The AVQI with extended representativity:

external validity and diagnostic precision with 1058 voice samples

Ben Barsties ¹,²
Youri Maryn, PhD ¹,³,⁴

¹ Faculty of Medicine and Health Sciences, University of Antwerp, Belgium.
² Medical School, Hochschule Fresenius University of Applied Sciences, Hamburg, Germany
³ European Institute for ORL, Sint-Augustinus Hospital, Antwerp, Belgium.
⁴ Faculty of Education, Health & Social Work, University College Ghent, Belgium.
The program: Acoustic Voice Quality Index

What is AVQI and what are the developments?
http://www.iqsol.biz/

Next step of development in AVQI: AVQI 03.01
http://www.mrhandymanfranchise.co.uk/

External validation of AVQI 03.01
https://nl-nl.facebook.com/SecondTimeOfficial

Conclusion
http://stevenjdavies.co/trouble-phonics-conclusion/#
The Acoustic Voice Quality Index (i.e., AVQI)

What is AVQI:

• Voice assessment tool for clinical and research utility

• Objective-acoustic measurement

• Multivariate construct based on linear regression analysis that combines several acoustic markers to yield a single score that correlates reasonably with the auditory-perceptual judgment of overall dysphonia severity (i.e., G from GRBAS-scale)
The Acoustic Voice Quality Index (i.e., AVQI)

Figure 1. Oscillogram and narrowband spectrogram (window length = 0.03 s) of a concatenated voice sample. There are three areas. The left portion reflects the first two sentences of the “Papa en Marloes” text. The right area reflects the middle 3 s of a sustained /a/. Both samples were separated by 1 s of silence (area in the middle).
Figure 2. Oscillogram and narrowband spectrogram (window length = 0.03 s) of a concatenated voice sample after extraction by using the algorithm of Parsa and Jamison (2001) in only continuous speech. There are two areas. The left area reflects the concatenated voiced segments of the first two sentences of the “Papa en Marloes” text. The right area reflects the middle 3 s of a sustained /a/.
The Acoustic Voice Quality Index (i.e., AVQI)

- six-factor acoustic model based on linear regression analysis
  1. smoothed cepstral peak prominence (i.e., CPPs),
  2. harmonics-to-noise ratio (i.e., HNR),
  3. shimmer local (i.e., Shim),
  4. shimmer local dB (i.e., ShdB),
  5. general slope of the spectrum (i.e., Slope), and
  6. tilt of the regression line through the spectrum (i.e., Tilt)
AVQI development

Meta-analysis in overall voice quality and acoustic measurement (2009)

Why the combination of continuous speech and sustained vowel (2009, 2012)

The model Acoustic Voice Quality Index (i.e., AVQI) (2010)

Internal and external validity of AVQI (2010)
AVQI development

Cross-linguistic in German (2012)
Cross-linguistic in English with pediatric population (2012)
Test-retest and internal consistency (2013)
Cross-linguistic in further Germanic languages (2014)
AVQI development

AVQI completely in Praat (2015)

Validity of AVQI in the Korean language (In Submission)

Regulation of internal consistency of AVQI with balanced out proportions of the two different speech tasks (In Submission)

Higher representativity and validity of the new AVQI 03.01 (In Submission)

Cross-linguistic of AVQI 02.02 & AVQI 03.01 in Finnish

Cross-linguistic of AVQI 02.02 & AVQI 03.01 in English

Future
Does the internal consistency of the AVQI improve when the proportions of sustained phonation and voiced continuous speech are adapted to reach higher ecological validity?

- Yes, equalization showed balanced out influence of the final AVQI score

Barsties B, Maryn Y. The improvement of internal consistency of the Acoustic Voice Quality Index. In Submission
Equal proportion of the 2 speech types in AVQI

• Definition of equal proportion
  ➢ continuous speech with customized length (i.e., to correspond with three seconds of voiced continuous speech after extraction) plus three seconds of sustained phonation

• Hand-marked detection for customized length in continuous speech
  1. Whole text of each participant was completely extracted of all voiceless segments
  2. The customized cut-off point was found hand-marked in the original text without extraction (i.e., correspondence to the extracted first three sec from step one)
  3. Define the hand-marked cut-off point manually (i.e., using oscilogram- and narrowband spectrogram view, the pitch contour, and the auditory feedback)
  4. Control step: extra run of the customized hand-marked segment with acceptable tolerant margin of ± 0.1 sec across hand-marked judging

Barsties B, Maryn Y. The improvement of internal consistency of the Acoustic Voice Quality Index. In Submission
Equal proportion of the 2 speech types in AVQI

• Recommendation:
  ➢ Individual cut-off point not practicable:
    o an arbitrary point in the text that is almost impossible to detect with a computer algorithm being backtracked in the normal text with voiceless and voice segments
  ➢ Use of 34 syllables correspondence to hand-marked detection:
    o duration time
    o AVQI score
  ➢ New weighted AVQI equation:

\[
AVQI_{03.01} = (2.48722 - (0.256 \times \text{CPPs}) + (0.028 \times \text{HNR}) - (0.108 \times \text{Shim}) + (2.032 \times \text{ShdB}) + (0.002 \times \text{Slope}) - (0.02 \times \text{Tilt})) \times 3.246753
\]

Barsties B, Maryn Y. The improvement of internal consistency of the Acoustic Voice Quality Index. In Submission

© Ben Barsties & Youri Maryn, PhD
External Validation of AVQI 03.01

• **Subjects** (ENT caseload of the Sint-Jan General Hospital in Bruges, Belgium):
  - 970 patients with dysphonia (non-organic: n=221, organic: n=749)
  - 88 healthy subjects without any voice complaints and voice disorders
  - M: n=386, F: n=672

• **Voice Samples:**
  - First 34 syllables plus 3 sec [a:] in WAV format
  - AKG C420 head-mounted condenser microphone, Kay Pentax CSL 4500
  - Recordings implemented in a soundproof booth
  - SNR (Deliyski et al. 2005, 2006): mean=38.56 dB, SD3.78 dB

External Validation of AVQI 03.01

Auditory perceptual evaluation:

- Expert panel of 12 native Dutch speech-language therapists (M: n=3, F: n=9)
- Panel: specialized in voice disorders (range of profession experience between 4 to 41 years; mean=22.3 years, SD=11.4 years)
- Rating G from the GRBAS-scale (i.e., 0 = normal, 1 = slightly disordered, 2 = moderately disordered, 3 = severely disordered)
- Judgment circumstances:
  - ambient noise level lower than 40 dB (A)
  - Presenting voice samples individually at comfortable loudness level through headphones
  - Allowing to repeat each voice sample as often as necessary to make a final decision of judgment
  - Randomized voice samples in different sessions (i.e., each session about 250 voice samples)
  - After every 25th voice sample short break
  - Six anchor voices at the beginning and after every 25th voice sample
  - 104 randomly selected repeated voice samples at the end of the judgment without informing the listeners that stimuli were repeated
External Validation of AVQI 03.01

• Statistics:
  ➢ Reliability:
    o Intra-rater reliability: Cohen’s Kappa coefficient (i.e., Ck)
    o Inter-rater reliability: Fleiss Kappa coefficient (i.e., Fk)
    o Significant changes in reliability bootstrapping with 10,000 replications (i.e., considered statistically significant at \( p \leq 0.01 \))
  ➢ Establishing a homogenous and representative rater panel:
    1. No significant differences of Ck values between all pairs of raters
    2. All raters consider Ck \( \geq 0.41 \)
    3. Leftover raters with comparable high intra-rater reliability were used to find the significantly (i.e., considered statistically significant at \( p \leq 0.01 \)) homogenous panel with Fk \( \geq 0.41 \) using bootstrap-backward method
  ➢ Concurrent validity
    o Spearman rank-order correlation coefficient (i.e., \( r_s \))
    o Coefficient of determination (i.e., \( r^2_s \))
  ➢ Diagnostic accuracy
    o receiver operating characteristic (i.e., ROC): area under ROC-curve (i.e., \( A_{ROC} \)), sensitivity, specificity
    o Likelihood ratio (i.e., LR): likelihood ratio for a positive result (i.e., LR+), likelihood ratio for a negative result” (i.e., LR−)
Results: listener reliability

- **Intra-rater reliability:**
  - no significant differences in Ck values between all pairs of the 12 raters (t= 12.824, p= 0.306)
  - 11 raters: Ck between 0.41 to 0.58
  - 1 rater: Ck= 0.32 (=excluded)

- **Inter-rater reliability:**
  - 11 raters: Fk= 0.39 with significant differences
  - 4th round of bootstrap-backward method: Fk= 0.43 plus for the first time no significant differences
  - Final panel: 8 judges
Results: listener reliability

- Frequency distribution of the mean auditory-perceptual overall voice quality ratings (average of G-scores of the 8 identified judges) of the 1058 concatenated voice samples
Results: concurrent validity

- $r_s = 0.815, p = 0.000$
- 66.4% (i.e., $r^2_s = 0.664$) of the variance in $G_{mean}$ was accounted for by AVQI
Results: diagnostic accuracy

Statistics illustrating the AVQI’s ability to differentiate normophonia vs. dysphonia and validity to auditory-perceptual judgment in AVQI 03.01.

<table>
<thead>
<tr>
<th></th>
<th>$A_{ROC}$</th>
<th>AVQI Threshold</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>$LR^+$</th>
<th>$LR^-$</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVQI 03.01</td>
<td>0.923</td>
<td>2.43</td>
<td>0.785</td>
<td>0.932</td>
<td>11.54</td>
<td>0.23</td>
</tr>
</tbody>
</table>
Conclusion

• Confirm AVQI as a robust and ecologically valid measurement to objectify overall voice quality

• AVQI improved through the development in equal proportion of continuous speech and sustained phonation:
  ➢ Balanced out internal consistency of final AVQI result
  ➢ Improvement of ecological validity
  ➢ Higher results in concurrent validity than AVQI 02.02
  ➢ Higher results in diagnostic accuracy than AVQI 02.02
  ➢ Prove in two investigations with totally 1118 subjects in representative voice clinic population:
    ➢ different ages
    ➢ genders
    ➢ different types and degrees of voice quality
    ➢ including nonorganic as well as organic laryngeal pathologies
Contact

Ben Barsties: ben.barsties@t-online.de

Youri Maryn, PhD: yourimaryn@vvl.be
AVQI 02.02 vs. AVQI 03.01

- Differences in perceived judgment of evaluation procedure between AVQI 02.02 and AVQI 03.01

<table>
<thead>
<tr>
<th>N=100</th>
<th>17 syllables plus 3 sec [a:]</th>
<th>Mean syllables 35.5 plus 3 sec [a:]</th>
<th>93 syllables plus 3 sec [a:]</th>
</tr>
</thead>
<tbody>
<tr>
<td>G_{mean} (SD)</td>
<td>1.23 (0.72) *</td>
<td>1.37 (0.70)</td>
<td>1.41 (0.65)</td>
</tr>
</tbody>
</table>

* Paired t-test: Significant differences among the other two groups (p=0.000)

Barsties B, Maryn Y. The impact of voice sample duration in the auditory-perceptual judgment of overall voice quality. In Submission
AVQI 02.02 vs. AVQI 03.01

<table>
<thead>
<tr>
<th>Patient</th>
<th>Sex</th>
<th>Age</th>
<th>Pathology</th>
<th>$G_{\text{mean}}$ in AVQI 02.02</th>
<th>$G_{\text{mean}}$ in AVQI 03.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>34</td>
<td></td>
<td>Polypoid mucosa (edema)</td>
<td>1.6</td>
<td>2.2</td>
</tr>
</tbody>
</table>