polarise scores and whether some judges consistently rate higher or lower than others, and if so why.

- Overall Grade would be an ideal parameter to use in future research into outcome in relation to surgery, non-surgical treatments and co-morbidities due to “good” mean reliability for ENT and SLT raters in this study. This would provide confirmation (or not) of its sensitivity to change.

- Research as to whether longer training and the use of auditory anchors during the rating task may improve reliability would also be fruitful. This may be particularly important if data collection is undertaken by ENT raters or SLT raters with less experience in voice perceptual assessment.

- Investigation of reliability in relation to different voice stimuli: some tracheoesophageal voices are more complex than others, whilst some may be more ‘typical’ of the various tonicity types and voices vary in perceived severity. An important future aim would be that one should study SLTs’ assessment of all varieties of tracheoesophageal voice in order to further insights into and reliability of rating more complex stimuli.

7.1.3 Aim 3

To examine the inter and intra-rater reliability of naïve raters in the perceptual assessment of tracheoesophageal voice using a modified form of the expert scale. A modified version of the professional scale was developed (section 3.3) for this purpose.

Key Findings

- "Good" levels of intra and inter rater reliability were only achieved for two parameters, "Overall Grade" and "Social Acceptability" (section 4.3).

- Intra rater agreement was not superior to inter rater in contrast to the opposite finding in the expert rater study (section 4.3).

- The baseline against which this group rate global parameters is unclear as they had no prior tracheoesophageal voice experience (section 6.3.2).

Future directions and clinical applications (section 6.3.3)

- Future investigations should ideally provide matched demographics (e.g. age, socioeconomic of naïve raters to the local community and to the patient population. This would permit SLTs to build up a more accurate picture of how patients will be perceived in their communication with their peers and other listeners they may encounter in daily life.

- Investigations of Intelligibility in more naturalistic contexts, including in the presence of environmental noise and with judges with normal versus decreased hearing would also add to our knowledge of the barriers faced by tracheoesophageal speakers' in daily living.

- The current scale was adapted from the professional scale to allow inter rater comparisons. This may have compromised content and construct validity and reliability especially for the uni-dimensional parameters in as far as naïve raters may judge according to a whole different set of parameters. An assessment specifically designed for this group with involvement of naïve judges from the inception stage would establish the optimal parameters and appropriate nomenclature for a dedicated naïve rater scale. Target population involvement would naturally potentially enhance validity too. As part of such a study investigation as to whether altered scale/scoring format (in particular direct magnitude estimation) would enhance reliability would be desirable. If after adjustments to which parameters were rated and in which way there were still unacceptable levels of rater agreement in this group it may indicate that the psychoacoustic skills required to rate voice s in these instances are not feasible for untrained, naïve listeners.

- Investigations with the NeAR (Gould et al 2012) would permit finer analysis at to whether and how naïve and professional judge rankings vary (as outlined in section 2.2.5). Any rater type discrepancies could be analysed in relation to underlying uni-dimensional parameter patterns as these are the parameters that can be ones that can be influenced by surgical or other treatment options.

7.1.4 Aim 4

To examine the patient and carer perspective of SVR voice outcomes using a modified version of the scale that had been developed for clinicians.

This aspect of the study was incorporated into the inter rater comparisons in Aim 5.

Future directions and clinical applications (section 6.4.1 and 6.4.2)

Patient and carer perspectives should be incorporated into routine clinical practice. This is particularly crucial as patients and carers report from real life communicative contexts whereas SLTs typically assess in a quiet idealised clinical setting. It would be important to investigate the factors that people with no prior knowledge/anchors for tracheoesophageal voice base their judgements upon, how one should assess these and how they alter over time.

7.1.5 Aim 5

To examine the relationship between expert, naïve, carer and patient rating of tracheoesophageal voice.

Key Findings (section 6.4)

- The greatest consensus was between patients and carers, who attained "good" agreement for three parameters. Only two other comparator groups (Naïve versus Expert SLT and Naïve versus ENT surgeons) achieved "good" reliability and for only one parameter, Overall Grade. This marked discrepancy between rater groups' judgement of Overall Grade potentially relates to the perspective (e.g. accepting impairment in relation to cancer survival and context (e.g. assessing against real life experience in contrast to controlled, research voice stimuli) of the person assessing the voice.

Future directions and clinical applications

- Whilst important facts have been established around rating, investigations involving further and different groups of listeners are desirable to increase insights into how different groups may differ in their patterns of rating. For instance, future studies could examine whether, and if so, what the nature is of systematic bias according to rater type; whether different groups utilize the scales and scale ranges differently. Any demonstrated effect of the latter type of systematic variation should be seen in relation to perspective and not a limitation of the scale (section 6.4.3 b).

- Defining the “best” SVR outcomes according to rater type. This may vary considerably between groups and again is an issue of perspective rather than a limitation of the scale.

- Initial focus group feedback suggested there may be influences of accent and other supralayrngeal voice features on Social Acceptability and Overall Grade judgements according to different rater types. This was not tackled in detail in this study but could be a subject of future research.

Appendix A. Guidance notes for the Sunderland Tracheoesophageal Perceptual Scale

1. Overall Severity Scale

Voice quality is **not** compared to normal voice for a laryngeal speaker. Rate the voice in comparison to your internal reference point of voice potential for SVR speakers.

0. Excellent - The best voice achievable for a SVR speaker; the voice quality you would judge to be the optimal outcome after laryngectomy.

1. Good – Some aspect(s) observed prevents you judging the voice as falling into the optimal outcome group.

2. Adequate - Some aspect(s) mean the voice cannot be rated as good.

3. Poor – The worst outcome for a SVR speaker.

Section A - Voice Quality Parameters

2. Perceptual Tonicity

Tonic

0. Neutral tone; neither lax nor tight.

Hypotonic (tone laxer than tonic)

1. Mildly laxer compared to tonic (Lee Marvin voice, like creak).

2. Moderately lax compared to tonic; voice may have 'echoing' sound of resonance of voice in the inflated hypotonic area. Creaky, lax feature and low pitch.

3. Severe hypotonicity for laryngectomy, but would be classed as good outcome for a jejunum or stomach graft. Obvious echoing resonance. Whisper quality is evident in the lax, inflated area. Low pitched.

4. Usually only jejunum/stomach pull-up patients display this degree of hypotonicity. The voice is severely whispery and has reduced volume compared to hypotonic 3. Echoing Resonance in the ballooning, inflated hypotonic area is severe.

5. Aphonic whisper. This differs from the aphonia in a stenosed neopharynx as air is passing through larger, laxer, ballooning area with less turbulence than a tight stenosed area. Tight stenosed voice sounds more like tense aphonia in a patient with a larynx. The volume is severely reduced. Intermittent gurgly phonation may occur due to vibration of secretions.

Hypertonic (tone tighter than tonic)

1. Mildly tenser than tonic. Quality sounds more like a dysphonic voice (in patient with a larynx). No strain.

2. Moderately tenser than tonic, but not to the degree that would be considered sufficient for botulinum toxin. Strain is evident but only mild. Volume may be reduced or louder than normal. No major effect on fluency.

3. Definitely hypertonic, moderately strained or whisper quality. Mild effect on fluency.

4. Marked hypertonic quality that is unpleasant to listen to. Voice is still functional but with marked strain and markedly reduced fluency.

5. Severe hypertonicity, fluency is severely affected and intermittent total spasm may occur. The voice is normally non-functional or cannot be used for all communication needs due to the strain required for phonation.

Stenosis

Stenosis is **not** rated 1-5; it could only be rated as a separate parameter from tonicity in its most marked form in the pilot study. Stenosis +5 should be used if no tonicity is judged to be present due to extensive neopharyngeal fibrosis. Marked stenosis causes a rigid, immobile neopharynx. Stenosis (+5) is characterized as an aphonic whisper that gives the impression of a scarred, tight neopharynx with resonance of the whisper in a rigid tube with no vibrating neoglottis. Strain may be a feature if the diameter of the rigid area is narrow. The voice often sounds similar to that of a laryngeal speaker with aphonia; N.B. hypotonic -5 has a lower resonance and is a lax aphonia. Stenotic voice quality is always associated with dysphagia for solids.

3. Strain

The amount of audible effort you perceive the patient requires to produce voice.

0. No perceived effort.

1. Mild.

2. Moderate.

3. Severe, usually associated with marked hypo/hypertonicity.

4. Wetness/Gurgliness

The perceptual feature of secretions bubbling in the neopharynx on voicing. If an intermittent feature, rate at its most severe.

0. No audible vibration of secretions.

1. Mild.

2. Moderate.

3. Severe - usually associated with jejunal grafts and hypotonicity +3 → +5. May occur with dysphagia if pooling of secretions or liquid bolus in stenosis or pouch/pseudoepiglottis.

5. Impairment of Volume

0. Conversational volume of voice judged to be within the same limits as expected for normal conversational volume for a laryngeal speaker.

1. Mildly impaired volume.

2. Moderately impaired volume.

3. Severely impaired volume reserved for voice that is whisper only Aphonia +5/-5/Stenosis.

6. Social Acceptability

If you are judging social acceptability to be impaired because of regional accent, please mark this on the rating form.

0. Social acceptability is the optimal level possible for a SVR speaker.

1. Mild impairment , e.g. mildly gurgly quality, strain etc.

2. Moderate impairment; obviously qualitatively different to a laryngeal speaker and not aesthetically pleasant.

3. Severe impairment of acceptability. “General public” would tend to turn or stare if they heard this voice e.g. marked stoma blast, echoing deep jejunal voice, severe hypertonic strain. The type of voice outcome you would dread if this subject were your relative. This parameter has the potential to link with one or more of the other parameters on the scale.

7. Whisper

The perceptual impression of whisperiness in the voice quality.

0. No whisper quality audible.

1. Mild whisper quality.

2. Moderate.

3. Severe. Total aphonia.

8. Intelligibility

Ease of understanding the speaker that would be expected for a normal laryngeal speaker, in a one to one speaking situation with no background noise.

1. Mild impairment of intelligibility.

2. Moderate impairment of intelligibility.

3. Severe impairment of intelligibility.

9. Stoma Noise

0. Stoma noise is judged to be absent.

1. Intermittent mild stoma noise; rate in this category even if a brief instance of mild stoma noise is audible in the sample.

2. Constant stoma noise even if you judge it as being relatively quiet or mild.

3. Constantly audible stoma noise that is marked and may compete with oral speech.

10. Fluency

0. Fluency within normal limits for a typical laryngeal speaker.

1. Mildly impaired fluency compared to a typical laryngeal speaker.

2. Moderate impaired fluency - 5 – 10 syllable phrasing per breath group.

3. Severely impaired fluency - phrasing of 5 syllables or less.

11. Articulatory Precision

0. “Average” precision of articulation as defined as a score of 1 or 2 out of 6 in the Vocal Profile Analysis.

1. Habitually lax articulation (score of 3 or 4 on the VPA).

2. Markedly habitual lax articulation (VPA 5 or 6) or mild dysarthria. A naïve listener would describe this as “mumbling” due to lack of precision or “slurring” due to mild dysarthria.

3. Moderate - severe dysarthria or a very severe premorbid articulation disorder; there would be marked intelligibility issues even if the subject were still a laryngeal speaker.

12. Positive Features (Diction / Paralinguistics)

This category is difficult to define succinctly. Certain speakers have speech diction, intonation and/or pause features that have an overall positive effect but are not part of the voice signal. These have the potential to positively affect the judgment of naïve listeners.

0. No specific positive features are judged to be present.

1. Positive features that are above average in comparison to laryngectomy peers; prosody (intonation) judged as present.

2. Excellent phrasing, diction and intonation.

3. Outstanding features such as noted in newsreaders, with normal or almost normal intonation present.

Parameters for Speech and Language Therapists only

13. Accent

This category is rated to control for naïve listeners unconsciously judging voices due to accent.

0. No discernible regional accent.

1. Mild accent.

2. Moderate accent.

3. Marked accent.

14. Poor Reader

This parameter is rated to control for naïve listeners assigning lower marks to poorer readers whose literacy skills affect their ability to read aloud the Rainbow passage.

0. No problems with reading aloud.

1. Mild problem with reading aloud.

2. Moderate problem.

3. Severe problem.

Appendix B. The Sunderland Tracheoesophageal Perceptual Scale

<u>**VOICE SAMPLE NUMBER :**</u> Male □ Female □

1. Overall Voice Rating

0	1	2	3
Excellent	Good	Adequate	Poor

Section A – Voice Quality Parameters

2. Perceptual Voice Tonicity

5	4	3	2	1	0	1	2	3	4	5

Hypo ← **Tonic** → **Hyper**

→ **Stenosis**

3. Strain (Audible Effort for Voicing):

0	1	2	3
	Mild	**Moderate**	**Severe**

4. "Wetness" (gurgliness) of Voice Quality

0	1	2	3
	Mild	**Moderate**	**Severe**

5. Impairment of volume

0 1 2 3

Mild Moderate Severe

6. Impairment of Social Acceptability of Voice

0 1 2 3

Mild Moderate Severe

7. Whisper

0 1 2 3

Mild Moderate Severe

Section B – Parameters not related to Voice Quality

8. Impairment of intelligibility

0 1 2 3

Mild Moderate Severe

9. Stoma Blast

0 1 2 3

Mild Moderate Severe

10. Impairment of fluency

0 1 2 3

Mild Moderate Severe

11. Impairment of articulatory precision

0 1 2 3

Mild Moderate Severe

12. Positive features of articulation (paralinguistics/diction)

0 1 2 3

Neutral good excellent outstanding

13. Accent

0 1 2 3

Mild Moderate Severe

14. Poor Reader

0 1 2 3

Mild Moderate Severe

Appendix C. Naïve rater information sheet

Laryngectomy Voice Quality Research

Naïve Listeners – Information Sheet

Thank you for agreeing to take part in this research project.

You will hear 55 voice samples. You will hear each sample twice before we move on to the next sample. All the voice samples are from people who have had throat surgery because of cancer. We are trying to see if there is agreement as to whether certain voices are judged as being more desirable than others.

If there is agreement between raters we can then look and see what the stronger and weaker voices have in common – it is possible surgery or the type of valve could be modified to try to make voice quality after surgery better.

No patients will know how you rate their voice. You may feel you want to give speakers a higher rating as it seems unkind to give them a poor score when they are trying so hard to speak. **Please** avoid this, as we need to know just how they sound to you.

Most speakers you hear are between 48 and 65 years of age. Whether they are male or female is marked at the top of each speaker's rating sheet.

As a guideline, imagine how you would feel about having a voice like the speaker you are rating (or if it was your partner, mother, father etc who sounded like that).

A few people get the best voice it is possible to achieve after removal of the voice box. The research assistant will play this voice to you so you know how the best speakers will sound to give you a baseline.

This voice would be rated as excellent or good by staff who work in a head & neck cancer unit.

It is entirely up to you how you rate each voice – you may feel all the voices are adequate or poor, as you may be shocked at how throat cancer patients end up sounding. If this is the case please mark this down – you are reflecting the views of people in the general public. It may be possible to change the voice of future patients if we can identify the worst voices and look at reasons why they have turned out this way.

You will first rate the voice quality as:

Excellent Good Adequate Poor

Then the social acceptability:

This may or may not be the same as the first rating. This reflects how you see others reacting to it – the way it is or is not attractive and would be pleasing or unpleasant to listen to. This is similar to how different people view accents e.g. some people love to hear certain accents but find some grating and hard to listen to.

The second part of the scale may be difficult for you, as you are not trained in analysing voices. Some speakers' voices sound:

Strained
Gurgly (wet)
Not loud enough
Whispery
Hard to understand

Or some can sound like a combination of 2 or 3.

Do not worry about leaving these boxes blank. The rule is – if it strikes you straight away tick the box. You may want to add extra ticks if it is so striking you feel it needs highlighting.

If you are not sure do not worry – we are mainly looking for the voice features that would be striking and obvious.

Once again many thanks for giving up your time to take part in this research.

Anne Hurren

Appendix D. Naïve rating scale

Untrained Listener Rating Scale

VOICE SAMPLE NUMBER ________

Male □ Female □

1. Overall Voice Rating

Excellent □ Good □ Adequate □ Poor □

2. Social Acceptability

Excellent □ Good □ Adequate □ Poor □

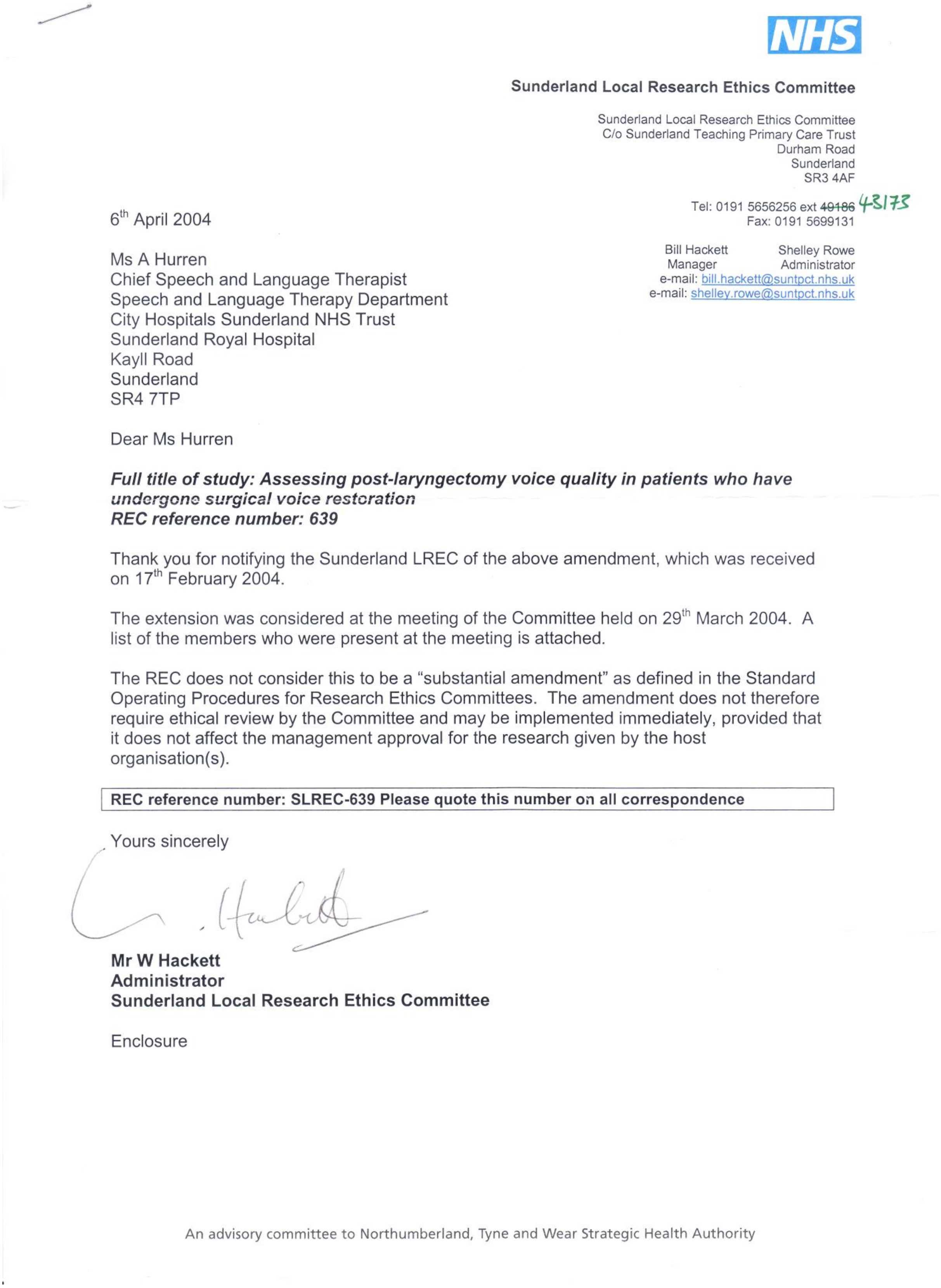
Descriptions

Is it …

Gurgly □ Strained □ Whispery □

Not Loud Enough □ Hard To Understand □

Appendix E. Ethics Committee approval letter



NHS

Sunderland Local Research Ethics Committee

Sunderland Local Research Ethics Committee
C/o Sunderland Teaching Primary Care Trust
Durham Road
Sunderland
SR3 4AF

Tel: 0191 5656256 ext ~~40186~~ 43173
Fax: 0191 5699131

Bill Hackett	Shelley Rowe
Manager	Administrator

e-mail: bill.hackett@suntpct.nhs.uk
e-mail: shelley.rowe@suntpct.nhs.uk

6th April 2004

Ms A Hurren
Chief Speech and Language Therapist
Speech and Language Therapy Department
City Hospitals Sunderland NHS Trust
Sunderland Royal Hospital
Kayll Road
Sunderland
SR4 7TP

Dear Ms Hurren

Full title of study: Assessing post-laryngectomy voice quality in patients who have undergone surgical voice restoration
REC reference number: 639

Thank you for notifying the Sunderland LREC of the above amendment, which was received on 17th February 2004.

The extension was considered at the meeting of the Committee held on 29th March 2004. A list of the members who were present at the meeting is attached.

The REC does not consider this to be a "substantial amendment" as defined in the Standard Operating Procedures for Research Ethics Committees. The amendment does not therefore require ethical review by the Committee and may be implemented immediately, provided that it does not affect the management approval for the research given by the host organisation(s).

REC reference number: SLREC-639 Please quote this number on all correspondence

Yours sincerely

Mr W Hackett
Administrator
Sunderland Local Research Ethics Committee

Enclosure

An advisory committee to Northumberland, Tyne and Wear Strategic Health Authority

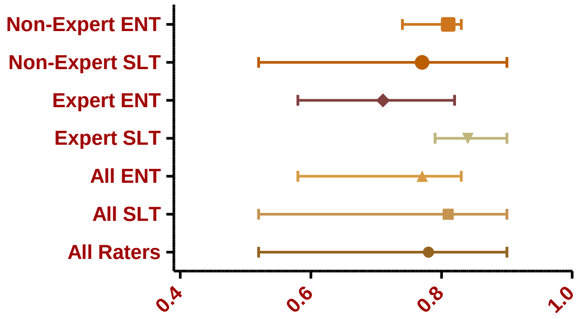
Appendix F. “The Rainbow” reading passage

The Rainbow

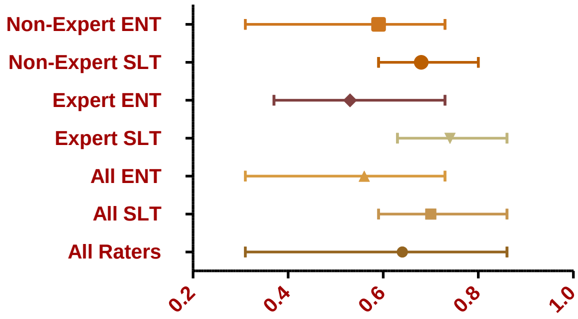
When the sunlight strikes raindrops in the air, they act like a prism and form a rainbow. The rainbow is a division of white light into many beautiful colours. These take the shape of a long round arch, with its path high above and its two ends apparently beyond the horizon. There is, according to legend, a boiling pot of gold at one end, people look, but no one ever finds it. When a man looks for something beyond his reach, his friends say he is looking for a pot of gold at the end of the rainbow.

Appendix G. Intra rater forest plots for professional judges

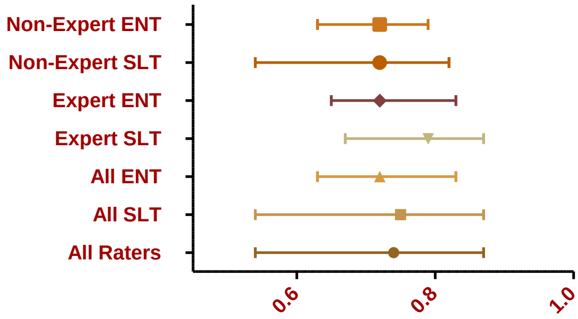
Overall Grade Intra Rater. Weighted Kappa Mean and Range.



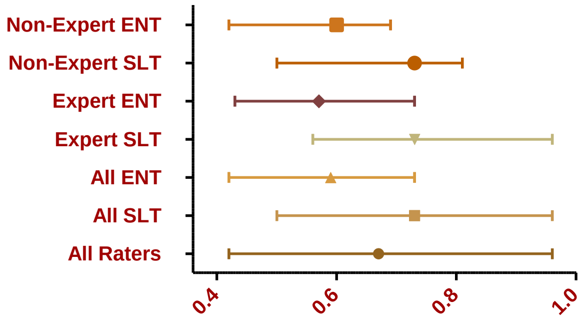
Tonicity Intra Rater. Weighted Kappa Mean and Range.



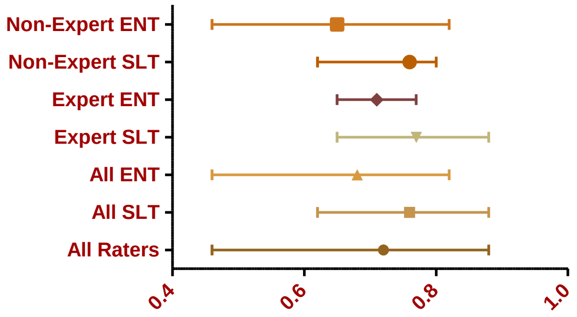
Strain Intra Rate. Weighted Kappa Mean and Range.



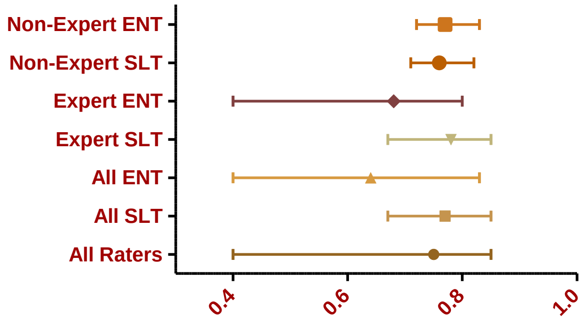
Wetness Intra Rater. Weighted Kappa Mean and Range.



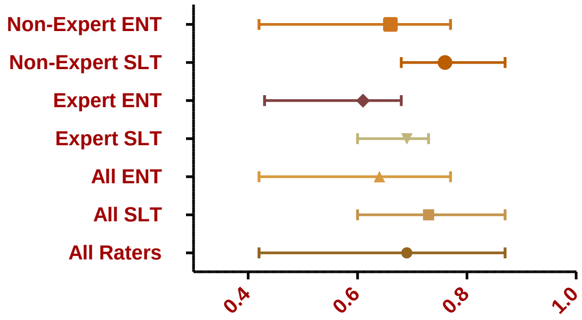
Impairment of Volume Intra Rater. Weighted Kappa Mean and Range.



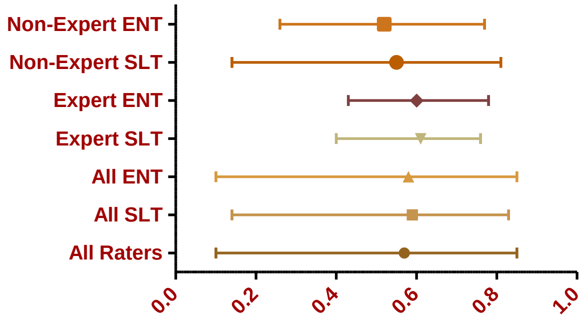
Impairment of Social Acceptability Intra Rater. Weighted Kappa Mean and Range.



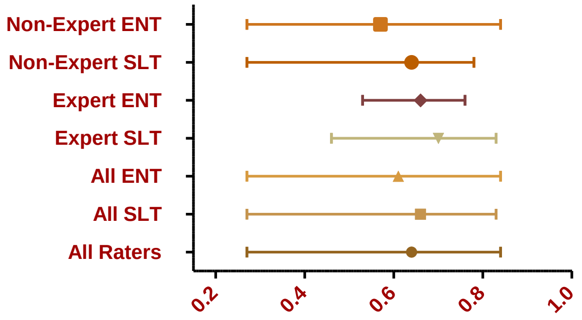
Whisper Intra Rater. Weighted Kappa Mean and Range.



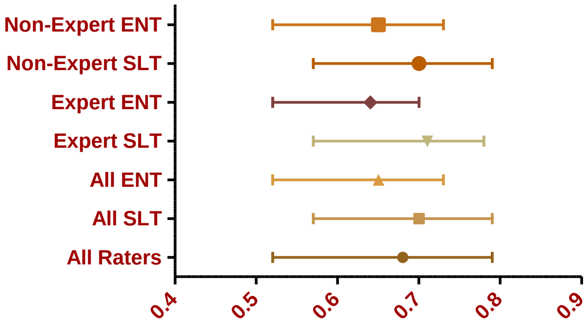
Impairment of Intelligibility Intra Rater. Weighted Kappa Mean and Range.



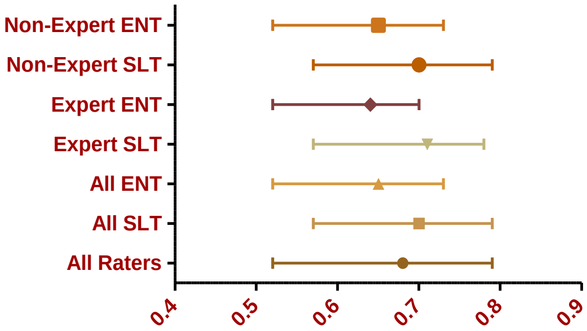
Stoma Noise Intra Rater. Weighted Kappa Mean and Range.



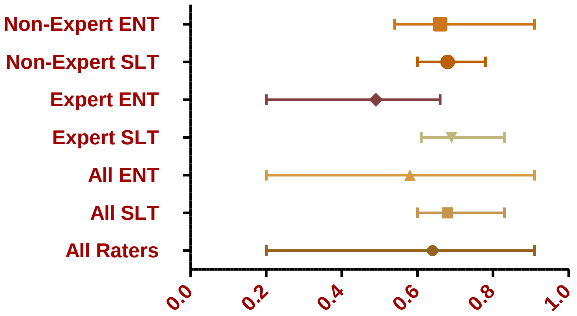
Impairment of Fluency Intra Rater. Weighted Kappa Mean and Range.



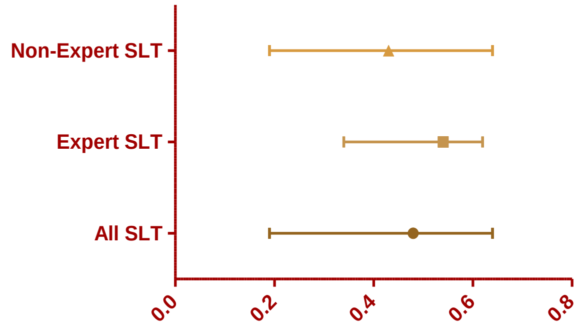
Impairment of Articulation Intra Rater. Weighted Kappa Mean and Range.



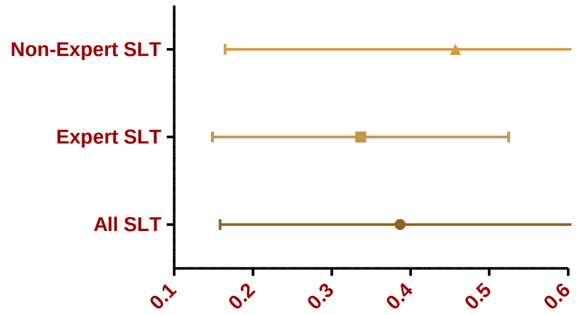
Positive Paralinguistic Features Intra Rater. Weighted Kappa Mean and Range.



Accent SLT Intra-agreement. Weighted Kappa Mean and Range.

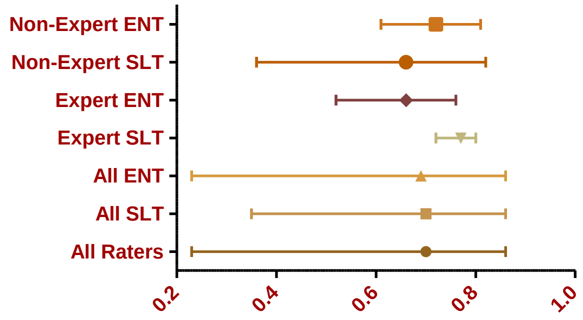


Poor Reader Intra Rater. Weighted Kappa Mean and Range.

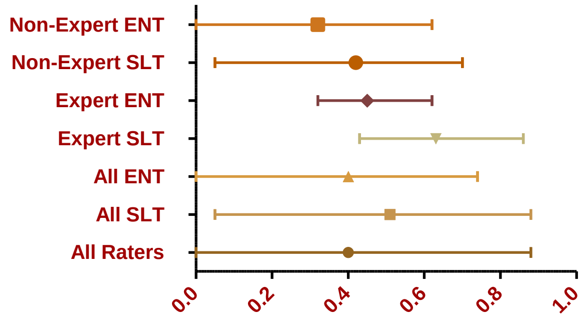


Appendix H. Inter rater forest plots for professional judges

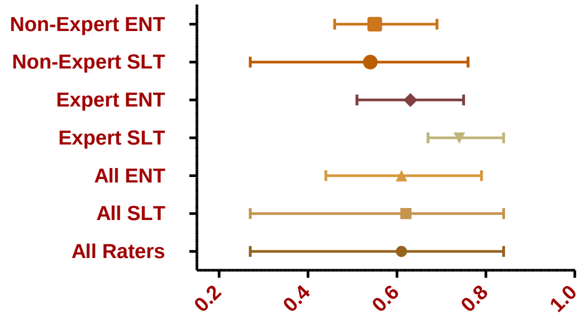
Overall Grade Inter Rater. Weighted Kappa Mean and Range.



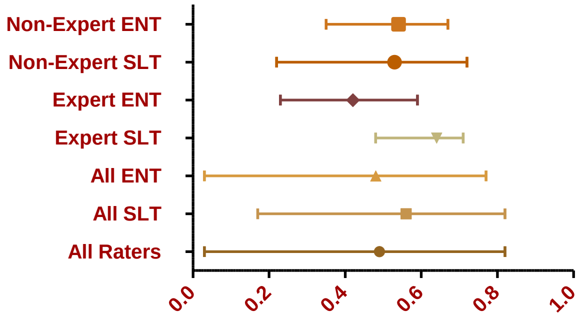
Tonicity Inter Rater. Weighted Kappa Mean and Range.



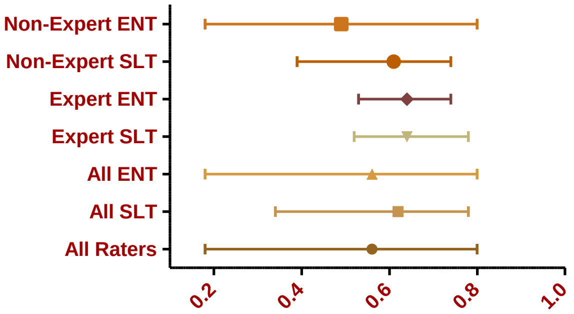
Strain Inter Rater. Weighted Kappa Mean and Range.



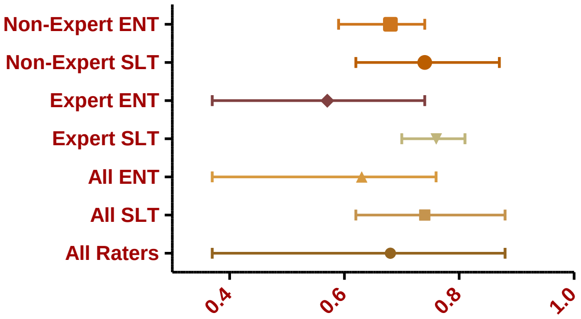
Wetness Inter Rater. Weighted Kappa Mean and Range.



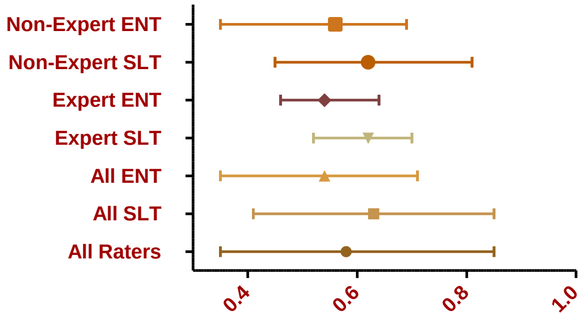
Impairment of Volume Inter Rater. Weighted Kappa Mean and Range.



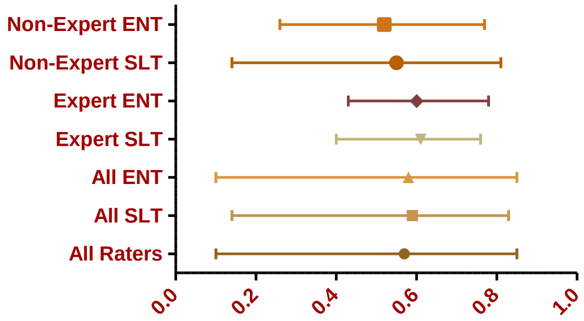
Impairment of Social Acceptability Inter Rater. Weighted Kappa Mean and Range.



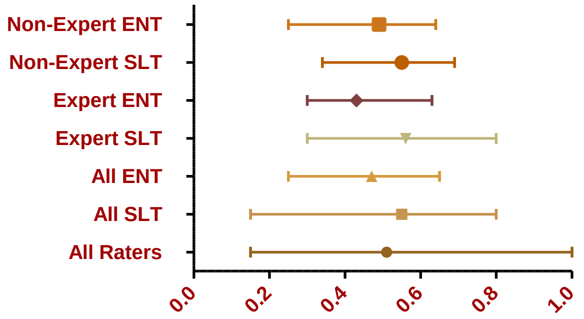
Whisper Inter Rater. Weighted Kappa Mean and Range.



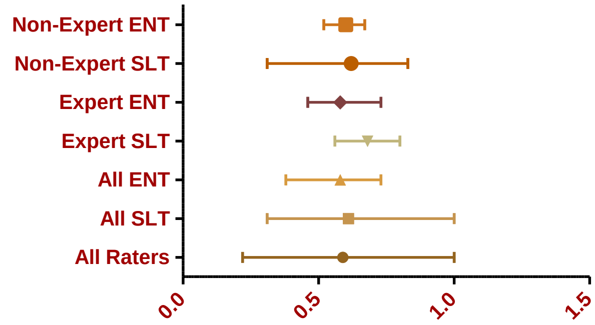
Impairment of Intelligibility Inter Rater. Weighted Kappa Mean and Range.



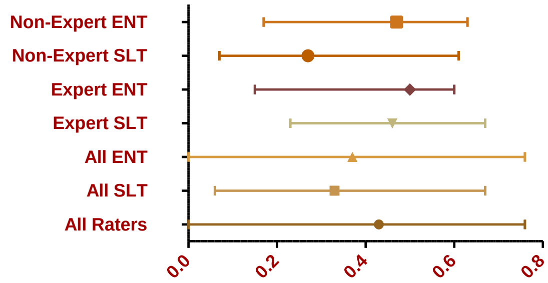
Stoma Noise Inter Rater. Weighted Kappa Mean and Range.



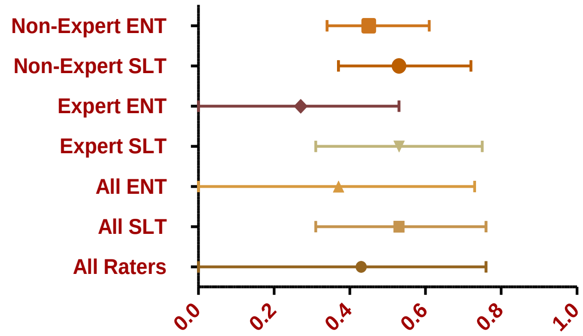
Impairment of Fluency Inter Rater. Weighted Kappa Mean and Range.



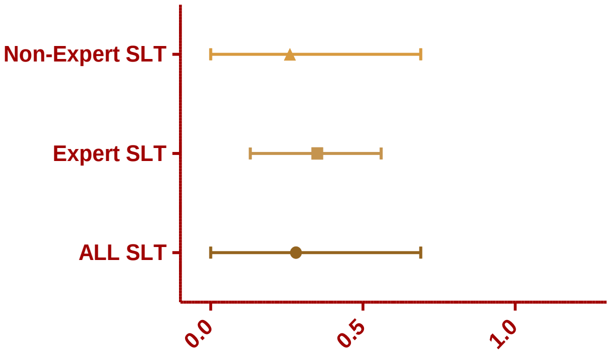
Impairment of Articulation Inter Rater. Weighted Kappa Mean and Range.



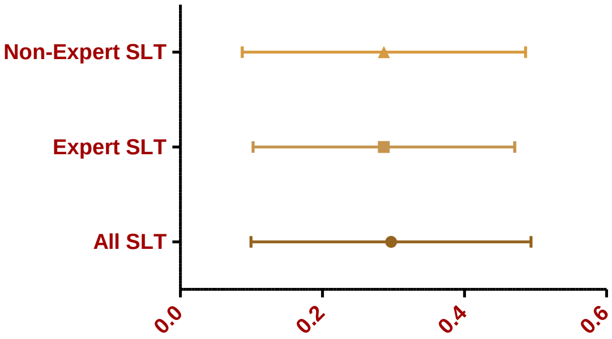
Positive Paralinguistic Features Inter Rater. Weighted Kappa Mean and Range.



Accent Inter Rater SLTs. Weighted Kappa Mean and Range.



Poor Reader Inter Rater. Weighted Kappa Mean and Range.



Appendix I. Comprehensive results tables for professional raters

* "good" or above level of agreement (Landis and Koch 1977)

Parameter 1: Overall Grade

Inter Rater

Rater type	Weighted Kappa Range (Mean)
All	0.52 - 0.90 (0.78)*
All SLT	0.52 - 0.90 (0.80)*
All ENT	0.58 - 0.83 (0.77)*
Expert SLT	0.79 - 0.90 (0.84)*
Expert ENT	0.58 - 0.82(0.71)*
Non Expert SLT	0.52 - 0.90 (0.77)*
Non Expert ENT	0.74 - 0.83 (0.81)*

Inter Rater

Rater type	Weighted Kappa Range (Mean)
All	0.23 - 0.86 (0.70)*
All SLT	0.35 - 0.86 (0.70)*
All ENT	0.23 - 0.86 (0.69)*
Expert SLT	0.72 - 0.80 (0.77)*
Expert ENT	0.52 - 0.76 (0.66)*
Non Expert SLT	0.36 - 0.82 (0.66)*
Non Expert ENT	0.61 - 0.81 (0.72)*

Parameter 2: Tonicity

Intra Rater

Rater type	Weighted Kappa Range (Mean)
All	0.31-0.86 (0.64)*
All SLT	0.59-0.86 (0.70)*
All ENT	0.31-0.73 (0.56)
Expert SLT	0.63-0.86 (0.74)*
Expert ENT	0.37-0.73 (0.53)
Non Expert SLT	0.59-0.80 (0.68)*
Non Expert ENT	0.31-0.73 (0.59)

Inter Rater

Rater type	Weighted Kappa Range (Mean)
All	0.00-0.88 (0.40)
All SLT	0.05-0.88 (0.51)
All ENT	0.00-0.74 (0.40)
Expert SLT	0.43-0.86 (0.63)*
Expert ENT	0.32-0.62 (0.45)
Non Expert SLT	0.05-0.70 (0.42)
Non Expert ENT	0.00-0.62 (0.32)

Parameter 3: Strain

Intra-Rater

Rater type	Weighted Kappa Range (Mean)
All	0.54-0.87 (0.74)*
All SLT	0.54-0.87 (0.75)*
All ENT	0.63-0.83 (0.72)*
Expert SLT	0.67-0.87 (0.79)*
Expert ENT	0.65-0.83 (0.72)*
Non Expert SLT	0.54-0.82 (0.72)*
Non Expert ENT	0.63-0.79 (0.72)*

Inter Rater

Rater type	Weighted Kappa Range (Mean)
All	0.27-0.84 (0.61)*
All SLT	0.27-0.84 (0.62)*
All ENT	0.44-0.79 (0.61)*
Expert SLT	0.67-0.84 (0.74)*
Expert ENT	0.51-0.75 (0.63)*
Non Expert SLT	0.27-0.76 (0.54)
Non Expert ENT	0.46-0.69 (0.55)

Parameter 4 : Wetness

Inter Rater

Rater type	Weighted Kappa Range (Mean)
All	0.42-0.96 (0.67)*
All SLT	0.50-0.96 (0.73)*
All ENT	0.42-0.73 (0.59)
Expert SLT	0.56-0.96 (0.73)*
Expert ENT	0.43-0.73 (0.57)
Non Expert SLT	0.50-0.81 (0.73)*
Non Expert ENT	0.42-0.69 (0.60)*

Inter Rater

Rater type	Weighted Kappa Range (Mean)
All	0.03-0.82 (0.49)
All SLT	0.17-0.82 (0.56)
All ENT	0.03-0.77 (0.48)
Expert SLT	0.48-0.71 (0.64)*
Expert ENT	0.23-0.59 (0.42)
Non Expert SLT	0.22-0.72 (0.53)
Non Expert ENT	0.35-0.67 (0.54)

Parameter 5: Impairment of Volume

Intra Rater

Rater type	Weighted Kappa Range (Mean)
All	0.46-0.88 (0.72)*
All SLT	0.62-0.88 (0.76)*
All ENT	0.46-0.82 (0.68)*
Expert SLT	0.65-0.88 (0.77)*
Expert ENT	0.65-0.77 (0.71)*
Non Expert SLT	0.62-0.80 (0.76)*
Non Expert ENT	0.46-0.82 (0.65)*

Inter Rater

Rater type	Weighted Kappa Range (Mean)
All	0.18-0.80 (0.56)
All SLT	0.34-0.78 (0.62)*
All ENT	0.18-0.80 (0.56)
Expert SLT	0.52-0.78 (0.64)*
Expert ENT	0.53-0.74 (0.64)*
Non Expert SLT	0.39-0.74 (0.61)*
Non Expert ENT	0.18-0.80 (0.49)

Parameter 6 : Impairment of Social Acceptability

Intra Rater

Rater type	Weighted Kappa Range (Mean)
All	0.40-0.85 (0.75)*
All SLT	0.67-0.85 (0.77)*
All ENT	0.40-0.83 (0.64)*
Expert SLT	0.67-0.85 (0.78)*
Expert ENT	0.40-0.80 (0.68)*
Non Expert SLT	0.71-0.82 (0.76)*
Non Expert ENT	0.72-0.83 (0.77)*

Inter Rater

Rater type	Weighted Kappa Range (Mean)
All	0.37-0.88 (0.68)*
All SLT	0.62-0.88 (0.74)*
All ENT	0.37-0.76 (0.63)*
Expert SLT	0.70-0.81 (0.76)*
Expert ENT	0.37-0.74 (0.57)
Non Expert SLT	0.62-0.87 (0.74)*
Non Expert ENT	0.59-0.74 (0.68)*

Parameter 7: Whisper

Inter Rater

Rater type	Weighted Kappa Range (Mean)
All	0.42-0.87 (0.69)*
All SLT	0.60- 0.87 (0.73)*
All ENT	0.42-0.77 (0.64)*
Expert SLT	0.60-0.73 (0.69)*
Expert ENT	0.43-0.68 (0.61)*
Non Expert SLT	0.68-0.87 (0.76)*
Non Expert ENT	0.42-0.77 (0.66)*

Inter Rater

Rater type	Weighted Kappa Range (Mean)
All	0.35-0.85 (0.58)
All SLT	0.41-0.85 (0.63)*
All ENT	0.35-0.71 (0.54)
Expert SLT	0.52-0.70 (0.62)*
Expert ENT	0.46-0.64 (0.54)
Non Expert SLT	045-0.81 (0.62)*
Non Expert ENT	0.35-0.69 (0.56)

Parameter 8: Impairment of intelligibility

Intra Rater

Rater type	Weighted Kappa Range (Mean)
All	0.26-0.86 (0.68)*
All SLT	0.60-0.86 (0.72)*
All ENT	0.26-0.80 (0.64)*
Expert SLT	0.67-0.86 (0.73)*
Expert ENT	0.51-0.75 (0.63)*
Non Expert SLT	0.65-0.80 (0.71)*
Non Expert ENT	0.26-0.80 (0.65)*

Inter Rater

Rater type	Weighted Kappa Range (Mean)
All	0.10-0.85 (0.57)
All SLT	0.14-0.83 (0.59)
All ENT	0.10-0.85 (0.58)
Expert SLT	0.40-0.76 (0.61)*
Expert ENT	0.43-0.78 (0.60)*
Non Expert SLT	0.14-0.81 (0.55)
Non Expert ENT	0.26-0.77 (0.52)

Parameter 9 : Stoma Noise

Intra Rater

Rater type	Weighted Kappa Range (Mean)
All	0.27-0.84 (0.64)*
All SLT	0.27-0.83 (0.66)*
All ENT	0.27-0.84 (0.61)*
Expert SLT	0.46-0.83 (0.70)*
Expert ENT	0.53-0.76 (0.66)*
Non Expert SLT	0.27-0.78 (0.64)*
Non Expert ENT	0.27-0.84 (0.57)

Inter Rater

Rater type	Weighted Kappa Range (Mean)
All	0.15-1.0 (0.51)
All SLT	0.15-0.80 (0.55)
All ENT	0.25-0.65 (0.47)
Expert SLT	0.30-0.80 (0.56)
Expert ENT	0.30-0.63 (0.43)
Non Expert SLT	0.34-0.69 (0.55)
Non Expert ENT	0.25-0.64 (0.49)

Parameter 10: Impairment of Fluency

Inter Rater

Rater type	Weighted Kappa Range (Mean)
All	0.52-0.79 (0.68)*
All SLT	0.57-0.79 (0.70)*
All ENT	0.52-0.73 (0.65)*
Expert SLT	0.57-0.78 (0.71)*
Expert ENT	0.52-0.70 (0.64)*
Non Expert SLT	0.57-0.79 (0.70)*
Non Expert ENT	0.52-0.73 (0.65)*

Inter Rater

Rater type	Weighted Kappa Range (Mean)
All	0.22-1.0 (0.59)
All SLT	0.31-1.0 (0.61)*
All ENT	0.38-0.73 (0.58)
Expert SLT	0.56-0.80 (0.68)*
Expert ENT	0.46-0.73 (0.58)
Non Expert SLT	0.31-0.83 (0.62)*
Non Expert ENT	0.52-0.67 (0.60)*

Parameter 11: Impairment of Articulatory Precision

Intra Rater

Rater type	Weighted Kappa Range (Mean)
All	0.00-0.74 (0.47)
All SLT	0.00-0.74 (0.45)
All ENT	0.27-0.69 (0.50)
Expert SLT	0.32-0.74 (0.51)
Expert ENT	0.27-0.39 (0.33)
Non Expert SLT	0.00-0.71 (0.40)
Non Expert ENT	0.45-0.69 (0.63)*

Inter Rater

Rater type	Weighted Kappa Range (Mean)
All	0.00-0.76 (0.43)
All SLT	0.06-0.67 (0.33)
All ENT	0.00-0.76 (0.37)
Expert SLT	0.23-0.67 (0.46)
Expert ENT	0.15-0.60 (0.50)
Non Expert SLT	0.07-0.61 (0.27)
Non Expert ENT	0.17-0.63 (0.47)

Parameter 12 : Positive Paralinguistic Features

Intra Rater

Rater type	Weighted Kappa Range (Mean)
All	0.20-0.91 (0.64)*
All SLT	0.60-0.83 (0.68)*
All ENT	0.20-0.91 (0.58)
Expert SLT	0.61-0.83 (0.69)*
Expert ENT	0.20-0.66 (0.49)
Non Expert SLT	0.60-0.78 (0.68)*
Non Expert ENT	0.54-0.91 (0.66)*

Inter Rater

Rater type	Weighted Kappa Range (Mean)
All	0.00-0.76 (0.43)
All SLT	0.31-0.76 (0.53)
All ENT	0.00-0.73 (0.37)
Expert SLT	0.31-0.75 (0.53)
Expert ENT	0.00-0.53 (0.27)
Non Expert SLT	0.37-0.72 (0.53)
Non Expert ENT	0.34-0.61 (0.45)

Parameter 13: Accent

Intra Rater N.B. SLT raters only

Rater type	Weighted Kappa Range (Mean)
All SLT	0.19-0.64 (0.48)
Expert SLT	0.34-0.62 (0.54)
Non Expert SLT	0.19-0.64 (0.43)

Inter Rater

Rater type	Weighted Kappa Range (Mean)
All SLT	0.00-0.69 (0.28)
Expert SLT	0.13-0.56 (0.35)
Non Expert SLT	0.00-0.69 (0.26)

Parameter 14: Poor Reader

Intra Rater N.B. SLT raters only

Rater type	Weighted Kappa Range (Mean)
All SLT	0.00-1.0 (0.45)
Expert SLT	0.00-0.65 (0.36)
Non Expert SLT	0.00-1.0 (0.37)

Inter Rater

Rater type	Weighted Kappa Range (Mean)
All SLT	0.00-0.67 (0.22)
Expert SLT	0.00-0.63 (0.23)
Non Expert SLT	0.00-0.67 (0.19)

Appendix J. Comprehensive results tables for naïve raters

* "good" or above level of agreement (Landis and Koch 1977)

Intra Rater

Parameter	Weighted Kappa Range (Mean)
Overall Grade	0.58-0.84 (0.72) *
Social Acceptability	0.48-0.78 (0.69) *
Intelligibility	0.13-0.73 (0.52)
Volume	0.00-0.37 (0.09)
Whisper	0.43-0.71 (0.53)
Gurgliness (wet)	0.18-0.65 (0.36)
Strain	0.00-0.45 (0.20)

Inter Rater

Parameter	Weighted Kappa Range (Mean)
Overall Grade	0.30-0.89 (0.62)*
Social Acceptability	0.38-0.84 (0.63)*
Intelligibility	0.15-0.90 (0.59)
Volume	0.00-0.74 (0.32)
Whisper	0.12-0.70 (0.41)
Gurgliness (wet)	0.09-0.59 (0.31)
Strain	0.00-0.59 (0.24)

Appendix K. The Patient and Carer rating scale

The carer version differs only in terms of pronouns included as marked by asterisk.

1. **Overall I would say my/ *my relative/friend's voice is:**

Excellent Good Adequate Poor

2. **The social acceptability of my/ *my relative/friend's voice is: (this means how you think other people feel about your voice especially people you meet who are not close members of your family):**

Excellent Good Adequate Poor

3. **The volume of my/ *my relative/friend's voice compared to my/ *his/her needs is:**

Excellent Good Adequate Poor

4. **Would you say your/ *your relative/friend's voice is:**

Whispery No Yes

Gurgly No Yes

Strained No Yes

Not loud enough No Yes

5. **How would you rate the intelligibility of your/ *your relative/friend's voice?**

Excellent Good Adequate Poor

Appendix L. Comprehensive inter rater reliability results tables to compare Patient, Carer, Naïve Expert SLT and Expert ENT judges.

* "good" or above level of agreement (Landis and Koch 1977)

Patient versus Carer raters

Patient versus Carer	Weighted Kappa 1 score only no mean score
Overall Grade	0.57
Social Acceptability	0.58
Intelligibility	0.74*
Volume	0.63*
Whisper	0.29
Gurgliness (wet)	0.62*
Strain	0.40

Patient versus Naive raters

Patient versus Naive	Weighted Kappa Range (Mean)
Overall Grade	0.22-0.43 (0.33)
Social Acceptability	0.12-0.37 (0.23)
Intelligibility	0.08-0.36 (0.19)
Volume	0.05-0.23 (0.15)
Whisper	0.09-0.45 (0.28)
Gurgliness (wet)	0.00-0.29 (0.18)
Strain	0.00-0.38 (0.14)

Patient versus Expert SLT

	Weighted Kappa Range (Mean)
Overall Grade	0.23-0.39 (0.34)
Social Acceptability	0.22-0.46 (0.33)
Intelligibility	0.25-0.47 (0.39)
Volume	0.26-0.41 (0.30)
Whisper	0.12-0.30 (0.21)
Gurgliness (wet)	0.03-0.25 (0.14)
Strain	0.07-0.31 (0.21)

Patient versus Expert ENT

	Weighted Kappa Range (Mean)
Overall Grade	0.19-0.35 (0.30)
Social Acceptability	0.20-0.41 (0.28)
Intelligibility	0.21-0.53 (0.43)
Volume	0.20-0.41 (0.33)
Whisper	0.00-0.30 (0.14)
Gurgliness (wet)	0.00-0.25 (0.13)
Strain	0.00-0.25 (0.07)

Carer versus Naïve

	Weighted Kappa Range (Mean)
Overall Grade	0.23-0.58 (0.41)
Social Acceptability	0.20-0.56 (0.38)
Intelligibility	0.02-0.33 (0.17)
Volume	0.00-0.15 (0.03)
Whisper	0.00-0.41 (0.15)
Gurgliness (wet)	0.00-0.34 (0.20)
Strain	0.00-0.27 (0.11)

Carer versus Expert SLT

	Weighted Kappa Range (Mean)
Overall Grade	0.37-0.61 (0.48)
Social Acceptability	0.38-0.61 (0.51)
Intelligibility	0.33-0.47 (0.40)
Volume	0.34-0.61 (0.45)
Whisper	0.22-0.46 (0.33)
Gurgliness (wet)	0.00-0.17 (0.08)
Strain	0.06-0.26 (0.17)

Carer versus Expert ENT

	Weighted Kappa Range (Mean)
Overall Grade	0.25-0.45 (0.36)
Social Acceptability	0.26-0.51 (0.44)
Intelligibility	0.24-0.62 (0.46)
Volume	0.41-0.52 (0.49)
Whisper	0.07-0.27 (0.16)
Gurgliness (wet)	0.00-0.35 (0.16)
Strain	0.00-0.09 (0.05)

Naïve versus Expert SLT

	Weighted Kappa Range (Mean)
Overall Grade	0.39-0.88 (0.67) *
Social Acceptability	0.26-0.85 (0.60) *
Intelligibility	0.07-0.72 (0.36)
Volume	0.00-0.25 (0.20)
Whisper	0.00-0.63 (0.32)
Gurgliness (wet)	0.00-0.55 (0.28)
Strain	0.00-0.56 (0.22)

Expert ENT versus Naïve

	Weighted Kappa Range (Mean)
Overall Grade	0.31-0.83 (0.63) *
Social Acceptability	0.22-0.82 (0.53)
Intelligibility	0.00-0.64 (0.19)
Volume	0.00-0.27 (0.07)
Whisper	0.05-0.45 (0.22)
Gurgliness (wet)	0.00-0.60 (0.25)
Strain	0.00-0.53 (0.16)

References

ABE, H., YONEKAWA, H., OHTA, F. & IMAIZUMI, S. 1986. Reproducibility of hoarse voice psychoacoustic evaluation. *Japanese Journal of Logopaedics and Phoniatrics,* 27, 168-177.

ACKERSTAFF, A. H., HILGERS, F. J. M., AARONSON, N. K. & BALM, A. J. M. 1994. Communication, functional disorders and lifestyle changes after total laryngectomy. *Clinical Otolaryngology,* 19, 295-300.

AHMAD, I., KUMAR, B. N., RADFORD, K., O'CONNELL, J. & BATCH, A. J. G. 2000. Surgical voice restoration following ablative surgery for laryngeal and hypopharyngeal carcinoma. *Journal of Laryngology and Otology,* 114, 522-525.

ALAM, D. S., VIVEK, P. P. & KMIECIK, J. 2008. Comparison of voice outcomes after radial forearm free flap reconstruction versus primary closure after laryngectomy. *Otolaryngology-Head and Neck Surgery,* 129, 240-244.

ALBIRMAWY, O. A., EL-GUINDY, A. S., ELSHEIKH, M. N., SAAFAN, M. E. & DARWISH, M. E. 2009. Effect of primary neopharyngeal repair on acoustic characteristics of tracheoesophageal voice after total laryngectomy. *Journal of Laryngology and Otology,* 123, 426-433.

ALLAN, W., BURGESS, L., HURREN, A., MARSH, R., SAMUEL, P. R. & SMALL, P. K. 2009. Oesophageal function in tracheoesophageal fistula speakers after laryngectomy. *Journal of Laryngology and Otology,* 123, 666-672.

ANTHONY, J. P., SINGER, M. I., DESCHLER, D. G., DOUGHERTY, E. T., REED, C. G. & KAPLAN, M. J. 1994. Long-Term Functional Results after Pharyngoesophageal Reconstruction with the Radial Forearm Free-Flap. *American Journal of Surgery,* 168, 441-445.

AWAN, S. N. & LAWSON, T. L. 2009. The Effect of Anchor Modality on the Reliability of Vocal Severity Ratings. *Journal of Voice,* 23, 341-352.

BARTKO, J. J. 1991. Measurement and reliability: statistical thinking considerations. *Schizophrenia Bulletin,* 17, 483-489.

BASSICH, C. J. & LUDLOW, C. L. 1986. The use of perceptual methods by new clinicians for assessing voice quality. *Journal of Speech and Hearing Disorders,* 51, 125-133.

BELE, I. V. 2005. Reliability in perceptual analysis of voice quality. *Journal of Voice,* 19, 555-573.

BENTZEN, N., GULD, A. & RASMUSSEN, H. 1976. X-Ray Videotape Studies of Laryngectomized Patients. *Journal of Laryngology and Otology,* 90, 655-666.

BLOM, E. C., PAULOSKI, B. R. & HAMAKER, R. C. 1995. Functional outcome after surgery for prevention of pharyngospasms in tracheoesophageal speakers. Part I: Speech Characteristics. *Laryngoscope,* 105**,** 1093-1103.

BLOM, E. D. 2000. Current status of voice restoration following total laryngectomy. *Oncology-New York,* 14**,** 915-922.

BLOM, E. D. 2003. Some comments on the escalation of tracheoesophageal voice prosthesis dimensions. *Archives of Otolaryngology-Head & Neck Surgery,* 129**,** 500-502.

BLOM, E. D., SINGER, M. I. & HAMAKER, R. C. 1986. A prospective study of tracheoesophageal speech. *Archives of Otolaryngology-Head & Neck Surgery,* 112**,** 440-7.

BLOM, J. G. & KOOPMANS-VAN BEINUM, F. J. An investigation concerning the judgment criteria for the pronunciation of Dutch I. Proceedings of the Institute of Phonetic Sciences, Amsterdam, 1973. 1-24.

BLOM, J. G. & VAN HERPT, L. W. A. The evaluation of jury judgments on pronunciation quality. Proceedings of the Institute of Phonetic Sciences, Amsterdam., 1976. 31-47.

BOON-KAMMA, B. 2001. *Verstaanbaarheid na totale laryngectomie. Report 136.* Amsterdam.

BRIDGES, A. 1991a. Acceptability ratings and intelligibility scores of alraryngeal speakers by three listener groups. *British Journal of Disorders of Communication,* 26**,** 325-335.

BRIDGES, A. 1991b. Perception of pitch contours in single words in alaryngeal speech. *British Journal of Disorders of Communication,* 26**,** 317-324.

BROK, H. A. J., STROEVE, R. J., COPPER, M. P. & SCHOUWENBURG, P. F. 1998. The treatment of hypertonicity of the pharyngo-oesophageal segment after laryngectomy. *Clinical Otolaryngology,* 23**,** 302-307.

BROWN, D. H., HILGERS, F. J. M., IRISH, J. C. & BALM, A. J. M. 2003. Postlaryngectomy voice rehabilitation: State of the art at the millennium. *World Journal of Surgery,* 27**,** 824-831.

CAMILLERI, A. E. & MACKENZIE, K. 1992. The acceptability of secondary tracheoesophageal fistula creation in long standing laryngectomees. *Journal of Laryngology and Otology,* 106**,** 231-233.

CANTU, E., RYAN, W. J., TANSEY, S. & JOHNSON, C. S. 1998. Tracheoesophageal puncture speech: Predictors of success and social validity ratings. *American Journal of Otolaryngology,* 19**,** 12-17.

CARDING, P. N., CARLSON, E., EPSTEIN, R., MATHIESON, L. & SHEWELL, C. 2000. Formal perceptual evaluation of voice quality in the United Kingdom. *Log Phon Vocol,* 25**,** 133-138.

CARDING, P. N., WILSON, J. A., MACKENZIE, K. & DEARY, I. J. 2009. Measuring voice outcomes: state of the science review. *Journal of Laryngology and Otology,* 123, 823-829.

CHAN, K. M. K., LI, M., LAW, T. Y. & YIU, E. M. L. 2012. Effects of immediate feedback on learning auditory perceptual voice quality evaluation. *International Journal of Speech-Language Pathology,* 14, 363-369.

CHAN, K. M. K. & YIU, E. M. L. 2002. The effects of anchors and training on the reliability of perceptual voice evaluation. *Journal of Speech, Language and Hearing Research,* 45, 111-126.

CHAN, K. M. K. & YIU, E. M. L. 2006. Comparison of two perceptual voice evaluation training programs for naive listeners. *Journal of Voice,* 20, 229-241.

CHEESMAN, A. D., KNIGHT, J., MCIVOR, J. & PERRY, A. 1986. Tracheo-oesophageal 'puncture speech'. An assessment technique for failed oesophageal speakers. *Journal of Laryngology & Otology,* 100, 191-9.

CHODOSH, P. L., GIANCARLO, H. R. & GOLDSTEIN, J. 1984. Pharyngeal myotomy for vocal rehabilitation postlaryngectomy. *Laryngoscope,* 94, 52-57.

CHONE, C. T., GRIPP, F. M., SPINA, A. L. & CRESPO, A. N. 2005. Primary versus secondary tracheoesophageal puncture for speech rehabilitation in total laryngectomy: Long-term results with indwelling voice prosthesis. *Otolaryngology-Head and Neck Surgery,* 133, 89-93.

CHONE, C. T., SEIXAS, V. O., PAES, L. A., GRIPP, F. M., TEIXEIRA, C., ANDREOLLO, N. A., SPINA, A. L., QUAGLIATO, E., BARCELOS, I. K. H. & CRESPO, A. N. 2008. Use of computerized manometry for the detection of pharyngoesophageal spasm in tracheoesophageal speech. *Otolaryngology-Head and Neck Surgery,* 139, 449-452.

CLARK, J. G. 1985. Alaryngeal speech intelligibility and the older listener. *Journal of Speech and Hearing Disorders,* 50, 60-65.

CLARKE, P. 2005. Macmillan SVR project handouts and personal communication.

CLEVENS, R. A., ESCLAMADO, R. M., HARTSHORN, D. O. & LEWIN, J. S. 1993. Voice rehabilitation after total laryngectomy and tracheoesophageal puncture using nonmuscle closure. *Annals of Otology Rhinology and Laryngology,* 102, 792-796.

COHEN, J. 1968. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin,* 70, 213-220.

COX, R. M. & MCDANIEL, D. M. 1984. Intelligibility ratings of continuous discourse: Application to hearing aid selection. *Journal of the Acoustical Society of America,* 76, 758-766.

CRARY, M. A. & GLOWASKY, A. L. 1996. Using botulinum toxin A to improve speech and swallowing function following total laryngectomy. *Archives of Otolaryngology-Head & Neck Surgery,* 122, 760-763.

CRONBACH, L. J. 1990. *Essentials of Psychological Testing,* New York, Harper and Row.

CULLINAN, W. L., BROWN, C. S. & BLALOCK, P. D. 1986. Ratings of intelligibility of esophageal and tracheoesophageal speech. *Journal of Communication Disorders,* 19, 185-195.

CULLINAN, W. L., E.M., P. & WILLIAMS, D. E. 1963. Comparison of procedures for scaling severity of stuttering *Jounal of Speech and Hearing Research,* 6, 187-194.

CULLINAN, W. L., PRATHER, E. M. & WILLIAMS, D. E. 1963 Comparison of procedures for scaling severity of stuttering. *Journal of Speech and Hearing Research,* 6, 187-194.

CUMBERWORTH, V. L., O'FLYNN, P., PERRY, A., BLEACH, N. R. & CHEESMAN, A. D. 1992. Surgical voice restoration after laryngopharyngectomy with free radial forearm flap repair using a Blom-Singer prosthesis. *Journal of the Royal Society of Medicine,* 85, 760-1.

DAMROSE, J. F., GOLDMAN, S. N., GROESSL, E. J. & ORLOFF, L. A. 2004. The impact of long term botulinum toxin injections on symptom severity in patients with spasmodic dysphonia. *Journal of Voice,* 18, 415-422.

DAMSTE, P. H. & LERMAN, J. W. 1969. Configuration of the neoglottis: An x-ray study. *folia Phoniatrica* 21, 347-38.

DAMSTE, P. H. & LERMAN, J. W. 1969. Configuration of the neoglottis: An x-ray study. *Folia Phoniatrica,* 21, 347-38.

DAVIS, P. J. 1981. Rehabilitation after total laryngectomy. *Medical Journal of Australia,* 1, 396-400.

DAY, A. M. B. & DOYLE, P. C. 2010. Assessing self-reported measures of voice disability in tracheoesophageal speakers. *Journal of Otolaryngology-Head & Neck Surgery,* 39, 762-768.

DE BODT, M., VAN DE HEYNING, P. H., WUYTS, F. L. & LAMBRECHTS, L. 1996. The perceptual evaluation of voice disorders. A*cta Oto-rhino-laryngologica Belg.,* 50, 283-291.

DE BODT, M. S., WUYTS, F. L., VAN DE HEYNING, P. H. & CROUX, C. 1997. Test-retest study of the GRBAS scale: Influence of experience and professional background on perceptual rating of voice quality. *Journal of Voice,* 11, 74-80.

DE KROM, G. 1994. Consistency and reliability of voice quality ratings for different types of speech fragments. *Journal of Speech and Hearing Research,* 37, 985-1000.

DE MADDALENA, H. & ZENNER, H. P. Evaluation of speech intelligibility after prosthetic voice restoration by a standardized telephone test. *In:*

ALGABA, J., ed. 6th International Congress on Surgical and Prosthetic Voice Restoration After Total Laryngectomy/2nd EGFL Conference, Sep 29-Oct 01 1995 San Sebastian, Spain. Elsevier Science Publ B V, 183-187.

DE RAUCOURT, D., RAME, J. P., DALIPHARD, F., LE PENNEE, D., BEQUIGNON, A. & LUQUET, A. 1998. Rehabilitation vocale par implant phonatoire etude de 62 patients ayant 5 ans de recul. *Rev Laryngol Otol Rhinol,* 119, 297-300.

DEJONCKERE, P. H., OBBENS, C., DE MOOR, G. M. & WIENECKE, G. H. 1993. Perceptual evaluation of dysphonia: Reliability and Relevance. *Folia Phoniatrica,* 45, 76-83.

DEJONCKERE, P. H., REMACLE, M., FRESNEL-ELBAZ, E., WOISARD, V., CREVIER, L. & MILLET, B. 1998. Reliability and clinical relevance of perceptual evaluation of pathological voices. *Rev Laryngol Otol Rhinol,* 119, 247-248.

DELSUPEHE, K., ZINK, I., LEJAEGERE, M. & DELAERE, P. 1998. Prospective randomized comparative study of tracheoesophageal voice prosthesis: Blom-Singer versus Provox. *Laryngoscope,* 108, 1561-1565.

DESCHLER, D. G., DOHERTY, E. T., ANTHONY, J. P., REED, C. G. & SINGER, M. I. 1994. Tracheoesophageal voice following tubed free radial forearm flap reconstruction of the neopharynx. *Annals of Otology Rhinology and Laryngology,* 103, 929-936.

DESCHLER, D. G., DOHERTY, E. T., REED, C. G., HAYDEN, R. E. & SINGER, M. I. 2000. Prevention of pharyngoesophageal spasm after laryngectomy with a half-muscle closure technique. *Annals of Otology Rhinology and Laryngology,* 109**,** 514-518.

DESCHLER, D. G., DOHERTY, E. T., REED, C. G. & SINGER, M. I. 1998. Quantitative and qualitative analysis of tracheoesophageal voice after pectoralis major flop reconstruction of the neopharynx. *Otolaryngology-Head and Neck Surgery,* 118**,** 771-776.

DESCHLER, D. G., DOHERTY, E. T., REED, C. G. & SINGER, M. I. 1999. Effects of sound pressure levels on fundamental frequency in tracheoesophageal speakers. *Otolaryngology-Head and Neck Surgery,* 121**,** 23-26.

DESCHLER, D. G. & GRAY, S. T. 2004. Tracheoesophageal speech following laryngopharyngectomy and pharyngeal reconstruction. *Otolaryngologic Clinics of North America,* 37**,** 567-+.

DIEDRICH, W. M. & YOUNGSTROM, K. A. 1966. *Alaryngeal Speech,* Springfield, Illinois, CC Thomas.

DONEGAN, J. O., GLUCKMAN, J. L. & SINGH, J. 1981. Limitations of the Blom Singer technique for voice restoration. *Annals of Otology Rhinology and Laryngology,* 90**,** 495-497.

DOYLE, P. C. 1997. Voice refinement following conservation surgery for cancer of the larynx: A conceptual framework for treatment intervention. *American Journal of Speech-Language Pathology,* 6**,** 27-35.

DOYLE, P. C., DANHAUER, J. L. & REED, C. G. 1988. Listeners perceptions of consonants produced by esophageal and tracheoesophageal talkers. *Journal of Speech and Hearing Disorders,* 53**,** 400-407.

DOYLE, P. C. & EADIE, T. L. 2005. The perceptual nature of alaryngeal voice and speech. *In:* DOYLE, P. C. & KEITH, R. L. (eds.) *Contemporary considerations in the treatment and rehabilitation of Head and Neck Cancer: Voice, Speech and Swallowing.*: Pro-Ed Inc, USA.

DOYLE, P. C. & EADIE, T. L. 2005. Pharygoesophageal Segment Function: A Review and Reconsideration. *In:* DOYLE, P. C. & KEITH, R. L. (eds.) *Contemporary considerations in the treatment and rehabilitation of Head and Neck Cancer: Voice, Speech and Swallowing.*: Pro-Ed Inc, USA.

DOYLE, P. C., SWIFT, R. & HAAF, R. G. 1989. Effects of listener sophistication on judgments of tracheoesophageal talker intelligibility. *Journal of Communication Disorders,* 22**,** 105-113.

DWORKIN, J. P., MELECA, R. J., OH, C. & SIMPSON, M. L. 2002. Use of esophageal videoendoscopy for the differential diagnosis and treatment of poor tracheoesophageal speech. *Journal of Medical Speech-Language Pathology,* 10**,** 133-141.

DWORKIN, J. P., MELECA, R. J., SIMPSON, M. L., ZORMEIER, M., GARFIELD, I., JACOBS, J. & MATHOG, R. H. 1999. Vibratory characteristics of the pharyngoesophageal segment in total laryngectomees. *Journal of Medical Speech-Language Pathology,* 7**,** 1-18.

EADIE, T., NICOLICI, C., BAYLOR, C., ALMAND, K., WAUGH, P. & MARONIAN, N. 2007. Effect of experience on judgments of adductor spasmodic dysphonia. *Ann Otol Rhinol Laryngol,* 116**,** 695-701.

EADIE, T., SROKA, A., WRIGHT, D. R. & MERATI, A. 2011. Does knowledge of medical diagnosis bias auditory-perceptual judgments of dysphonia? *Journal of Voice,* 25**,** 420-429.

EADIE, T. L. & BAYLOR, C. R. The effect of perceptual training on inexperienced listeners' judgments of dysphonic voice. 34th Annual Symposium on Care of the Professional Voice, Jun 2005 Philadelphia, PA. Mosby-Elsevier, 527-544.

EADIE, T. L. & DOYLE, P. C. 2002a. Direct magnitude estimation and interval scaling of naturalness and severity in tracheoeosphageal (TE) speakers. *Jspeech Lang Hear Res* 45**,** 1088-1096.

EADIE, T. L. & DOYLE, P. C. 2002b. Direct magnitude estimation and interval scaling of pleasantness and severity in dysphonic and normal speakers. *Journal of the Acoustical Society of America,* 112**,** 3014-3021.

EADIE, T. L. & DOYLE, P. C. 2004. Auditory-perceptual scaling and quality of life in tracheoesophageal speakers. *Laryngoscope,* 114**,** 753-759.

EADIE, T. L. & DOYLE, P. C. 2005a. Scaling of voice pleasantness and acceptability in tracheoesophageal speakers. *Journal of Voice,* 19**,** 373-383.

EADIE, T. L. & DOYLE, P. C. 2005c. Quality of life in male tracheoesophageal (TE) speakers. *Journal of Rehabilitation Research and Development,* 42**,** 115-124.

EADIE, T. L. & DOYLE, T. C. 2005b. Classification of dysphonic voice: Acoustic and auditory-perceptual measures. *Journal of Voice,* 19**,** 1-14.

EADIE, T. L., DOYLE, T. C., HANSEN, K. & BEAUDIN, T. G. 2008. Influence of speaker gender on listener judgments of tracheoesophageal speech. *Journal of Voice,* 22**,** 43-57.

EADIE, T. L., KAPSNER, M., ROSENZWEIG, J., WAUGH, P., HILLEL, A. & MERATI, A. 2010a. The role of experience on judgments of dysphonia. *Journal of Voice,* 24**,** 564-573.

EADIE, T. L. & KAPSNER-SMITH, M. 2011. The effect of listener experience and anchors on judgments of dysphonia. *Journal of Speech Language and Hearing Research,* 54**,** 430-447.

EADIE, T. L., VAN BOVEN, L., STUBBS, K. & GIANNINI, E. 2010b. The effect of musical background on judgments of dysphonia. *Journal of Voice,* 24**,** 93-101.

EDELS, Y. 1983. Pseudo-voice: Its Theory and Practice. *In:* EDELS, Y. (ed.) *Laryngectomy Diagnosis to Rehabilitation.*London: Croom Helm.

ERGUN, G. A., KAHRILAS, P. & LOGEMANN, J. A. 1993. Interpretation of pharyngeal manometric recordings; limitations and variability. *Diseases of the esophagus,* 6**,** 11-16.

EVANS, E., CARDING, P. & DRINNAN, M. 2009. The Vocal Handicap Index with post-laryngectomy male voices. *International Journal of Language & Communication Disorders,* 44**,** 575-586.

FAGEL, W. P. F., VAN HERPT, L. W. A. & BOVES, L. 1983. Analysis of the perceptual qualities of Dutch speakers' voice and pronunciation. *Speech Communication,* 2**,** 315-326.

FAIRBANKS, G. 1960. *Voice and Articulation Drillbook,* New York, Harper Row.

FERRER RAMIREZ, M. J., GUALLART DOMENECH, F. G., BROTONS DURBAN, S., CARRASCO LLATAS, M., ESTELLES FERRIOL, E. & LOPEZ MARTINEZ, R. 2001. Surgical Voice Restoration after total laryngectomy: Long term results. *European Archives of Oto-Rhino-Laryngology,* 258**,** 463-466.

FEX, S. 1992. Perceptual Evaluation. *Journal of Voice,* 6, 155-158.

FINIZIA, C., DOTEVALL, H., LUNDSTROM, E. & LINDSTROM, J. 1999. 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,* 125, 157-163.

FINIZIA, C., LINDSTROM, J. & DOTEVALL, H. 1998. Intelligibility and perceptual ratings after treatment for laryngeal cancer: Laryngectomy versus radiotherapy. *Laryngoscope,* 108, 138-143.

FUJIMOTO, P. A., MADISON, C. L. & LARRIGAN, L. B. 1991. The effects of a tracheostoma valve on the intelligibility and quality of tracheoesophageal speech. *Journal of Speech and Hearing Research,* 34, 33-36.

GARTH, R. J. N., MCRAE, A. & EVANS, P. H. R. 1991. Tracheoesophageal puncture - a review of problems and complications. *Journal of Laryngology and Otology,* 105, 750-754.

GELFER, M. P. 1988. Perceptual analysis of voice: development and use of rating scales. *Journal of Voice,* 2, 320-326.

GERRATT, B. R. & KREIMAN, J. 2000. Theoretical and methodological development in the study of pathological voice quality. *Journal of Phonetics,* 28, 335-342.

GERRATT, B. R., KREIMAN, J., ANTONANZASBARROSO, N. & BERKE, G. S. 1993. Comparing internal and external standards in voice quality judgments. *Journal of Speech and Hearing Research,* 36, 14-20.

GOULD, J., WAUGH, J., CARDING, P. & DRINNAN, M. 2012. A new voice rating tool for clinical practice. *Journal of Voice,* 26, e163-e170.

GREEN, G. & HULTS, M. 1982. Preferences for 3 types of alaryngeal speech. *Journal of Speech and Hearing Disorders,* 47, 141-145.

GUILDFORD, J. P. 1954. *Psychometric Methods,* New York, McGraw-Hill.

HAMADE, R., HEWLETT, N. & SCANLON, E. 2006. A quantitative and qualitative evaluation of an automatic occlusion device for tracheoesophageal speech: The Provox FreeHands HME. *Clinical Linguistics & Phonetics,* 20, 187-193.

HAMAKER, R. C. & BLOM, E. D. 2003. Botulinum neurotoxin for pharyngeal constrictor muscle spasm in tracheoesophageal voice restoration. *Laryngoscope,* 113, 1479-1482.

HAMMARBERG, B., FRITZELL, B., GAUFFIN, J. & SUNDBERG, J. 1986. Acoustic and perceptual analysis of vocal dysfunction. *Journal of Phonetics,* 14, 533-547.

HAMMARBERG, B., FRITZELL, B., GAUFFIN, J., SUNDBERG, J. & WEDIN, L. 1980. Perceptual and acoustic correlates of abnormal voice qualities. *Acta Oto-Laryngologica,* 90**,** 441-451.

HEATON, J. M., SANDERSON, D., DUNSMORE, I. R. & PARKER, A. J. 1996. Speech assessment of patients using three types of indwelling tracheo-oesophageal voice prostheses. *Journal of Laryngology and Otology,* 110**,** 343-347.

HELOU, L. B., SOLOMON, N. P., HENRY, L. R., COPPIT, G. L., HOWARD, R. S. & STOJADINOVIC, A. 2010. The role of listener experience of Consensus Auditory- Perceptual Evaluation of Voice (CAPE-V) Ratings of post-thyroidectomy voice. *Journal of Voice,* 19**,** 248-258.

HILGERS, F. J. M. & BALM, A. J. M. 1993. Long-term results of vocal rehabilitation after total laryngectomy with the low-resistance, indwelling Provox(tm) voice prosthesis system. *Clinical Otolaryngology,* 18**,** 517-523.

HILLMAN, R. E., WALSH, M. J. & HEATON, J. T. 2005. Laryngectomy Speech Rehabilitation: A Review of Outcomes. *In:* DOYLE, P. C. & KEITH, R. L. (eds.) *Contemporary considerations in the treatment and rehabilitation of Head and Neck Cancer: Voice, Speech and Swallowing.*. Pro-Ed Inc, USA.

HIRANO, M. 1981. *Clinical Examination of Voice,* New York, Springer.

HIRANO, M. 1989. Objective evaluation of the human voice - Clinical Aspects. *Folia Phoniatrica,* 41, 89-144.

HO, M. W., HOUGHTON, L., GILLMARTIN, E., JACKSON, S. R., LANCASTER, J., JONES, T. M., BLACKBURN, T. K., HOMER, J. J., LOUGHRAN, S., ASCOTT, F. M. & SHAW, R. J. 2012. Outcomes following pharyngolaryngectomy reconstruction with the anterolateral thigh (ALT) free flap. *British Journal of Oral & Maxillofacial Surgery,* 50, 19-24.

HOFFMAN, H. T., FISCHER, H., VAN DEMARK, D., PETERSON, K. L., MCCULLOCH, T. M., KARNELL, L. H. & FUNK, G. F. 1997. Botulinum toxin after total laryngectomy. *Head and Neck,* 19, 92-97.

HOFFMAN, H. T. & MCCULLOCH, T. M. 1998. Botulinum Neurotoxin for Tracheoesophageal Voice Failure. *In:* BLOM, E. D., SINGER, M. I. & HAMAKER, R. C. (eds.) *Tracheoesophageal Voice Restoration Following Total Laryngectomy.* San Diego: Singular Publishing Group.

HOTZ, M. A., BAUMANN, A., SCHALLER, I. & ZBAREN, P. 2002. Success and predictability of Provox prosthesis voice rehabilitation. *Archives of Otolaryngology-Head & Neck Surgery,* 128, 687-691.

HUI, Y., WEI, W. I., YUEN, P. W., LAM, L. K. & HO, W. K. 1996. Primary closure of the pharyngeal remnant after total laryngectomy: How much residual mucosa is sufficient? *Laryngoscope,* 106, 490-495.

HURREN, A. 1997. Perceptual assessment of voice quality after total laryngectomy. *Head and Neck Oncology Conference.* Nottingham.

HURREN, A., HILDRETH, A. J. & CARDING, P. N. 2009. Can we perceptually rate alaryngeal voice? Developing the Sunderland Tracheoesophageal Voice Perceptual Scale. *Clinical Otolaryngology,* 34**,** 533-538.

ISMAN, K. A. & O'BRIEN, C. J. 1992. Videofluoroscopy of the pharyngoesophageal segment during tracheoesophageal and esophageal speech. *Head and Neck-Journal for the Sciences and Specialties of the Head and Neck,* 14**,** 352-358.

ISSING, W. J., FUCHSHUBER, S. & WEHNER, M. 2001. Incidence of tracheo-oesophageal fistulas after primary voice rehabilitation with the Provox or the Eska-Herrmann voice prosthesis. *European Archives of Oto-Rhino-Laryngology* 258**,** 240-242.

IWAI, H., TSUJI, H., TACHIKAWA, T., INOUE, T., IZUMIKAWA, M., YAMAMICHI, K. & YAMASHITA, T. 2002. Neoglottic formation from posterior pharyngeal wall conserved in surgery for hypopharyngeal cancer. *Auris Nasus Larynx,* 29**,** 153-157.

IWARSSON, J. & PETERSEN, N. R. 2012. Effects of consensus training on the reliability of auditory perceptual ratings of voice quality. *Journal of Voice,* 26**,** 304-312.

JONES, A. S., ROLAND, N. J., HUSBAND, D., HAMILTON, J. W. & GATI, I. 1996. Free revascularized jejunal loop repair following total pharyngolaryngectomy for carcinoma of the hypopharynx: Report of 90 patients. *British Journal of Surgery,* 83**,** 1279-1283.

JONGMANS, P., VAN AS, C. J., POLS, L. C. W. & HILGERS, F. J. M. An introduction to the assessment of intelligibility of tracheoesophageal speech. I.F.A., 2003 Institute of Phonetic Sciences, University of Amsterdam, Netherlands. 185-196.

JONGMANS, P., WEMPE, T. G., VAN TINTEREN, H., HILGERS, F. J. M., POLS, L. C. W. & VAN AS-BROOKS, C. J. 2010. Acoustic analysis of the voiced-voiceless distinction in Dutch tracheoesophageal speech. *Journal of Speech Language and Hearing Research,* 53**,** 284-297.

KAO, W. W., MOHR, R. M., KIMMEL, C. A., GETCH, C. & SILVERMAN, C. 1994. The Outcome and techniques of primary and secondary tracheoesophageal puncture. *Archives of Otolaryngology-Head & Neck Surgery,* 120**,** 301-307.

KARLEN, R. G. & MAISEL, R. H. 2001. Does primary tracheoesophageal puncture reduce complications after laryngectomy and improve patient communication? *American Journal of Otolaryngology,* 22**,** 324-328.

KAZI, R., KANAGALINGAM, J., VENKITARAMAN, R., PRASAD, V., CLARKE, P., NUTTING, C. M., RHYS-EVANS, P. & HARRINGTON, K. J. 2009. Electroglottographic and perceptual evaluation of tracheoesophageal speech. *Journal of Voice,* 23**,** 247-254.

KAZI, R., KIVERNITI, E., PRASAD, V., VENKITARAMAN, R., NUTTING, C. M., CLARKE, P., RHYS-EVANS, P. & HARRINGTON, K. J. 2006a. Multidimensional assessment of female tracheoesophageal prosthetic speech. *Clinical Otolaryngology,* 31**,** 511-517.

KAZI, R., SINGH, A., DE CORDOVA, J., AL-MUTAIRY, A., CLARKE, P., NUTTING, C., RHYS-EVANS, P. & HARRINGTON, K. 2006c. Validation of a voice prosthesis questionnaire to assess valved speech and its related issues in patients following total laryngectomy. *Clinical Otolaryngology,* 31**,** 404-410.

KAZI, R., SINGH, A., DE CORDOVA, J., CLARKE, P., HARRINGTON, K. J. & RHYS-EVANS, P. 2005. A new self-administered questionnaire to determine patient experience with voice prostheses (Blom-Singer valve). *Journal of Post-Graduate Medicine and Quality of Life,* 51**,** 253-259.

KAZI, R., SINGH, A., MULLAN, G. P. J., VENKITARAMAN, R., NUTTING, C. M., CLARKE, P., RHYS-EVANS, P. & HARRINGTON, K. J. 2006b. Can objective parameters derived from videofluoroscopic assessment of post-laryngectomy valved speech replace current subjective measures? An e-tool-based analysis. *Clinical Otolaryngology,* 31**,** 518-524.

KEARNS, K. P. & SIMMONS, N. N. 1988. Interobserver reliability and perceptual ratings - More than meets the ear. *Journal of Speech and Hearing Research,* 31**,** 131-136.

KEMPSTER, G. B., GERRATT, B. R., ABBOTT, K. V., BARKMEIER-KRAEMER, J. & HILLMAN, R. E. 2009. Consensus Auditory-Perceptual Evaluation of Voice: Development of a standardized clinical protocol. *American Journal of Speech-Language Pathology,* 18**,** 124-132.

KENDALL, T. S. 2009. *Speech Rate, Pause, and Linguistic Variation: An Examination Through the Sociolinguistic Archive and Analysis Project.* PhD, Duke University, USA.

KIRCHNER, J. A., SCATLIFF, J. H., DEY, F. L. & SHEDD, D. P. 1963. The pharynx after laryngectomy: Changes in its structure and function. *Laryngoscope,* 73**,** 18-33.

KOYBASIOGLU, A., OZ, O., USLU, S., ILERI, F., INAL, E. & UNAL, S. 2003. Comparison of pharyngoesophageal segment pressure in total laryngectomy patients with and without pharyngeal neurectomy. *Head and Neck-Journal for the Sciences and Specialties of the Head and Neck,* 25**,** 617-623.

KREIMAN, J. 1997. Listening to voices: Theory and practice in voice perception research. *In:* JOHNSON K. AND MULLINEX, J. W. (ed.) *Talker Variability in Speech Processing.* San Diego: Academic Press Inc.

KREIMAN, J. & GERRATT, B. R. 1996. The perceptual structure of pathologic voice quality. *Journal of the Acoustical Society of America,* 100**,** 1787-1795.

KREIMAN, J. & GERRATT, B. R. 1998. Validity of rating scale measures of voice quality. *Journal of the Acoustical Society of America,* 104**,** 1598-1608.

KREIMAN, J. & GERRATT, B. R. 2000. Sources of listener disagreement in voice quality assessment. *Journal of the Acoustical Society of America,* 108**,** 1867-1876.

KREIMAN, J. & GERRATT, B. R. 2011. Comparing two methods for reducing variability in voice quality measurements. *Journal of Speech Language and Hearing Research,* 54**,** 803-812.

KREIMAN, J., GERRATT, B. R. & ANTONANZAS-BARROSO, N. 2007. Measures of the glottal source spectrum. *Journal of Speech Language and Hearing Research,* 50**,** 595-610.

KREIMAN, J., GERRATT, B. R. & BERKE, G. S. 1994. The multidimensional nature of pathological vocal quality. *Journal of the Acoustical Society of America,* 96**,** 1291-1302.

KREIMAN, J., GERRATT, B. R. & ITO, M. 2007. When and why listeners disagree in voice quality assessment tasks. *Journal of the Acoustical Society of America,* 122**,** 2354-2364.

KREIMAN, J., GERRATT, B. R., KEMPSTER, G. B., ERMAN, A. & BERKE, G. S. 1993. Perceptual evaluation of voice quality - Review, tutorial, and a framework for future-research. *Journal of Speech and Hearing Research,* 36**,** 21-40.

KREIMAN, J., GERRATT, B. R. & PRECODA, K. 1990. Listener experience and perception of voice quality. *Journal of Speech and Hearing Research,* 33**,** 103-115.

KREIMAN, J., GERRATT, B. R., PRECODA, K. & BERKE, G. S. 1992. Individual-differences in voice quality perception. *Journal of Speech and Hearing Research,* 35**,** 512-520.

LANDIS, J. R. & KOCH, G. G. 1977. The measurement of observer agreement for categorical data. *Biometrics,* 33**,** 159-174.

LAVER, J. 1980. *The Phonetic Description of Voice Quality,* London, Cambridge University Press.

LEE, A., WHITEHILL, T. L. & CIOCCA, V. 2009. Effect of listener training on perceptual judgement of hypernasality. *Clinical Linguistics & Phonetics,* 23**,** 319-334.

LEE, M., DRINNAN, M. & CARDING, P. 2005. The reliability and validity of patient self-rating of their own voice quality. *Clinical Otolaryngology,* 30**,** 357-361.

LEWIN, J. S. 2001. Further experience with botox injection for tracheoesophageal speech failure. *Head and Neck***,** 456-466.

LIU, C., YU, J., N., W., CHEN, R., CHANG, H., LI, H., TSAI, C., YANG, Y. & LU, C. 1998. Emotional symptoms are secondary to the voice disorder in patients with spasmodic dysphonia. *General Hospital Psychiatry,* 20, 255-259.

LUNDSTROM, E., HAMMARBERG, B., MUNCK-WIKLAND, E. & EDSBORG, N. 2008. The pharyngoesophageal segment in laryngectomees-videoradiographic, acoustic and voice quality perceptual data. *Log Phon Vocol,* 33, 115-125.

MACKEY, L. S., FINN, P. & INGHAM, R. J. 1997. Effect of speech dialect on speech naturalness ratings: A systematic replication of Martin, Haroldson, and Triden (1984). *Journal of Speech Language and Hearing Research,* 40, 349-360.

MACLEAN, J., SZCZESNIAK, M., COTTON, S., COOK, I. & PERRY, A. 2011. Impact of a laryngectomy and surgical closure technique on swallow biomechanics and dysphagia severity. *Otolaryngology-Head and Neck Surgery,* 144, 21-28.

MAHIEU, H. F., ANNYAS, A. A., SCHUTTE, H. K. & VAN DER JAGT, E. J. 1987. Pharyngoesophageal myotomy for vocal rehabilitation of laryngectomees. *Laryngoscope,* 97, 451-457.

MARYN, Y., ROY, N., DE BODT, M., VAN CAUWENBERGE, P. & CORTHALS, P. 2009. Acoustic measurement of overall voice quality: A meta-analysis. *Journal of the Acoustical Society of America,* 126, 2619-2634.

MAX, L., DE BRUYN, W. & STEURS, W. 1997. Intelligibility of oesophageal and tracheo-oesophageal speech: preliminary observations. *European Journal of Disorders of Communication,* 32, 429-440.

MAX, L., STEURS, W. & DEBRUYN, W. 1996. Vocal capacities in esophageal and tracheoesophageal speakers. *Laryngoscope,* 106, 93-96.

MCAULIFFE, M. J., WARD, E. C., BASSETT, L. & PERKINS, K. 2000. Functional speech outcomes after laryngectomy and pharyngolaryngectomy. *Archives of Otolaryngology-Head & Neck Surgery,* 126, 705-709.

MCCOLL, D. A. 2006. Intelligibility of tracheoesophageal speech in noise. *Journal of Voice,* 20, 605-615.

MCDOWELL, I. & NEWELL, C. 1996. *Measuring Health A Guide to Rating Scales and Questionnaires,* New York, Oxford University Press Inc.

MCDOWELL, I. & NEWELL, C. 2006. *Measuring Health: A guide to rating scales and questionnaires,* New York, Oxford University Press.

MCHENRY, M. 2011. An Exploration of Listener Variability in Intelligibility Judgments. *American Journal of Speech-Language Pathology,* 20, 119-123.

MCIVOR, J., EVANS, P. H. R., PERRY, A. & CHEESMAN, A. D. 1990. Radiological assessment of post laryngectomy speech. *Clinical Radiology,* 41, 312-316.

MELECA, R. J., DWORKIN, J. P., ZORMEIER, M. M., SIMPSON, M. L., SHIBUYA, T. & MATHOG, R. H. 2000. Videostroboscopy of the pharyngoesophageal segment in laryngectomy patients treated with botulinum toxin. *Otolaryngology-Head and Neck Surgery,* 123, 38-43.

MILFORD, C. A., PERRY, A., MUGLISTON, T. A. & CHEESMAN, A. D. 1988. A British experience of surgical voice restoration as a primary procedure. *Archives of Otolaryngology, Head and Neck Surgery,* 114, 1419-1421.

MILLET, B. & DEJONCKERE, P. H. 1998. What determines the differences in perceptual rating of dysphonia between experienced raters? *Folia Phoniatrica Et Logopaedica,* 50, 305-310.

MIRALLES, J. L. & CERVERA, T. 1995. Voice intelligibility in patients who have undergone laryngectomies. *Journal of Speech and Hearing Research,* 38, 564-571.

MISONO, S., MERATI, A. L. & EADIE, T. L. 2012. Developing auditory-perceptual judgment reliability in Otolaryngology residents. *Journal of Voice,* 26, 358-364.

MOERMAN, M., MARTENS, J. P., CREVIER-BUCHMAN, L., DE HAAN, E., GRAND, S., TESSIER, C., WOISARD, V. & DEJONCKERE, P. 2006. The INFVo perceptual rating scale for substitution voicing:

development and reliability. *European Archives of Oto-Rhino-Laryngology,* 263, 435-439.

MOERMAN, M., MARTENS, J. P. & DEJONCKERE, P. 2004. Application of the Voice Handicap Index in 45 patients with substitution voicing after total laryngectomy. *European Archives of Oto-Rhino-Laryngology,* 261, 423-428.

MOHRI, M., YOSHIFUJI, M., KINISHI, M. & AMATSU, M. 1994. Neoglottic activity in tracheoesophageal phonation. *Auris Nasus Larynx,* 21, 53-58.

MOON, J. B. & WEINBERG, B. 1987. Aerodynamic and myoelastic contributions to tracheoesophageal voice production. *Journal of Speech and Hearing Research,* 30, 387-395.

MOST, T., TOBIN, Y. & MIMRAN, R. C. 2000. Acoustic and perceptual characteristics of esophageal and tracheoesophageal speech production. *Journal of Communication Disorders,* 33, 165-181.

MUNOZ, J., MENDOZA, E., FRESNEDA, M. D. & CARBALLO, G. 2002. Perceptual analysis in different voice samples: Agreement and reliability. *Perceptual and Motor Skills,* 94, 1187-1195.

MURRAY, D. J., NOVAK, C. B. & NELIGAN, P. C. 2008. Fasciocutaneous free flaps in pharyngolaryngo-oesophageal reconstruction: a critical review of the literature. *Journal of Plastic Reconstructive and Aesthetic Surgery,* 61, 1148-1156.

NAGLE, K. F. & EADIE, T. L. 2012. Listener effort for highly intelligible tracheoesophageal speech. *Journal of Communication Disorders,* 45, 235-245.

NIEBOER, G. L. J., DE GRAAF, T. & SCHUTTE, H. K. 1988. Esophageal voice quality judgements by means of the semantic differential. *Journal of Phonetics,* 16, 417-436.

NUNALLY, J. C. J. 1970. *Introduction to Psychological Measurement,* New York, McGraw-Hill.

NUNALLY, J. C. J. 1978. *Psychometric Theory,* New York, McGraw-Hill.

O'LEARY., I. 1988. A preliminary report on the use of the Groningen tracheoesophageal valve in laryngectomy patients in the UK. *Head and Neck Oncology*. Nottingham.

OATES, J. 2009. Auditory-perceptual evaluation of disordered voice quality pros, cons and future directions. *Folia Phoniatrica Et Logopaedica,* 61, 49-56.

O'LEARY, I. 1988. A preliminary report on the use of the Groningen tracheo-oesophageal valve in laryngectomy patients in Sheffield, U.K. National Head and Neck Oncology Conference, 1988 Nottingham, U.K.

O'LEARY, I. K., HEATON, J. M., CLEGG, R. T. & PARKER, A. J. 1994. Acceptability and intelligibility of tracheoesophageal speech using the Groningen valve. *Folia Phoniatrica Et Logopaedica,* 46, 180-187.

OLSON, N. R. & CALLAWAY, E. 1990. Nonclosure of pharyngeal muscle after laryngectomy. *Annals of Otology Rhinology and Laryngology,* 99, 507-508.

OLTHOFF, A., MRUGALLA, S., LASKAWI, R., FROHLICH, M., STUERMER, I., KRUSE, E., AMBROSCH, P. & STEINER, W. 2003. Assessment of irregular voices after total and laser surgical partial laryngectomy. *Archives of Otolaryngology-Head & Neck Surgery,* 129, 994-999.

OMORI, K. & KOJIMA, H. 1999. Neoglottic vibration in tracheoesophageal shunt phonation. *European Archives of Oto-Rhino-Laryngology,* 256, 501-505.

OMORI, K., KOJIMA, H., NONOMURA, M. & FUKUSHIMA, H. 1994. Mechanism of tracheoesophageal shunt phonation. *Archives of Otolaryngology-Head & Neck Surgery,* 120, 648-652.

OP DE COUL, B. M. R., ACKERSTAFF, A. H., VAN AS, C. J., VAN DEN HOOGEN, F. J. A., MEEUWIS, C. A., MANNI, J. J. & HILGERS, F. J. M. 2005. Quality of life assessment in laryngectomized individuals: Do we need additions to standard questionnaires in specific clinical research projects? *Clinical Otolaryngology,* 30, 169-175.

OP DE COUL, B. M. R., HILGERS, F. J. M., BALM, A. J. M., TAN, I. B., VAN DEN HOOGEN, F. J. A. & VAN TINTEREN, H. 2000. A decade of postlaryngectomy vocal rehabilitation in 318 patients - A single institution's experience with consistent application of Provox indwelling voice prostheses. *Archives of Otolaryngology-Head & Neck Surgery,* 126, 1320-1328.

OP DE COUL, B. M. R., VAN DEN HOOGEN, F. J. A., VAN AS, C. J., MARRES, H. A. M., JOOSTEN, F. B. M., MANNI, J. J. & HILGERS, F. J. M. 2003. Evaluation of the effects of primary myotomy in total laryngectomy on the neoglottis with the use of quantitative videofluoroscopy. *Archives of Otolaryngology-Head & Neck Surgery,* 129, 1000-1005.

PATEL, R. S., MAKITIE, A. A., GOLDSTEIN, D. P., GULLANE, P. J., BROWN, D., IRISH, J. & GILBERT, R. W. 2009. Morbidity and functional outcomes following gastro-omental free flap reconstruction of circumferential pharyngeal defects. *Head and Neck-Journal for the Sciences and Specialties of the Head and Neck,* 31, 655-663.

PERRY, A. 1989. *Vocal Rehabilitation after Total Laryngectomy.* Leicester.

PINDZOLA, R. H. & CAIN, B. H. 1988. Acceptability ratings of tracheoesophageal speech. *Laryngoscope,* 98, 394-397.

RAMACHANDRAN, K., ARUNACHALAM, P. S., HURREN, A., MARSH, R. L. & SAMUEL, P. R. 2003. Botulinum toxin injection to improve voice post-laryngectomy: The Sunderland experience *Journal of Laryngology and Otology,* 117, 544-548.

ROBB, G. & LEWIN, J. S. 2003. Speech and swallowing outcomes in reconstructions of the pharynx and cervical esophagus. *Head and Neck-Journal for the Sciences and Specialties of the Head and Neck,* 25, 232-244.

ROBBINS, J., FISHER, H. B., BLOM, E. C. & SINGER, M. I. 1984. A comparative acoustic study of normal, esophageal, and tracheoesophageal speech production. *Journal of Speech and Hearing Disorders,* 49, 202-210.

ROBILLARD SHULTZ, J. & HARRISON, J. 1992. Defining and predicting tracheoesophageal puncture success. *Archives of Otolaryngology-Head & Neck Surgery,* 118, 811-816.

ROGERSON, J. & DODD, B. 2005. Is there an effect of dysphonic teachers' voices on children's processing of spoken language? *Journal of Voice,* 19, 47-60.

SANDERSON, R. J., ANDERSON, S. J., DENHOLM, S. & KERR, A. I. G. 1993. The assessment of alaryngeal speech. *Clincal Otolaryngology,* 18, 181-183.

SAPIR, S., ARONSON.A.E. & THOMAS, J. E. 1986. Judgment of voice improvement after recurrent laryngeal nerve section for spastic dysphonia: Clinician versus patients. *Ann Otol Rhinol Laryngol* 95, 137-141.

SCHIAVETTI, N. & METZ, D. E. 1997. Stuttering and the measurement of speech naturalness. *In:* CURLEE, R. F. & SIEGEL, G. M. (eds.)

Nature and treatment of stuttering: New directions. second ed. Boston: Allyn and Bacon.

SCHUSTER, M., HADERLEIN, T., NOTH, E., LOHSCHELLER, J., EYSHOLDT, U. & ROSANOWSKI, F. 2006. Intelligibility of laryngectomees' substitute speech: automatic speech recognition and subjective rating. *European Archives of Oto-Rhino-Laryngology,* 263, 188-193.

SCHUSTER, M., TOY, H., LOHSCHELLER, J., EYSHOLDT, U. & ROSANOWSKI, F. 2005. Quality of life and voice handicap of laryngectomees using tracheoesophageal substitute voice. *Laryngo-Rhino-Otologie,* 84, 101-107.

SEDORY, S. E., HAMLET, S. L. & CONNOR, N. P. 1989. Comparisons of perceptual and acoustic characteristics of tracheoesophageal and excellent esophageal speech. *Journal of Speech and Hearing Disorders,* 54, 209-214.

SHIPP, T. 1967. Frequency, duration and perceptual measures in relation to judgments of alaryngeal speech acceptability. *Journal of Speech and Hearing Research,* 10, 417-427.

SHIPP, T. 1970. EMG of pharyngoesophageal musculature during alayrngeal voice production. *Journal of Speech and Hearing Research,* 13, 184-192.

SHRIVASTAV, R. 2006. Multidimensional scaling of breathy voice quality: Individual differences in perception. *Journal of Voice,* 20, 211-222.

SHRIVASTAV, R., SAPIENZA, C. M. & NANDUR, V. 2005. Application of psychometric theory to the measurement of voice quality using rating scales. *Journal of Speech Language and Hearing Research,* 48, 323-335.

SILVERMAN, A. H. & BLACK, M. J. 1994. Efficacy of primary tracheoesophageal puncture in laryngectomy rehabilitation. *Journal of Otolaryngology,* 23, 370-377.

SIMPSON, I. C., SMITH, J. C. S. & GORDON, M. T. 1972. Laryngectomy: The influence of muscle reconstruction on the mechanism of oesophageal voice production. *Journal of Laryngology and Otology,* 86, 961-990.

SINGER, M. I. & BLOM, E. D. 1981. Selective myotomy for voice restoration after total laryngectomy. *Archives of Otolaryngology-Head & Neck Surgery,* 107, 670-673.

SINGER, M. I., BLOM, E. D. & HAMAKER, R. C. 1986. Pharyngeal plexus neurectomy for alaryngeal speech rehabilitation. *Laryngoscope,* 96, 50-53.

SOFRANKO, J. L. & PROSEK, R. A. 2012. The effect of experience on classification of voice quality. *Journal of Voice,* 26, 299-303.

STREINER, D. L. & NORMAN, G. R. 1995. *Health Measurement Scales A Practical Guide to Their Development and Use,* Oxford, Oxford University Press Inc.

TARDYMITZELL, S., ANDREWS, M. L. & BOWMAN, S. A. 1985. Acceptability and intelligibility of tracheoesophageal speech. *Archives of Otolaryngology-Head & Neck Surgery,* 111**,** 213-215.

TERRELL, J. E., LEWIN, J. S. & ESCLAMADO, R. 1995. Botulinum toxin injection for postlaryngectomy tracheoesophageal speech failure. *Otolaryngology - Head and Neck Surgery,* 113**,** 788-791.

TRUDEAU, M. D. 1987. A comparison of the speech acceptibility of good and excellent esophageal and tracheoesophageal speakers. *Journal of Communication Disorders,* 20**,** 41-49.

TSAO, Y. C., WEISMER, G. & IQBAL, K. 2006. Interspeaker variation in habitual speaking rate: Additional evidence. *Journal of Speech Language and Hearing Research,* 49**,** 1156-1164.

VAN AS, C. J. 2001. *Tracheoesophageal speech: A multidimensional assessment of voice quality*. University of Amsterdam.

VAN AS, C. J., KOOPMANS-VAN BEINUM, F. J., POLS, L. C. W. & HILGERS, F. J. M. 2003. Perceptual evaluation of tracheoesophageal speech by naive and experienced judges through the use of semantic differential scales. *Journal of Speech Language and Hearing Research,* 46**,** 947-959.

VAN AS-BROOKS, C. J., HILGERS, F. J. M., KOOPMANS-VAN BEINUM, F. J. & POLS, L. C. W. 2005. Anatomical and functional correlates of voice quality in tracheoesophageal speech. *Journal of Voice,* 19**,** 360-372.

VAN DEN HOOGEN, F. J. A., VAN DEN BERG, R. J. H., OUDES, M. J. & MANNI, J. J. 1998. A prospective study of speech and voice rehabilitation after total laryngectomy with the low-resistance Groningen, Nijdam and Provox voice prostheses. *Clinical Otolaryngology,* 23**,** 425-431.

VAN DER TORN, M., VAN GOGH, C. D. L., VERDONCK-DE LEEUW, I. M. D., FESTEN, J. M., VERKERKE, G. J. & MAHIEU, H. F. 2006. Assessment of alaryngeal speech using a sound-producing voice prosthesis in relation to sex and pharyngoesophageal segment tonicity. *Head and Neck-Journal for the Sciences and Specialties of the Head and Neck,* 28**,** 400-412.

VAN WEISSENBRUCH, R., KUNNEN, M., ALBERS, F. W. J., VAN CAUWENBERGE, P. B. & SULTER, A. M. 2000. Cineradiography of the pharyngoesophageal segment in post-laryngectomy patients. *Annals of Otology Rhinology and Laryngology,* 109**,** 311-319.

VLANTIS, A. C., GREGOR, R. T., ELLIOT, H. & OUDES, M. 2003. Conversion from a non-indwelling to a Provox (R) 2 indwelling voice prosthesis for speech rehabilitation: comparison of voice quality and patient preference. *Journal of Laryngology and Otology,* 117**,** 815-820.

WALSHE, M., MILLER, N., LEAHY, M. & MURRAY, A. 2008. Intelligibility of dysarthric speech: Perceptions of speakers and listeners. *International Journal of Language & Communication Disorders,* 43**,** 633-648.

WANG, C. P., TSENG, T. C., LEE, R. C. & CHANG, S. Y. 1997. The techniques of nonmuscular closure of hypopharyngeal defect following total laryngectomy: The assessment of complication and pharyngoesophageal segment. *Journal of Laryngology and Otology,* 111**,** 1060-1063.

WARD, E. C., HANCOCK, K., LAWSON, N. & VAN AS-BROOKS, C. J. 2011. Perceptual characteristics of tracheoesophageal speech production using the new indwelling Provox Vega voice prosthesis: A randomized controlled crossover trial. *Head and Neck-Journal for the Sciences and Specialties of the Head and Neck,* 33**,** 13-19.

WARD, E. C., KOH, S. K., FRISBY, J. & HODGE, R. 2003. Differential modes of alaryngeal communication and long-term voice outcomes following pharyngolaryngectomy and laryngectomy. *Folia Phoniatrica Et Logopaedica,* 55**,** 39-49.

WATSON, J. B. & WILLIAMS, S. E. 1987. Laryngectomees and non-laryngectomees perceptions of 3 methods of alaryngeal voicing. *Journal of Communication Disorders,* 20**,** 295-304.

WEBB, A. L. 2005. *An evaluation of the reliability, validity and responsiveness of selected perceptual assessments of outcome for use with voice disorders.* PhD thesis, Newcastle University..

WEBB, A. L., CARDING, P. N., DEARY, I. J., MACKENZIE, K., STEEN, I. N. & WILSON, J. A. 2007. Optimising outcome assessment of voice interventions, I: reliability and validity of three self-reported scales. *Journal of Laryngology and Otology,* 121**,** 763-767.

WEBB, A. L., CARDING, P. N., DEARY, I. J., MACKENZIE, K., STEEN, N. & WILSON, J. A. 2004. The reliability of three perceptual evaluation scales for dysphonia. *European Archives of Oto-Rhino-Laryngology,* 261**,** 429-434.

WETMORE, S. J., KRUEGER, K. & WESSSON, K. 1981. The Singer-Blom speech rehabilitation procedure. *Laryngoscope,* 91**,** 1109-1117.

WETMORE, S. J., RYAN, S. P., MONTAGUE, J. C., KRUEGER, K., WESSON, K., TIRMAN, R. & DINER, W. 1985. Location of the vibratory segment in tracheoesophageal speakers. *Otolaryngology-Head and Neck Surgery,* 93**,** 355-361.

WILLIAMS, S. E. & WATSON, J. B. 1985. Differences in speaking proficiencies in 3 laryngectomee groups. *Archives of Otolaryngology-Head & Neck Surgery,* 111**,** 216-219.

WILLIAMS, S. E. & WATSON, J. B. 1987. Speaking proficiency variations according to method of alaryngeal voicing. *Laryngoscope,* 97**,** 737-739.

WILSON, D. K. 1987. *Voice Problems in Children,* Baltimore, Williams and Wilkins.

WILSON, J. A. 1997. Upper oesophageal sphincter manometry. *In:* EVANS, D. F. & BUCKTON, G. K. (eds.) *Clinical Measurement in Gastroenterology* Oxford: Blackwell Science Ltd. .

WILSON, J. A., WEBB, A., CARDING, P. N., STEEN, I. N., MACKENZIE, K. & DEARY, I. J. The Voice Symptom Scale (VoiSS) and the Vocal Handicap Index (VHI):A comparison of structure and content. 4th Congress of the European-Laryngological-Society, Sep 2002 Brussels, BELGIUM. Blackwell Publishing Ltd, 169-174.

WUYTS, F. L., DE BODT, M. & VAN DE HEYNING, P. H. 1999. Is the reliability of a visual analogue scale higher than an ordinal scale? An experiment with the GRBAS scale for the perceptual evaluation of voice. *Journal of Voice,* 13, 508-517.

YANG, C. C., LEE, J. C., WU, K. C. & CHANG, S. H. 2011. Voice and speech outcomes with radial forearm free flap-accompanied phonation tube after total pharyngolaryngectomy of hypopharyngeal cancer. *Acta Oto-Laryngologica,* 131, 847-851.

YIU, E. M. L., CHAN, K. M. K. & MOK, R. S. M. 2007. Reliability and confidence in using a paired comparison paradigm in perceptual voice quality evaluation. *Clinical Linguistics & Phonetics,* 21, 129-145.

ZORMEIER, M. M., MELECA, R. J., SIMPSON, M. L., DWORKIN, J. P., KLEIN, R., GROSS, M. & MATHOG, R. H. 1999. Botulinum toxin injection to improve tracheoesophageal speech after total laryngectomy. *Otolaryngology-Head and Neck Surgery,* 120, 314-319.

ZRAICK, R. I., KEMPSTER, G. B., CONNOR, N. P., THIBEAULT, S., KLABEN, B. K., BURSAC, Z., THRUSH, C. R. & GLAZE, L. E. 2011. Establishing validity of the Consensus Auditory-Perceptual

Evaluation of Voice (CAPE-V). *American Journal of Speech-Language Pathology,* 20, 14-22.