The validity of the Acoustic Breathiness Index in the evaluation of breathy voice quality: A Meta-Analysis

Ben Barsties v. Latoszek1,2 | Geun-Hyo Kim3 | Jonathan Delgado Hernández4 | Kiyohito Hosokawa5,6,7 | Marina Englert8,9 | Katrin Neumann2 | Svetlana Hetjens10

Abstract

Background: The evaluation of voice quality with acoustic measurements is useful to objectify the diagnostic process. Particularly, breathiness was highly evaluated and the Acoustic Breathiness Index (ABI) might have promising features.

Objective of review: The goal of the present meta-analysis is to quantify, from existing cross-validation studies, the evidence for the diagnostic accuracy of ABI, including its sensitivity and specificity.

Type of review: Meta-analysis.

Search strategy: We searched in MEDLINE, Google Scholar and Science Citation Index, and as manual search for the term Acoustic Breathiness Index from inception to February 2020. Studies were included that used equal proportion of continuous speech and sustained vowel segments, a recording hardware with a sufficient standard for voice signal analyses, the software Prat for signal processing and the customised Prat script, and two groups of subjects (vocally healthy and voice-disordered). Furthermore, the diagnostic accuracy of ABI was measured.

Evaluation method: The primary outcome variable was ABI. The score ranged from 0 to 10 with varying thresholds according to different languages to determine the absence or presence of breathiness. A meta-analysis was performed according to the Preferred Reporting Items for Systematic Reviews and Meta-analyses of diagnostic test accuracy study guidelines. Data were extracted, and the risk of bias was assessed using the QUADAS-2 tool. The pooled sensitivity and specificity of ABI were determined using a summary receiver operating characteristic (SROC) approach to calculate also a weighted threshold value of ABI with its sensitivity and specificity.

Results: A total of 34 unique citations were screened, and 10 full-text articles were reviewed, including six studies. In total, 3603 voice samples were considered for further analysis separating into 467 vocally healthy and 3136 voice-disordered voice samples. The pooled sensitivity was 0.84 (95% CI, 0.83-0.85), and the pooled specificity was...
1 | INTRODUCTION

The assessment of voice quality is a crucial part in voice diagnostics, complicated by the fact that voice quality is an imprecisely defined term. Voice quality is generally described as perceptually multidimensional construct which cannot be measured by a single parameter. Various diagnostic protocols use different terms in the perceptual evaluation of voice quality such as grade or overall severity, hoarseness, roughness, breathiness, asthenia and strain.

In case of major subtypes of perceived abnormal voice quality, the terms such as breathiness, roughness and strain have been received wide acceptance. Breathiness can be defined as turbulent noise that is excessively high in frequency during phonation resulting from air leakage during glottal closure. Roughness is the result of irregular vocal fold vibrations characterised by frequent, rapid and random changes in the regular vocal fold movement patterns. Strain, also described as pressedness, means increased vocal effort or “hyperfunction” and can be attributed to an unintentional excessive contraction of phonation, breathing, articulation and neck muscles during phonation leading to increased subglottal pressure and an increased degree of vocal fold adduction during phonation.

Next to overall voice quality, the intra- and inter-rater reliability of these dimensions of voice quality reached the highest values both within and among single raters and within rater panels. However, there is still a notable variability of intra- and inter-rater reliability in all auditory-perceptual features caused by different factors.

In addition and as an alternative to auditory-perceptual judgement, acoustic measurements of the voice signal have a high potential in the assessment of voice quality because they are based on objective measurements and are therefore less-biased by subjective factors. Moreover, they are the most frequently used diagnostic tool in voice research for identifying voice anomalies. Furthermore, the earlier described pitfalls in the judgement of perceived auditory features can be overcome with acoustic methods, thus increasing the reliability and validity of voice assessments. Acoustic methods used so far are based on time, frequency, amplitude and quefrency domains to analyse acoustically the voice signal. However, recent research has shown that the combination of several acoustic parameters in one model or index provides better diagnostic accuracy and higher concurrent validity than single metrics in the evaluation of perceived features of voice quality.

0.92 (95% CI, 0.89-0.94). The area under the curve of the SROC curve of this analysis showed an excellent value of 0.94. The weighted ABI threshold was determined at 3.40 (sensitivity: 0.86, 95% CI, 0.84-0.87; specificity: 0.90, 95% CI 0.88-0.92).

Conclusions: The results confirm the ABI as robust and valid objective measure for evaluating breathiness.

Keypoints
- To objectify the diagnostic accuracy of the Acoustic Breathiness Index (ABI) for the identification of abnormal voice quality.
- A meta-analysis identified 6 of 152 studies with 467 typical and 3136 pathological voice samples that were included in a joint analysis. The area under the curve of the summary receiver operating characteristic curve of this analysis revealed an excellent value of 0.94; the weighted ABI threshold was 3.40 (sensitivity 0.86, specificity 0.90).
- ABI is a valid diagnostic tool to objectify breathiness of phonation across languages.

Yet, there are still problems to develop a highly valid equivalent for roughness and strain based on acoustic measurements. Because roughness is quite complex in nature, it is difficult to capture all variants of the rough vocal sound (eg multiphonia, acoustic irregularities or vocal fry) with a single model.

The correlations between perceived strain and spectral-cepstral and other acoustic measures revealed in some studies moderate to high results, but an evaluation of the diagnostic accuracy of these methods and the validity for the perceived strain is still lacking. Hence, there is currently no acoustic index or model that can accurately assess perceived strain for clinical or research purposes.

For breathiness, an adequate model seems to have been developed. The Acoustic Breathiness Index (ABI) is a multiparametric nine-variable acoustic measure to quantify the degree of breathiness with a single score based on concatenated samples of continuous speech and the sustained vowel/a/. This index has been developed by the first author. Both speech tasks with voiced segments of continuous speech of three seconds and intermediate vowel segments with sustained vowel/a/ of also three seconds are necessary for a high ecological validity of the voice quality assessment. The initial study of ABI showed a strong correlation between ABI scores and perceived breathiness judgements, which was confirmed with cross-correlations statistic. Furthermore, the diagnostic accuracy (ability of the measure to discriminate between the target condition and health) was high both in terms of sensitivity and in terms of specificity.
Neither age and gender nor roughness significantly affects ABI in the evaluation of natural voices. In addition, the ABI also indicates highly sensitive therapy-related voice quality changes.

The ABI score ranges from 0 to 10, and the higher an ABI score, the more severe breathiness and vice versa. The initial study of ABI investigated a Dutch-speaking population, and the reported threshold was 3.44. Because continuous speech is part of the ABI, inter-language phonetic differences may influence the outcome of an acoustic-based voice quality measurement. Therefore, cross-validation studies are needed to investigate the ABI's level of validity for different languages. In particular, the diagnostic accuracy of ABI, including its sensitivity and specificity, is of primary importance in determining its language-related thresholds.

Thus, the aim of the meta-analysis presented here was (a) to estimate the diagnostic accuracy of this tool together with its sensitivity and specificity levels, based on cross-validation studies and (b) to calculate an overall weighted ABI threshold of all included studies.

2 | METHOD

2.1 | Search strategy

Firstly, a systematic literature search was performed to identify studies in electronic databases. We searched in MEDLINE, Google Scholar and Science Citation Index starting with the first publication of ABI, which was published in the year 2017 to February 2020.

Secondly, a manual search was utilised in grey literature sources. The hand search strategy of relevant scientific reports for the meta-analysis was done in seven languages (English, German, Dutch, Japanese, Korean, Spanish and Brazilian Portuguese). Except for finding grey and non-English literature, a manual search was performed because electronic databases do not always include relevant search terms in the titles or abstracts, or publications are not indexed with terms that allow them to be easily identified as relevant works for the present study. The Acoustic Breathiness Index was used as the only search term, because it has existed as a proper term without overlapping meanings since 2017.

2.2 | Inclusion and exclusion of studies

Studies were included that used the ABI analyses according to equal processing methodology of the initial ABI study: equal proportion of continuous speech and sustained vowel segments in acoustic analyses, recording hardware meeting sufficient standard for voice signal analyses, using the software Praat for signal processing and the customised Praat script for ABI analysis, and consideration of two groups of subjects (a vocally healthy group and a heterogenic voice-disordered group). Potential studies, which will be included in the present meta-analysis, had one of the objectives to analyse the validity levels of ABI. Therefore, and because the present study focused on the diagnostic accuracy of the ABI, other validity aspects such as concurrent validity and internal validity were excluded.

Inter-study differences in data acquisition were tolerated in the present meta-analysis. Each study has unique acoustical settings and configurations in terms of hardware, limiting the comparability between studies. However, many recent studies which used acoustic measurements for its investigation followed recommendations for hardware and recording circumstances to increase comparability of differences in room acoustics, microphone type and placement.

2.3 | Critical appraisal of included studies and data abstraction

Data were abstracted, and the quality of each study was appraised using a customised data abstraction form. The QUADAS-2 tool was applied to assess the risk of bias and applicability to the research question. Based on the QUADAS-2 tool, each article was evaluated for risk of bias in 4 domains. High risk of bias in each domain was defined as follows: (a) patient selection, the study enrolment was not consecutive; (b) index test, the ABI results were based on unequal proportion of continuous speech and sustained vowel segments, the recording hardware and recording environment are insufficient for voice signal analyses, and the software Praat and the customised Praat script were not used for signal processing and ABI calculation; (c) reference standard, the reference standard for assessing abnormal voice quality does not consider several factors of high reliability such as a rater panel, a blinding procedure and anchor voices for the raters, and consideration of fatigue, attention and low concentration level by the listeners who have to evaluate hundreds of voice samples; and (d) flow and timing, the interval between reference standard and ABI based on different voice sample recordings.

With regard to the assessment of applicability, each article was evaluated for low and high concern for applicability to the research question. Using the patient selection, index test and reference standard domains, we defined low applicability concern as follows: (a) patient selection, there is an acceptable range between vocally healthy and voice-disordered subjects; (b) index test, the ABI was performed according to the standards of the initial study to develop ABI; and (c) reference standard, the diagnosis of voice-disordered subjects was based on multidimensional voice assessments from otolaryngology.

We reported our methods and findings in accordance with Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) standards regarding the diagnostic test accuracy studies guidelines.

2.4 | Statistical analysis

Statistical analyses were completed using SAS software, release 9.4 (Cary, USA), and Meta-DiSc software, release 1.4. First, the true-positive, true-negative, false-positive and false-negative
values were calculated using SAS software. To explain these factors according to the present meta-analysis, the two assessments of auditory-perceptual judgement of breathiness and ABI are further described. A voice sample was considered without breathiness when a rater panel scored a voice sample with a mean breathiness result <0.5 (negative). A breathy voice was considered as mean breathy ≥0.50 (positive). Abnormal breathiness degree ratings ranged from ≥0.50 to ≤3. The higher the perceived breathiness score, the more severe the degree of breathiness. ABI's threshold depends on the findings of the inter-language differences from the potential included studies for the meta-analysis. The calculated positive and negative values were determined as follows and as for the original Dutch voice samples. If the breathiness judgement was >0.5 and the ABI score was greater than 3.44, the result was classified as true positive. If the breathiness judgement was <0.5 and the ABI score was less than 3.44, the result was deemed true negative. If the breathiness judgement was less than 0.5 and the ABI value was greater than 3.44, the result was considered false positive. If the breathiness judgement was >0.5 and the ABI value was less than 3.44, the result was regarded as false negative.

Second, the results of positive and negative values of all included studies were transmitted to the program Meta-DiSc for further analysis. This program analysed the sensitivity and specificity values for each study and the pooled sensitivity and specificity with their 95% confidence intervals. The results were plotted in forest plots and receiver operating characteristic (ROC) curve. The heterogeneity of studies was calculated using chi-squared heterogeneity test and the I² index. Heterogeneity was confirmed for P-values <0.05. An I² index value between 0% and 25% represents insignificant heterogeneity; >25%-50% low heterogeneity; >50%-75% moderate heterogeneity; and >75% high heterogeneity. For the pooling method, the random-effects model was used. In case of heterogeneity between the studies, this model takes this factor into account. To assess the diagnostic accuracy of the pooled analysis, the summary receiver operating characteristic (SROC) curve with its area under the curve (AUC) and index Q* was conducted. A model with poor performances in diagnostic accuracy has an AUC = 0.5 or lower. When AUC reaches a value of 0.8, there is 80% chance that the model will be able to distinguish between positive values and negative values. The following definition rules for the interpretation of AUC values were used: If AUC = 0.5: no class separation, if 0.7 ≤ AUC ≤0.8: acceptable class separation, if 0.8 ≤ AUC <0.9: excellent class separation, and if AUC ≥ 0.9: outstanding class separation.

The Q* index represents the most ideal point in the ROC curve space equalising ABI's sensitivity and specificity levels which have been presented as close as possible to the top left-hand corner of the ROC diagram. The definitions of Q* index results are comparable with the AUC interpretations.

Third, a weighted threshold of ABI with its sensitivity and specificity was calculated based on the results of all included studies using SAS software. The weighting procedure is based on the Youden Index, which produces the best threshold provided by the maximum of sensitivity + specificity -1.

3 | RESULTS

3.1 | Study characteristics

Figure 1 shows the details of exclusion and inclusion of studies using a flow chart according to the PRISMA statement. We screened a total of 34 unique citations, and 10 full-text articles were reviewed, including 6 studies. Table 1 reports the characteristics of six included studies of ABI's diagnostic accuracy obtained from 3136 subjects with voice disorders and 467 vocally healthy subjects.

Six different languages were investigated of the included studies to assess the ABI's stability and validity across different phonetic structures for Germanic language groups (Dutch and German), Roman language groups (Spanish and Brazilian Portuguese), Atlantic language group (Korean) and Japonic language group (Japanese). The intra-rater reliability of breathiness ratings of the different panels was sufficient in all studies, and the inter-rater reliability reached acceptable agreement in five of six studies. Just the study with Brazilian Portuguese samples reached a low inter-rater reliability level of Fleiss kappa <0.41.

The correlation between ABI scores and auditory-perceptual evaluation of breathiness provides information about the concurrent validity aspect of the ABI, as expressed by the Spearman rank-order correlation coefficient (r_s). The r_s values ranged from 0.746 to 0.890 corresponding to a high correlation between ABI and the auditory-perceptual judgement of breathiness.

The evaluation of the risk of bias of the included studies is shown in Table 2. The features of the included studies were judged to be at low risk of bias, although some confounding was possible for one study of the patient selection according to applicability concerns because the reference standard was proportionally high for vocally healthy voices.

3.2 | Diagnostic accuracy of ABI

Figure 2 shows the forest plots with pooled sensitivity of 0.84 (95% CI: 0.83 to 0.85) and pooled specificity of 0.92 (95% CI: 0.89 to 0.94), as well as heterogeneity statistics. There was a high heterogeneity in sensitivity (I² = 89.7%, P < .001) and a low heterogeneity in specificity (I² = 45.5, P = .103).

Figure 3 reports the SROC curve with excellent values for AUC = 0.94 and Q* index = 0.88.

The weighted threshold for ABI was 3.40 with a sensitivity of 0.86 (95% CI: 0.84 to 0.87) and specificity of 0.90 (95% CI: 0.88 to 0.92).
DISCUSSION

The objective assessment of breathiness with acoustic measurements such as ABI seems to be sufficient compared with auditory-perceptual judgements. The literature search yielded six studies that evaluated the validity of ABI in recent years. For this purpose, the present meta-analysis investigated the diagnostic accuracy of ABI. The pooled sensitivity of the six included studies was acceptable, and the specificity was quite high. Although there was found high heterogeneity in pooled sensitivity but acceptable homogeneity in pooled specificity, the diagnostic accuracy of ABI was sufficient to high. This meta-analysis with more than 3600 included voice samples verifies that the ABI as an objective acoustic measure identifies pathological voices with high diagnostic accuracy. Its pooled sensibility was good at 0.84, and its pooled specificity was very good at 0.92. The high specificity of the ABI guarantees that almost all subjects with absence of breathiness are correctly classified as such. Also, the vast majority, but not all cases, of voice disorders are detected by the ABI. This cannot be expected from a single measure either. For an even higher diagnostic accuracy, further measures are required, optimally as a combination of different dimensions, which also meets the requirements of the protocol for the diagnosis of voice disorders.\textsuperscript{4,6,7,35} Nevertheless, given the low correlation between perceptive assessment of voice quality by the
| Study                        | Language   | N   | P   | T   | Recording equipment                                                                 | No. of judges | Intra-rater reliability | Inter-rater reliability | Concurrent validity | Threshold |
|-----------------------------|------------|-----|-----|-----|-------------------------------------------------------------------------------------|---------------|-------------------------|------------------------|----------------------|-----------|
| Barsties v. Latoszek et al | Dutch      | 88  | 954 | 1042| AKG C420 head-mounted condenser microphone and the Computerized Speech Lab model 4500 | 4             | CK\text{range} = 0.43  | Fk = 0.41              | r_s = 0.840           | 3.44      |
| Delgado Hernández et al    | Spanish    | 47  | 136 | 183 | AKG CS44L head-mounted condenser microphone and the audio interface Alesis IO2 express | 5             | CK\text{mean} = 0.578  | Fk = 0.426              | r_s = 0.826           | 3.40      |
| Hosokawa et al             | Japanese   | 55  | 288 | 343 | Head-worn microphone SE50 (Samson Technologies Corp.) and a linear PCM recorder H4n (Zoom Corp.) | 5             | CK\text{mean} = 0.621  | Fk = 0.492              | r_s = 0.890           | 3.44      |
| Englert et al              | Brazilian Portuguese | 37  | 113 | 150 | AKG C420 head-mounted condenser microphone and Focusrite iTrack Solo               | 5             | CK\text{range} = 0.520  | Fk = 0.353              | r_s = 0.746           | 2.94      |
| Barsties v. Latoszek et al | German     | 43  | 175 | 218 | AKG CS44L head-mounted condenser microphone and Focusrite iTrack Solo soundcard      | 3             | CK\text{mean} = 0.560  | Fk = 0.480              | r_s = 0.850           | 3.42      |
| Kim et al                  | Korean     | 197 | 1470| 1667| AKG Perception 220 microphone and the Computerized Speech Lab model 4500           | 3             | ICC = 0.896             | ICC = 0.825             | r_s = 0.870           | 3.69      |

Abbreviation: CK, Cohen kappa; Fk, Fleiss kappa; ICC, intraclass correlation coefficient; N, number of normal voice samples without diagnosed laryngeal pathology; P, number of pathological or dysphonic voice samples; r_s, Spearman’s correlation coefficient; T, total number of voice samples.
TABLE 2 QUADAS-2 assessment of methodologic quality of included studies

| Study                          | Risk of bias | Applicability concerns |
|-------------------------------|--------------|------------------------|
|                               | Patient      | Index test (ABI)       | Reference standard | Flow and timing |
|                               | selection    |                        |                    |                |
| Barsties v. Latoszek et al^29 | Low          | Low                    | Low                | Low            |
| Delgado Hernández et al^47    | Low          | Low                    | Low                | Low            |
| Hosokawa et al^21             | Low          | Low                    | Low                | Low            |
| Englert et al^68              | Low          | Low                    | Low                | High           |
| Barsties v. Latoszek et al^69 | Low          | Low                    | Low                | Low            |
| Kim et al^1^5                 | Low          | Low                    | Low                | Low            |

FIGURE 2 Forest plots from the meta-analysis of the six included studies of the ABI with pooled sensitivity and specificity next to their 95% confidence intervals, and heterogeneity statistics

FIGURE 3 SROC curve from the meta-analysis of the six included studies of the ABI's diagnostic accuracy results

Symmetric SROC
AUC = 0.9392
SE(AUC) = 0.0087
Q* = 0.8765
SE(Q*) = 0.0110
examiner, self-assessment by the person being examined and single acoustic measures, the high diagnostic accuracy of a single acoustic index is astonishing and gratifying.

Potential confounding factors such as hardware equipment and rater reliability did not appear to significantly influence the results of the acoustic measurements and auditory-perceptual assessments in the six studies analysed. The hardware equipment used followed the standard described by Patel et al for recording voice samples for acoustic voice quality analysis. All but one study had sufficient inter-rater reliability; the Englert et al study showed a lower inter-rater reliability than recommended by Landis and Koch (Fleiss kappa ≥ 0.41). Possibly, the relatively low inter-rater reliability in the Englert et al study also explains the smaller correlation between ABI and the assessments of perceived breathiness, because the study also shows the lowest concurrent validity between both measures. The study included voice samples with 43% normal, 41% mild, 11% moderate and 5% severe breathiness (high applicability concerns of patient selection). In other studies, the levels of breathiness were more balanced, especially for abnormal breathiness. For the validation of ABI as a voice diagnostic instrument, the full spectrum of abnormal breathiness should be represented in a balanced way to allow a clear distinction between normal and abnormal breathiness levels.

The present analysis covered four different language groups (Germanic, Roman, Atlantic and Japonic) with mostly two representatives for each language group. According to the present results, the ABI appears to be relatively robust to phonetic inter-language differences regardless of the language differences in the continuous speech part such as stress timing, syllable timing, mixed rhythm, intonation and complex syllable structure.

Compared to other linguistic studies on the ABI, the Korean study included the largest sample of voices from various laryngeal diseases. As demonstrated in the forest plots of the present meta-analysis, this study also had the largest power. The power of each study, which is included in the meta-analysis, reflects the magnitude or intensity of the parameters of interest in each study. The SROC analysis, which represents the performance of a diagnostic test based on data from a meta-analysis, confirms that the ABI provides highly results to identify breathy voice quality. The Q* index is useful to summarise the accuracy of screening tests and defines test thresholds. In our analysis, both AUC and Q* index indicate the high diagnostic power of the ABI.

The present results confirm that breathiness can be validly quantified with acoustic measures. Overall voice quality can be evaluated by other acoustic indices, such as the cepstral spectral index of dysphonia (CSID) and the acoustic voice quality index (AVQI), which require the recording of continuous speech and sustained vowel phonations. The CSID consists of two multiple regression-based mathematical estimates of the severity of dysphonia, with separate analysis of sustained phonation and continuous speech, using several cepstral and spectral-based measures in a commercial software of PENTAX Medical. It has proven to be a valid instrument for acoustic voice changes associated with voice disorders such as vocal fold paralysis, adductor spasmodic dysphonia, primary muscle tension dysphonia, benign vocal fold lesions, presbylaryngis and mutational falsetto. AVQI is a multivariate model in the freeware program Praat that includes six acoustic parameters demonstrating high validity in the evaluation of overall voice quality in heterogenic voice disorders. The recording and overall signal processing procedure and scale of AVQI are similar to ABI. Unlike CSID and AVQI, which assess the overall voice quality, ABI is particularly suitable as an acoustic measure for the evaluation of breathy voices, for example in case of benign vocal fold lesions, which are dominantly characterised by breathiness such as nodules with medium or large size, paralysis or paresis of the recurrent laryngeal nerve, and vocal fold bowing associated with presbyphonia.

5 | CONCLUSION

In summary, our results confirm the ABI as robust and valid objective measure for evaluating breathiness. Its diagnostic accuracy revealed in six different validation studies high to very high results for sensitivity and specificity. A weighted threshold of ABI to discriminate categorically between breathy and non-breathy voice quality of a subject’s voice was calculated at 3.40.

CONFLICT OF INTEREST

None to declare.

ACKNOWLEDGEMENTS

Open access funding enabled and organized by Projekt DEAL.

ORCID

Ben Barsties v. Latoszkek https://orcid.org/0000-0002-0086-8163
Marina Englert https://orcid.org/0000-0003-0598-8313

REFERENCES

1. Titze IR. Principles of voice production. Englewood Cliff, NJ, USA: Prentice Hall; 1994.
2. Maryn Y. Acoustic measurement of overall voice quality in sustained vowels and continuous speech [doctoral dissertation]. Ghent, Belgium: University of Ghent; 2010.
3. Hirano M. Psycho-acoustic evaluation of voice. In: Arnold GE, Winckel F, Wyke BD, eds. Disorders of human communication 5. Clinical examination of voice. Vienna, Austria: Springer-Verlag; 1981:81-84.
4. Dejonckere PH, Bradley P, Clemente P, et al. A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques. Guideline elaborated by the Committee on Phoniatrics of the European Laryngological Society (ELS). Eur Arch Otorhinolaryngol. 2001;258(2):77-82.
5. Kempster GB, Gerratt BR, Verdolini Abbott K, et al. Consensus auditory-perceptual evaluation of voice: development of a standardized clinical protocol. Am J Speech Lang Pathol. 2009;18(2):124-132.
6. Ziethe A, Patel R, Kunduk M, et al. Clinical analysis methods of voice disorders. Curr Bioinform. 2011;6(3):270-285.
7. Boominathan P, Samuel J, Arunachalam R, et al. Multiparametric voice assessment: Sri Ramachandra University protocol. Indian J Otolaryngol Head Neck Surg. 2014;66(Suppl. 1):246-251.
1. Shrivastav R, Evaluating voice quality. In: Ma EPM, Yiu EML, eds. *Handbook of voice assessments*. San Diego, CA: Singular Publishing Group; 2011:305-318.

2. Fukazawa T, El-Assuwoy A, Honjo I. A new index for evaluation of the turbulent noise in pathological voice. *J Acoust Soc Am*. 1988;83(3):1189-1193.

3. Dejonckere PH, Principal components in voice pathology. *Voice*. 1995;4:96-105.

4. Kramer E, Linder R, Schönweiler R. A study of subharmonics in connected speech material. *J Voice*. 2013;27(1):29-38.

5. Anand S, Kopf LM, Shrivastav R, et al. Objective indices of perceived vocal strain. *J Voice*. 2019;33(6):838-845.

6. Von Leden H, Moore P, Tincke R. Laryngeal vibrations: measurement of the glottis wave. Part 3: the pathologic larynx. *AMA Arch Otolaryngol*. 1960;71(1):16-35.

7. Shrivastav R. Evaluating voice quality. In: Ma EPM, Yiu EML, eds. *Handbook of voice assessments*. San Diego, CA: Singular Publishing Group; 2011:305-318.

8. de Krom G. Some spectral correlates of pathological breathy and rough voice quality for different types of vowel fragments. *J Speech Hear Res*. 1995;38(4):794-811.

9. Van Leden H, Moore P, Tincke R. Laryngeal vibrations: measurement of the glottis wave. Part 3: the pathologic larynx. *AMA Arch Otolaryngol*. 1960;71(1):16-35.

10. Dejonckere PH, Obbens C, de Moor GM, et al. Perceptual evaluation of dysphonia: reliability and relevance. *Folia Phoniatr (Basel)*. 1993;45(2):76-83.

11. Titze IR. Workshop on acoustic voice analysis: Summary statement. Denver, CO: National Center for Voice and Speech; 1995.

12. Kramer E, Linder R, Schönweiler R, A study of subharmonics in connected speech material. *J Voice*. 2013;27(1):29-38.

13. Anand S, Kopf LM, Shrivastav R, et al. Objective indices of perceived vocal strain. *J Voice*. 2019;33(6):838-845.

14. De Bodt MS, Wuyts FL, Van de Heyning PH, et al. Test-retest study of the GRBAS scale: influence of experience and professional background on perceptual rating of voice quality. *J Voice*. 1997;11(1):74-80.

15. Kelchner LN, Brehm SB, Weinrich B, et al. Perceptual evaluation of severe pediatric voice disorders: rater reliability using the consensus auditory perceptual evaluation of voice. *J Voice*. 2010;24(4):441-449.

16. Zalick RI, Kempeast GB, Connor NP, et al. Establishing validity of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V). *Am J Speech Lang Pathol*. 2011;20(1):14-22.

17. Kreiman J, Gerratt BR, Kempeast GB, et al. Perceptual evaluation of voice quality: review, tutorial, and a framework for future research. *J Speech Lang Hear Res*. 1993;36(1):21-40.

18. Gerratt BR, Kreiman J, Antonanzas-Barroso N, et al. Comparing internal and external standards in voice quality judgments. *J Speech Lang Hear Res*. 1993;36(1):14-20.

19. Bele IV. Reliability in perceptual analysis of voice quality. *J Voice*. 2005;19(4):555-573.

20. Roy N, Barkmeier-Kraemer J, Eadie T, et al. Evidence-based clinical voice assessment: a systematic review. *Am J Speech Lang Pathol*. 2013;22(2):212-226.

21. Schönweiler R, Hess M, Wübbelt P, et al. Novel approach to acoustic voice analysis using artificial neural networks. *J Assoc Res Otolaryngol*. 2000;1(4):270-282.

22. Linder R, Albers AE, Hess M, et al. Artificial neural network based classification to screen for dysphonia using psychoacoustic scaling of acoustic voice features. *J Voice*. 2008;22(2):155-163.

23. Godino-Llorente JI, Gómez-Vilda P, Cruz-Roldán F, et al. Pathological likelihood index as a measurement of the degree of voice normality and perceived hoarseness. *J Voice*. 2010;24(6):667-677.

24. Lowell SY, Kelley RT, Awan SN, et al. Spectral- and cepstral-based acoustic features of dysphonic, strained voice quality. *Ann Otol Rhinol Laryngol*. 2012;121(8):539-548.

25. Barsties v. Latoszek BB, Feller-Riesgo CA, et al. Acoustic breathiness index for the Japanese-speaking population: validation study and exploration of affecting factors. *J Speech Lang Hear Res*. 2019;62(8):2617-2631.

26. Barsties v. Latoszek B, Lehnhert B. Internal validation of the acoustic voice quality index version 03.01 and acoustic breathiness index. *Laryngorhinootologie*. 2018;97(9):630-635.

27. Dickerson S, Scherer R, Lefebvre C. Identifying relevant studies for systematic reviews. BMJ. 1994;309(6964):1286-1291.

28. Englert M, Lima L, Behlau M. Acoustic voice quality index and acoustic breathiness index: analysis with different speech material in the Brazilian Portuguese. *J Voice*. In Press.

29. Patel RR, Awan SN, Barkmeier-Kraemer J, et al. Recommended protocols for instrumental assessment of voice: American Speech-Language-Hearing Association Expert Panel to Develop a Protocol for Instrumental Assessment of Vocal Function. *Am J Speech Lang Pathol*. 2018;27(3):887-905.

30. Svec JG, Granqvist S. Guidelines for selecting microphones for human voice production research. *Am J Speech Lang Pathol*. 2010;19(4):356-368.

31. Titze IR, Winholtz WS. Effect of microphone type and placement on voice perturbation measurements. *J Speech Hear Res*. 1993;36(6):1177-1190.

32. Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med*. 2011;155(8):529-536.

33. Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *J Clin Epidemiol*. 2009;62(10):e1-e34.

34. McNenes MDF, Moher D, Thoms BD, et al. Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: the PRISMA-DTA statement. *JAMA*. 2018;319(4):388-396.

35. Zamora J, Abruá V, Muriel A, et al. Meta-DiSc: a software for meta-analysis of test accuracy data. *BMC Med Res Methodol*. 2006;6:31.

36. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ*. 2003;327(7414):557-560.

37. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials*. 1986;7(3):177-188.

38. Portney LG, Watkins MP. *Foundations of Clinical Research, Applications to Practice*, 2nd edn. Upper Saddle River, NJ: Prentice Hall; 2000.

39. Hosmer DW, Lemeshow S. *Applied Logistic Regression*, 2nd edn.: John Wiley & Sons, Inc; 2000:156-164.

40. Walter SD. Properties of the summary receiver operating characteristic (SROC) curve for diagnostic test data. *Stat Med*. 2002;21(9):1237-1256.

41. Schisterman EF, Perkins NJ, Liu A, et al. Optimal cut-point and its corresponding Youden Index to discriminate individuals using pooled blood samples. *Epidemiology*. 2005;16(1):73-81.

42. Delgado Hernández J, León Gómez NM, Jiménez A, et al. Validation of the Acoustic Voice Quality Index Version 03.01 and the Acoustic Breathiness Index in the Spanish language. *Ann Otol Rhinol Laryngol*. 2018;127(5):317-326.

43. Englert MT, Maryn Y, Behlau M, et al. Validation of the AVQI 03.01 and the ABI to the Brazilian Portuguese Language. Int. *Arch Otorhinolaryngol*. 2019;23:106.

44. Barsties v. Latoszek B, Lehnhert B, Janotte B. Validation of the Acoustic Voice Quality Index Version 03.01 and Acoustic Breathiness Index in German. *J Voice*. 2020;34(1):157.e17-157.e25.
51. Kim GH, Barsties v. Latoszek B, Lee YW. Validation of Acoustic Voice Quality Index Version 3.01 and Acoustic Breathiness Index in Korean Population. *J Voice*. In Press.

52. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159-174.

53. Frey LR, Botan CH, Friedman PG, et al. *Investigating Communication: An Introduction to Research Methods*. Englewood Cliffs, NJ: Prentice Hall; 1991.

54. Dejonckere PH, Crevier-Buchman L, Marie JP, et al. Implementation of the European Laryngological Society (ELS) basic protocol for assessing voice treatment effect. *Rev Laryngol Otol Rhinol*. 2003;124(5):279-283.

55. Hosokawa K, Barsties B, Iwahashi T, et al. Validation of the Acoustic Voice Quality Index in the Japanese Language. *J Voice*. 2017;31(2):260.e1-260.e9.

56. Peterson EA, Roy N, Awan SN, Merrill RM, Banks R, Tanner K. Toward validation of the cepstral spectral index of dysphonia (CSID) as an objective treatment outcomes measure. *J Voice*. 2013;27(4):401-410.

How to cite this article: Barsties v. Latoszek B, Kim G-H, Delgado Hernández J, et al. The validity of the Acoustic Breathiness Index in the evaluation of breathy voice quality: A Meta-Analysis. *Clin Otolaryngol*. 2021;46:31-40. [https://doi.org/10.1111/coa.13629](https://doi.org/10.1111/coa.13629)