Measuring voice outcomes: state of the science review

Citation for published version:

Digital Object Identifier (DOI):
10.1017/S0022215109005398

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Journal of Laryngology and Otology

Publisher Rights Statement:

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Measuring voice outcomes: state of the science review

P N Carding, J A Wilson, K MacKenzie*, I J Deary†

Abstract
Researchers evaluating voice disorder interventions currently have a plethora of voice outcome measurement tools from which to choose. Faced with such a wide choice, it would be beneficial to establish a clear rationale to guide selection. This article reviews the published literature on the three main areas of voice outcome assessment: (1) perceptual rating of voice quality, (2) acoustic measurement of the speech signal and (3) patient self-reporting of voice problems. We analysed the published reliability, validity, sensitivity to change and utility of the common outcome measurement tools in each area. From the data, we suggest that routine voice outcome measurement should include (1) an expert rating of voice quality (using the Grade-Roughness-Breathiness-Asthenia-Strain rating scale) and (2) a short self-reporting tool (either the Vocal Performance Questionnaire or the Vocal Handicap Index 10). These measures have high validity, the best reported reliability to date, good sensitivity to change data and excellent utility ratings. However, their application and administration require attention to detail. Acoustic measurement has arguable validity and poor reliability data at the present time. Other areas of voice outcome measurement (e.g. stroboscopy and aerodynamic phonatory measurements) require similarly detailed research and analysis.

Key words: Dysphonia; Voice Outcomes; Voice Disorders; Voice Quality; Voice Handicap; Self-Reported Voice Measures

Introduction
The rise of evidence-based medicine and ‘payment by results’ has driven a need for sophisticated and robust clinical outcome data in all areas of medicine. Outcome measurement tools need to have established reliability, validity, sensitivity to change and utility in order to be clinically useful and to enable confident assessment of a disorder as it changes following intervention.1 ‘Reliability’ refers to the internal consistency and stability of the tool,2 free from random error3 or unwanted variation.4 ‘Validity’ is concerned with the relevance of a tool or the extent to which the instrument measures what it purports to measure.2,3 If the reliability of a measure can be seen as its trustworthiness, then validity can be thought of as its truthfulness.5 ‘Sensitivity to change’ refers to an instrument’s responsiveness and ability to detect clinically important changes.6 ‘Utility’ is a measure of the ease of use of a tool for both the clinician and patient. Aspects such as patient discomfort and inconvenience (e.g. the time required to complete the task) are also important here.

Several areas of voice outcome measurement have been subjected to systematic international research over the past decade. Although different voice disorders may be treated with different types of intervention (e.g. pharmacological, surgical, behavioural, mechanical or psychological, or a combination), similar voice outcome measurements can be applied to all situations.

This review concentrates on three areas of voice outcome measurement that have been subjected to extensive research: (1) perceptual rating of voice quality, (2) acoustic measurement of the speech signal and (3) patient self-reporting of voice problems. The outcome measurement tools for each area are discussed with respect to their reliability, validity, sensitivity to change and utility. It is clear that there are a number of other areas of voice outcome measurement which require similarly detailed research – for example, endoscopic laryngeal interpretation (including stroboscopy) and aerodynamic phonatory measurements. However, to date no such data have been published.

Perceptual rating of voice quality
Auditory perceptual rating of voice quality involves an expert listener judging a voice sample according to various vocal parameters, and (in most cases)
marking the extent to which the voice deviates from a perceived ‘normal’ range.\(^7\) Perceptual voice quality rating is considered by all voice clinicians to be an essential outcome measure.\(^1\) There are a number of formal voice quality rating scales available. Three of the most commonly used scales in the UK are the Buffalo Voice Profile,\(^8\) the Vocal Profile Analysis scheme\(^9\) and the Grade-Roughness-Breathiness-Asthenia-Strain scale.\(^10\) An additional rating scale has recently emerged – the Consensus of Auditory Perceptual Evaluation Voice scale, which incorporates the parameters of grade, roughness, breathiness, asthenia and strain, and also allows for additional dimensions to be added.\(^11\)

**Reliability**

Several studies have established good reliability for the Grade-Roughness-Breathiness-Asthenia-Strain scale in the hands of expert users.\(^12\)–\(^14\) To our knowledge, there have not been any studies examining the reliability of the Consensus of Auditory Perceptual Evaluation Voice scale, the most recent perceptual voice quality rating scale suggested by the American Speech and Hearing Association. Webb et al. have provided the only evidence for the comparative internal consistency, repeatability and reliability of three commonly used voice quality rating scales.\(^15\) Webb and colleagues’ study was conducted under optimal rating conditions using seven highly experienced voice clinicians. A judgement of ‘overall’ voice severity was the most robust rated parameter in terms of inter- and intra-rater reliability (with reliability coefficients of 0.78 and 0.81, respectively). The Grade-Roughness-Breathiness-Asthenia-Strain scale was reliable across all parameters (inter-rater reliability coefficients ranged from 0.68 to 0.70 and intra-rater reliability coefficients from 0.69 to 0.79) except strain (with an inter-rater reliability coefficient of 0.48). Almost all of the component parameters of the Buffalo Voice Profile and Vocal Profile Analysis scales were found to have either poor or moderate reliability (i.e. below 0.50).

**Validity**

Perceptual voice rating has strong content validity, since most patients seek help for a voice disorder based on the sound of their voice. In addition, improvement in voice quality is the outcome by which interventions are judged to be successful.\(^1,7\) The criterion validity of perceptual voice rating (i.e. does it measure what it purports to measure?) has been demonstrated by highly significant correlations between Grade-Roughness-Breathiness-Asthenia-Strain scale ratings and self-perception and self-reporting scale scores.\(^16\) In this particular study, the strongest correlation (Spearman’s correlation coefficient of 0.32) was between the ‘overall’ grade of voice severity and the Vocal Performance Questionnaire total score (see below).

**Sensitivity to change**

To date, there has only been one quantitative study assessing the responsiveness to change (following intervention) of auditory perceptual voice quality ratings. Steen et al.\(^17\) compared effect sizes of the component parameters of the Grade-Roughness-Breathiness-Asthenia-Strain scale in a cohort of 144 patients following voice therapy and phonosurgery. For subjects undergoing voice therapy, there were significant, small-to-medium effect sizes. All of the Grade-Roughness-Breathiness-Asthenia-Strain scale parameters except roughness showed moderate effect sizes (the standard deviation (SD) ranged from 0.32 to 0.57). Roughness ratings generally showed less responsiveness to change following either voice therapy or surgery (effect size SDs ranged from 0.16 to 0.29).

**Utility**

Perceptual voice evaluation can be quick to perform and succinct, and the results easily communicable between clinicians. It is also non-invasive, readily available and can be performed ‘live’ in the clinic. However, when undertaking external validation, voice samples should be recorded using high quality recording equipment (preferably in a soundproof room). It is important to note that the task requires highly trained clinicians in order to be performed adequately. A review of practice amongst UK experts in voice perception analysis concluded that the absolute minimum requirement for observer voice assessment in the clinical setting was the use of the Grade-Roughness-Breathiness-Asthenia-Strain scale.\(^7\)

**Acoustic measures of voice quality**

Acoustic analysis of the voice signal involves computerised measurement of specific properties of the sound waveform as produced by the patient. For the purposes of voice outcome measurement, the three most commonly used acoustic parameters are ‘jitter’ (i.e. cycle-to-cycle frequency perturbation), ‘shimmer’ (i.e. cycle-to-cycle amplitude perturbation) and harmonics to noise ratio (an expression of aperiodic to periodic sound). In most published papers, these parameters are measured during ‘steady state’ vowel production.

**Reliability**

Steady state acoustic vowel analysis has been reported to have only moderate reliability, for both intra- and inter-system comparisons and repeated measures (i.e. within-subject) analysis.\(^18\)–\(^20\) Carding et al. studied a group of dysphonic patients and found that test–retest reliability (i.e. stability) coefficients were at best moderate for jitter (0.45 (95 per cent confidence interval (CI) = 0.23–0.70)) and shimmer (0.40 (95 per cent CI = 0.18–0.67)) and lower for harmonics to noise ratio (0.33 (95 per cent CI = 0.11–0.63)).\(^21\) The intra-class correlation coefficient for reliability improved when acoustic analysis was performed on non-dysphonic or near-normal (i.e. type one) voice signals (jitter = 0.73 (95 per cent CI = 0.58–0.85), shimmer = 0.55 (95 per cent CI = 0.35–0.74) and harmonics to noise...
carding

ceptual voice quality features; however, others (e.g.
exists between specific parameters and certain per-
highly complex correlation. 20,21 Furthermore, many
measurement, and have argued for a more represen-
tative measure of connected speech. 23 However,
may have high content validity. However, acoustic
within a typical voice pathology population may not
be analysable in this way.21 Criterion validity has
not been clearly established. Some authors (e.g.
Carding et al.) have suggested that a close correlation
exists between specific parameters and certain per-
ceptual voice quality features; however, others (e.g.
Carding et al.) have reported a less convincing and
highly complex correlation. 20,21 Furthermore, many
authors have debated the validity of steady state
vowel analysis for the purposes of voice outcome
measurement, and have argued for a more represen-
tative measure of connected speech. 23 However,
more complex speech signals are inherently more dif-
cult to analyse, and data are sparse.

Sensitivity to change

There is limited information on the comparative
sensitivity of acoustic voice analysis parameters for
measuring voice change. Carding et al. found poor-to-moderate effect sizes when assessing the
sensitivity of such parameters in detecting change
following treatment. 21 Following surgery, the effect
sizes (SD) for this assessment were: jitter = 0.32,
shimmer = 0.28 and harmonics to noise ratio =
0.34; those following voice therapy were: jitter =
0.47, shimmer = 0.34 and harmonics to noise
ratio = 0.32.

Utility

Good reliability of acoustic measurement is difficult
to achieve in moderately dysphonic (aperiodic)
voices and is of very limited value in cases of severely
dysphonic voice. The process of acquiring and ana-
lysing the speech sound signal is time-consuming
(approximately one hour per patient) and requires
considerable voice laboratory expertise.

Patient self-reporting

There are a number of voice-specific patient self-
reporting tools reported in the literature. Most of
the research activity over the past decade has concen-
trated on examining the Voice Handicap Index, the
Vocal Performance Questionnaire and the Voice
Symptom Scale. 24–26

Reliability

Several studies have examined the comparative
reliability of the Vocal Performance Questionnaire,
the Voice Handicap Index and the Voice Symptom
Scale. 16,17 In summary, based on assessment of 181
patients presenting with dysphonia, all three assess-
ment tools provided excellent internal consistency
(Cronbach’s coefficient = 0.81–0.95) and repeatabil-
ity (intra-class correlation coefficients: Voice Handi-
cap Index total = 0.83, Vocal Performance
Questionnaire = 0.75 and Voice Symptom Scale
total = 0.63). For baseline measures, therefore, cri-
teria other than reliability should direct the selection
of self-reporting tools.

Validity

Patient self-reporting has high content validity since,
unless patients are satisfied with their own voice,
little can claim to have been achieved in treatment.
Patient self-reporting also offers an opportunity to
obtain information about vocal handicap and disabil-
in, in addition to aspects of vocal quality. Further-
more, many dysphonic patients have a widely
fluctuating disorder (e.g. worse at the end of the
working day or the working week). Therefore, the
voice that is presented to the clinician in the voice
clinic may well not be representative of the overall
voice performance. 27 Self-reporting tools allow
the patient to give an overall voice rating, as opposed
to one based solely on vocal performance on the
day of consultation. 27

Criterion validity is more difficult to prove. A
central problem with many historic self-reporting
tools has been the physician-centred nature of their
derivation. Both the Voice Handicap Index and the
Vocal Performance Questionnaire suffer from this
limitation.

In this respect, the Voice Symptom Scale is con-
siderably superior to all previous voice self-reporting
tools, with 800 subjects participating in the final
development of the tool. 26,28 Criterion validity is
also affected by the internal component structure of
the self-reporting tool. Psychometric analysis of the
800 subjects’ Voice Symptom Scale responses
showed three distinct subscales: impairment (15
items), emotional response (eight items) and physical
symptoms (seven items). 28

In contrast, Rosen et al. assessed the Voice Handi-
cap Index and found a lack of statistically discreet
subscales. 29 Further factor analysis of the Voice Handi-
cap Index subscales revealed that only a single
factor was being measured. For these reasons, a
shorter, 10-item Voice Handicap Index was proposed.

Sensitivity to change

Several published studies have analysed the sensi-
tivity of self-reporting assessment tools for measur-
change following intervention. 16,17 Again, it
would appear that the Vocal Performance Question-
aire, Voice Handicap Index and Voice Symptom
Scale all show large effect sizes as regards sensitivity
to change following either voice therapy (SD results
being 1.04, 0.62 and 0.78, respectively) or surgery
(SD results being 0.82, 0.72 and 1.06, respectively).
In terms of sensitivity to change, the ability of the
Vocal Performance Questionnaire to demonstrate a treatment effect size of more than one (i.e. equal to the Voice Symptom Scale and somewhat higher than the Voice Handicap Index) is an impressive result for a short, 12-item questionnaire.

Utility

Deary et al. compared the Voice Handicap Index 10 (i.e. the shorter, 10-item version) with the Vocal Performance Questionnaire. Both were found to be similar, being short, convenient, internally consistent, uni-dimensional tools used to measure the severity of a voice disorder. Furthermore, Rosen et al. concluded that there was no benefit to using the full (30-item) version of the Voice Handicap Index rather than the shortened Voice Handicap Index 10. The use of an extended questionnaire (with a considerable risk of item redundancy) would appear to be required only for very specific reasons and requirements. In this latter case, it would appear that the Voice Symptom Scale may be most useful, since it has three discreet subscales.

Discussion

When measuring outcomes, the aim is to document significant change – i.e. change that is neither random nor unimportant. The established opinion is that voice outcome measurement should be multi-dimensional in nature. We have analysed the evidence base for three common types of voice outcome measurement tools: voice quality perceptual rating, acoustic measurement of the speech signal and patient self-reporting. We suggest that the selection of voice outcome measurement tool should be based on considerations of reliability, validity, sensitivity to change and utility. Whilst our research only extended into three areas of voice assessment, we would anticipate that a similar approach to the analysis of other tools (such as laryngeal endoscopy and stroboscopy, and aerodynamic phonatory measurement) may also yield valuable clinical information.

From our research findings, we recommend that routine voice outcome measurement should include (1) an expert rating of voice quality (probably using the Grade-Roughness-Breathiness-Asthenia-Strain scale) and (2) a short self-reporting tool (either the Vocal Performance Questionnaire or the Voice Handicap Index 10). These measures have high validity; the best reported reliability to date, good sensitivity to change and excellent utility ratings. These instruments are therefore likely to provide high quality outcome information irrespective of whether the treatment choice is phonosurgery, voice therapy, pharmacological therapy or a combination of several approaches.

The obvious limitation is that, in a clinical setting, expert rating of voice quality will probably be carried out by the treating clinician. We should remember that published studies relate only to blinded, controlled, independent evaluation of voice quality by an expert rater. The effect of clinical bias and the performance of less expert raters have not been fully examined.

However, this should not prevent clinicians from applying these measures in routine practice in order to determine the effectiveness of their treatments. Furthermore, with respect to voice quality ratings, we should not forget that clinician ratings may not always correlate with patient perceptions of their own voice quality scores.

The Voice Symptom Scale is certainly worth considering if a more detailed patient self-evaluation is required. The advantage of the Voice Symptom Scale over the Vocal Performance Questionnaire or the Voice Handicap Index 10 is that it includes a physical symptoms subscale. However, whilst information on these physical symptoms may be interesting to obtain, physical symptom subscale results do not seem to correlate with vocal outcome nearly as closely as different voice measures correlate with each other. A review of the impact of surgery according to the Voice Symptom Scale impairment subscale showed an effect size of one, with a corresponding Voice Symptom Scale emotional subscale of 0.69 but a physical symptoms subscale response of only 0.43. This result is perhaps predictable and indeed may be welcomed, as it suggests that these subscales may be obtaining information on an area of dysfunction which conventional strategies have yet to adequately address. Both the Voice Symptom Scale and the Vocal Performance Questionnaire are detailed in the appendices of this article.

Acoustic analysis of the speech signal would currently appear to have a limited clinical role. Reliability may be enhanced by recording and analysing multiple voice samples and averaging the results, but this is at the expense of utility. Perturbation measurements of selected vowel prolongations may be greatly enhanced by following a strict recording protocol. However, the value of this approach in measuring clinically useful change has yet to be established. Acoustic analysis of connected speech still appears to be in its infancy.

In a research context, there is no doubt that multi-dimensional analysis is best. Where high quality evidence exists, we should use it to guide our selection of the most robust voice outcome measures. However, limiting our data to that obtained by these tools only would be to the long term detriment of the development of knowledge in this area. For example, the general positive benefits of being a patient in a clinical trial mean that it would be very unwise to interpret research findings on the basis of self-reporting measures alone; however reliable they appear on statistical analysis. Clinical outcome data from laryngeal endoscopy, aerodynamic phonatory measurement and psychological impact assessment may all yield valuable data. It is however clear that these measures require considerable further attention, particularly with respect to reliability and sensitivity to change.

References

(c) Stops me doing a lot of things that involve using my voice
(d) Stops me doing most things that involve using my voice
(e) I can hardly do anything that involves me using my voice

7 In your opinion, do you think that your voice is ever difficult to hear or understand?
   (a) Not at all
   (b) A little difficult
   (c) Quite difficult
   (d) Very difficult
   (e) Extremely difficult

8 Do other people (e.g. close family) ever comment that your voice is difficult to hear or understand?
   (a) No comments
   (b) Occasional comments
   (c) Quite often there are comments
   (d) Frequent comments
   (e) Very frequent comments

9 Since your voice problem started, has your voice...?
   (a) Improved a lot
   (b) Improved a little
   (c) Not improved at all
   (d) Deteriorated a little
   (e) Deteriorated a lot

10 Since your voice problem started, have other people (e.g. close family) commented that your voice has improved?

Appendix 2. Voice Symptom Scale

Your name
Your date of birth
Today's date

Please circle one answer for each item
Please do not leave any blank items

1 Do you have difficulty attracting attention? Never Occasionally Some of the time Most of the time Always
2 Do you have problems singing? Never Occasionally Some of the time Most of the time Always
3 Is your throat sore? Never Occasionally Some of the time Most of the time Always
4 Is your voice hoarse? Never Occasionally Some of the time Most of the time Always
5 When talking in company do people fail to hear you? Never Occasionally Some of the time Most of the time Always
6 Do you lose your voice? Never Occasionally Some of the time Most of the time Always
7 Do you cough or clear your throat? Never Occasionally Some of the time Most of the time Always
8 Do you have a weak voice? Never Occasionally Some of the time Most of the time Always
9 Do you have problems talking on the telephone? Never Occasionally Some of the time Most of the time Always
10 Do you feel miserable or depressed because of your voice problem? Never Occasionally Some of the time Most of the time Always
11 Does it feel as if there is something stuck in your throat? Never Occasionally Some of the time Most of the time Always
12 Do you have swollen glands? Never Occasionally Some of the time Most of the time Always
13 Are you embarrassed by your voice problem? Never Occasionally Some of the time Most of the time Always
14 Do you find the effort of speaking tiring? Never Occasionally Some of the time Most of the time Always
15 Does your voice problem make you feel stressed and nervous? Never Occasionally Some of the time Most of the time Always
16 Do you have difficulty competing against background noise? Never Occasionally Some of the time Most of the time Always
17 Are you unable to shout or raise your voice? Never Occasionally Some of the time Most of the time Always
18 Does your voice problem put a strain on your family and friends? Never Occasionally Some of the time Most of the time Always
19 Do you have a lot of phlegm in your throat? Never Occasionally Some of the time Most of the time Always
20 Does the sound of your voice vary throughout the day? Never Occasionally Some of the time Most of the time Always
21 Do people seem irritated by your voice? Never Occasionally Some of the time Most of the time Always

(Continued)
Appendix 2. (Continued)

<table>
<thead>
<tr>
<th></th>
<th>Question</th>
<th>Never</th>
<th>Occasionally</th>
<th>Some of the time</th>
<th>Most of the time</th>
<th>Always</th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td>Do you have a blocked nose?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>Do people ask what is wrong with your voice?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>Does your voice sound creaky and dry?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>Do you feel you have to strain to produce voice?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>How often do you get throat infections?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>Does your voice ‘give out’ in the middle of speaking?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>Does your voice make you feel incompetent?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>Are you ashamed of your voice problem?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>Do you feel lonely because of your voice problem?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For office use:

Total Voice Symptom Scale score = . . . .
Impairment score (items 1, 2, 4, 5, 6, 8, 9, 14, 16, 17, 20, 23, 24, 25 & 27) (maximum 60) = . . . .
Emotional score (items 10, 13, 15, 18, 21, 28, 29 & 30) (maximum 32) = . . . .
Physical score (items 3, 7, 11, 12, 19, 22 & 26) (maximum 28) = . . . .

Please note that the Vocal Performance Questionnaire and Voice Symptom Scale are also available in electronic format (at http://www.entuk.org/clinical_outcomes/). This website also includes information about how to score the questionnaires, as well as several supporting publications.

Address for correspondence:
Prof Paul Carding,
Professor of Speech/Voice Pathology,
Dept of Speech, Voice and Swallowing,
Otolaryngology,
Freeman Hospital,
Newcastle upon Tyne NE7 7DN, UK.
E-mail: paul.carding@nuth.nhs.uk

Professor P Carding takes responsibility for the integrity of the content of the paper.
Competing interests: None declared