Perceptual Voice Quality Evaluation Tools: Normophonic, Pathological, and Sung Voices—A Scoping Review

*Amelia Pettirossi, †Lola Terny, *Nicolas Audibert, *,‡Lise Crevier-Buchman, §Véronique Delvaux, †Sophie Fagniart, ¶Bernard Harmegnies, †Kathy Huet, †Myriam Piccaluga, *Reina Remman, †Virginie Roland, and *Claire Pillot-Loiseau, *Paris, ‡Suresnes, France, and †§¶Mons, Belgium

SUMMARY: Objective. The objective of this scoping review is to provide an inventory of tools allowing the perceptual evaluation of voice, whether spoken or sung, and uttered in either a pedagogical or a pathological context. It provides details on these tools in terms of the gender and age of the speakers, the nature of the production context, the corpora collected, the language in which the data collection was conducted, and the most frequently considered attributes of voice.

Methods. Four online databases were searched, following *Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols* (PRISMA) guidelines. The descriptor words "perceptual" or "auditory," "assessment" or "evaluation," "voice" or "voice quality," and "rating scale" were principally taken into consideration.

Results. Our initial search revealed 891 studies, of which 105 sources met the eligibility criteria for data extraction. This scoping review highlights the existence of 19 tools that have a name, as well as 45 unnamed tools, most of which consist of four or six ordinal degrees. Most of these tools were developed to perceptually evaluate pathological adult voices, half of them in English. The GRBAS(I) and CAPE-V protocols dominate, but this inventory also shows a wide variety of tools, perceived tasks, evaluated attributes, and combinations thereof. All these results confirm the multi-dimensional aspect of voice quality, with some attributes that are rarely mentioned but deserve to be highlighted.

Conclusions. The findings of this review allow researchers, clinicians, and educators to understand the various factors and methods that affect the auditory-perceptual evaluations of voice quality. The results may also guide researchers and educators in developing training programs for less-experienced clinicians and educators to improve their confidence in the auditory-perceptual evaluation of voice. New trends for research in the field are pointed out and discussed.

Key Words: Voice quality– Perceptual assessment– Normophonic voice– Pathological voice– Sung voice.

INTRODUCTION

This scoping review aims to gain insight into the various perceptual evaluation tools developed to describe an individual's voice quality. It is the first comprehensive account of the full spectrum of auditory, perceptual evaluation methods, regardless of the evaluation context (pathological or not), its intended purpose (whether supporting voice change, monitoring vocal evolution over time, screening for potential disorders, guiding pedagogical orientation, characterizing specific vocal attributes, or informing identity profiling), the profile of the evaluated individuals (age, language background, cultural context, etc), and the tasks used to elicit speech material (spontaneous speech, singing, reading, producing sentences, words, syllables or logatomes, etc).

Perceptually based voice evaluation can be intuitive, immediate, and global but such informal practices, although sometimes fruitful, are highly dependent on the evaluator's personal experience and training and therefore turns out to be of limited reliability and validity. Various attempts have been made to avoid these limitations by use of devices based upon objective data varying in nature (acoustical, aerodynamical, or even drawn from medical imagery). Yet, auditory evaluation remains very popular. Within the field of speech-language pathology, several organizations, such as ASHA's Special Interest Group for Voice and Voice Disorders (SIG 3), the European Laryngological Society (ELS), and the Union of the European Phoniatricians (UEP), seek to promote the Evidence-Based Practice (EBP) approach and have contributed to several initiatives aimed at promoting evidence-based systematic reviews and developing practice guidelines.[1–6] Recently, a joint ELS-UEP consensus proposed a multidimensional minimal set for voice quality (VQ) evaluation, which includes perceptual assessment as one of its core components.[7]

In the field of vocal pedagogy, several initiatives promoting EBP have emerged, one emphasizing that research in this area is in its infancy.[8,9] However, it seems that voice assessment is not an essential aspect in this area of this definition. Yet, as Kreiman et al[10] emphasize, "*voice quality is fundamentally*

Accepted for publication February 16, 2026.

This work was supported by the investment program "France 2030" launched by the French Government and implemented by the University Paris Cité as part of its program "Initiative d'excellence" IdEx No. ANR-18-IDEX-0001 and by the Partenariats Hubert Curien (PHC) TOURNESOL No. 46241WH and No. 51884UJ.

From the *Laboratoire de Phonétique et Phonologie, UMR 7018 Université Sorbonne Nouvelle and CNRS, Paris, France; †Service de Métrologie et Sciences du Langage, Institut de Recherche en Sciences et Technologies du Langage, Université de Mons, Mons, Belgium; ‡Hôpital Foch: Service de Laryngologie-Phoniatrie, Université Paris Saclay, Suresnes, France; §Service de Métrologie et Sciences du Langage, Institut de Recherche en Sciences et Technologies du Langage & Fond National de La Recherche Scientifique, Université de Mons, Mons, Belgium; and the ¶Institut de Recherche en Sciences et Technologies du Langage, Université de Mons, Centre International de Phonétique Appliquée, Mons, Belgium.

Address correspondence and reprint requests to Amelia Pettirossi, Laboratoire de Phonétique et Phonologie, UMR 7018 Université Sorbonne Nouvelle and CNRS, 4 rue des Irlandais, 75005 Paris, France. E-mail: pettirossi.a@hotmail.fr

Journal of Voice, Vol xx, No xx, pp. xxx–xxx
0892-1997

https://doi.org/10.1016/j.jvoice.2026.02.030

perceptual in nature" (p. 21). Even among clinicians equipped with advanced acoustic or imaging technologies, auditory-perceptual evaluation remains the most common method[11] and is often viewed as the "gold standard."[12]

Ultimately, the challenge lies in retaining the benefits of perceptual expertise while mitigating its limitations. The most widely accepted solution has been to formalize perceptual judgment through structured, standardized procedures that specify conditions of use, rating methods, technical lexicon, quantification procedures, and so on, thereby increasing reliability and mutual intelligibility among specialists. The voice evaluation "tools" resulting from these efforts are precisely the objects of study in this article.

Terminology

It is a long-lasting observation that there is no consensus on the terminology used to describe perceptual assessment of voice.[10] As a result, terms such as "assessment," "scale," and "protocol" are frequently used interchangeably, despite underlying conceptual differences. We will therefore clarify the concepts employed and specify the terminology adopted in this review.

Defining "voice"

Although the meaning of "voice" may seem self-evident, its interpretation varies widely across disciplines. Harmegnies[13] proposed distinguishing three conceptions of the term. In the first, "voice" refers narrowly to the quasi-periodic sound generated by airflow through the vibrating vocal folds, the laryngeal tone, therefore electively focusing isolating on the larynx and neglecting the role of supraglottal structures. This view dominated medical and phonetic traditions due to the frequency and severity of laryngeal pathologies and the availability of easily accessible precise measurement techniques. However, this reductionist perspective restricts the study of voice to a limited physiological component.

The second conception, introduced by Abercrombie[14] under the term "voice quality," broadens the focus to include both laryngeal and supralaryngeal contributions. Abercrombie[14] defined voice quality as "*those characteristics which are present more or less all the time that a person is talking*" (p. 91) and Laver[15] described it as "*the characteristic auditory colouring of an individual speaker's voice*" (p. 1), resulting from both intrinsic factors (linked to anatomical structures) and extrinsic factors (under voluntary control). Crystal[16] and Nolan[17] further emphasized the interaction between physiological, linguistic, and paralinguistic determinants. Subsequent developments, including those proposed by Esling et al,[18] have refined and expanded this conceptual framework, making voice quality a robust theoretical and empirical object of study.

The third conception, broader and more eclectic, encompasses diverse reflections on "voice" as a vector of identity, imitation, or expression,[19] but lacks a consistent scientific definition.

Consequently, only the second acceptation, voice as "voice quality," provides a sufficiently precise, holistic, and empirically grounded foundation for rigorous investigation. This study therefore focuses on perceptual tools designed for the evaluation of voice quality.

Note that, so defined, "voice quality" designates the object of the evaluation, that is, the auditory-perceptual characteristics of the voice, whereas "quality of voice" refers to the evaluative process used to determine the degree of excellence (from "good" to "bad").

Naming evaluation

The term "evaluation" refers to the act of assigning a value to a given object; it can cover the widest range of operational aims and procedures, from the most primitive to the most sophisticated, which is why we adopt it here when several others appear in literature. The term "assessment" is sometimes used as a parasynonymous for "evaluation," particularly when several sources of information and/or multiple elementary tools are combined to obtain a multicriteria evaluation of the voice.

Very often, in the pursuit of objectification and intelligibility, evaluation is based on "measurement." In the sense of Allen and Yen,[20] "*measurement is the assigning of numbers to individuals in a systematic way as a means of representing properties of the individuals*" (p. 4). The term "scale" will be restricted to those cases when the numbers used for measurement are taken from a set whose possible values are ordered in a specific sequence corresponding to the magnitude of the property being measured, that is, ordinal values.

The term "protocol" refers to the entire concrete procedure. Decisions such as the choice of the speech material to be produced (eg, spontaneous speech, read speech, isolated speech sounds, etc), the level of measurement employed (eg, nominal, ordinal, interval, ratio, etc), and the processing of the collected measurements (eg, averaging—weighted or not—reference to preestablished standards, etc) are all components of an evaluation protocol.

As a general principle, although the voice quality may be perceived globally as a unified entity, it can be considered, also, as a composite phenomenon that can be approached from different points of view. We will call these characteristics "attributes" of the voice such as hoarseness or breathiness and so on. The mathematization of this conception might lead to considering evaluation as the outcome of a polynomial equation in which each attribute is assigned a numerical value, which explains why the expression "voice parameter" is commonly used in a metaphoric sense.

Appreciating the quality of evaluation tools

Measurement theory provides the conceptual foundations for judging the quality of an evaluative tool, and two concepts are central in this regard: "validity" and "reliability." Validity refers to the extent to which an evaluation tool measures what it is intended to measure. "Content validity" concerns how adequately the measured traits are represented by the components of the evaluation device (speech tasks, rating categories, definitions of attributes, etc). It is usually examined through "face validity" (the device appears appropriate to observers) and "logical validity" (the structure and choices underpinning the tool are theoretically sound). "Criterion-related validity" is the extent to which the results of a perceptual scale correlate with those of another, already-validated measure, for example, acoustic indices. "Predictive validity" is the ability of an assessment tool to predict future outcomes such as whether a screening procedure can identify individuals at risk of developing a voice disorder.

Reliability concerns stability and consistency of measurement. Derived from classical true score theory ($X = T + E$), it expresses the idea that an observed score (X) should reflect the "true" measure (ie, a vocal property) (T) with minimal influence from random measurement error (E). In voice evaluation, reliability is assessed through intra-rater reliability (the same evaluator should produce similar ratings when judging the same voice at different times) and inter-rater reliability: different evaluators using the same protocol should arrive at similar (or at least comparable) judgments.

Earlier work and aims of the study

To date, no study has provided an extensive inventory of the existing perceptual evaluation tools or offered an analysis of their methodological features and use in research, vocal education, or clinical practice. Nonetheless, several earlier contributions remain relevant for the present analysis. Kreiman et al[10] study constitutes a key contribution to the field of voice assessment, particularly through the development of a theoretical framework identifying sources of variability in perceptual judgments, along with recommendations for improving rating reliability and enhancing intra- and inter-listener agreement. A concept that is essentially put forward by Kreiman et al[10] for understanding the limitations in the reliability of auditory-perceptual evaluation is that listeners base their perceptual judgments on internalized mental representations that can differ between listeners as different individuals may have acquired different standards through their experience with voices. More recently, Suhail et al[21] presented a comparative table summarizing some advantages and shortcomings of a selection of tools relevant for assessing tracheoesophageal speech (including the GRBAS scale, the Hammarberg scale, the Buffalo Voice Profile, the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V), the Voice Activity and Participation Profile (VAPP), the Vocal Profile Analysis Scheme, and the Visual Sort and Rate method).

Carding et al[22] are certainly interesting in the comparison it offers between three specific voice evaluation tools: GRBAS, VPA, and Buffalo III. Nevertheless, in our view, its main interest lies in the list of criteria it uses for these purposes of comparative evaluation of the three tools. This guide for selecting a perceptual voice quality evaluation scheme includes 14 criteria: terms based on theoretical framework, training prerequisite, applicability to normal voice, abnormality rating, availability of audio tapes for listener training, laryngeal note rating, vocal tract ratings, prosodic features, evidence of intra- and inter-judge reliability, number of parameters, rating range, protocol format, approximate time required for administration, and applicability to voice and singing teachers. Indeed, this list could be used as the basis for an analysis grid allowing assessment of the potential value of any evaluation tool. This could allow practitioners and researchers to quickly form a good idea of the interest of any tool unknown to them but likely to interest them for its suitability for the objectives of care or research that concern them personally.

To our knowledge, there are very few perceptual assessment tools dedicated to the analysis of the singing voice. Henrich et al[23] attempted to develop a common terminology for describing the perceived quality of the lyrical voice in vocal pedagogy, vocal therapy, and musical acoustics, using a listening grid based on consensus terms and illustrative sound examples. This grid focused on the perception of vocal gesture or vocal technique, sound perception, and performance perception. However, it does not constitute a perceptual assessment tool as such. Nevertheless, the study has the merit of offering the result of a joint reflection on a comprehensive set of attributes related to the operatic singing voice.

Given the considerable diversity of perceptual voice quality assessment tools, and the limited awareness of those developed outside strictly clinical contexts (eg, sung esthetics, role selection in spoken or sung performance, and sociophonetic analyses), a comprehensive inventory was needed. This scoping review therefore provides an overview of the perceptual assessment tools developed to describe the quality of an individual's voice quality, offering a descriptive analysis of the instruments currently available across a wide range of clinical and nonclinical situations. Although numerous tools exist, only a few are widely known or routinely used, while many others remain underrecognized despite their potential relevance. This diversity reflects the multiplicity of points of view and access routes to perceptual vocal evaluation, a complexity that practitioners must consider when selecting appropriate tools.

Our research questions are as follows:

1. Which perceptual evaluation tools of voice quality are currently available, and how are they structured in terms of rating scales and assessed attributes?
2. How were these tools applied in studies on voice quality, considering the evaluated voices (singing voices, spoken voices, and pathological voices), the speakers' gender and age, the nature of the stimuli presented to raters, and the language in which the assessments were conducted?
3. Which attributes are most frequently considered across the identified evaluation tools?
4. How reliable and valid are the obtained ratings, considering intra- and inter-rater agreement as well as their link with acoustic measures?

METHODOLOGY

This review followed the methodological framework provided by the PRISMA Extension for Scoping Reviews

(PRISMA-ScR).[24] Scoping reviews are particularly well-suited for mapping the breadth and nature of evidence in a given field, especially when the literature is heterogeneous and not yet amenable to systematic synthesis. They allow for the inclusion of a wide range of study types, the identification of key concepts, gaps in knowledge, and variations in methodologies.

Given these characteristics, a scoping review was considered the most appropriate approach for our objectives. The research question was deliberately broad, encompassing different ways to qualify voice and aiming to compare perceptual tools across multiple dimensions. Moreover, the exploratory nature of this review was essential, as preliminary investigation revealed an unexpectedly large diversity of perceptual assessment tools, further confirming the relevance of a scoping methodology over a more restrictive systematic review.

Identification phase

As can be seen, the breadth of the range of possible forms of vocal evaluation is mainly due to the richness of the combinatory of basic choices to be made. This results in a certain instability of the specialized lexicon of voice evaluation. Indeed, depending on the subjective preponderance of one or the other of these choices, the author of an evaluation tool can, for example, name it “scale of evaluation of…,” “protocol of evaluation of…,” or even “assessment tool of…” and so on. Clearly the lexicon is open. This is of course a difficulty for any bibliographic-based scoping process, by definition based on the use of descriptor *words*. This will require, for this scoping review, to implement bibliographic search criteria open to this particular difficulty of the discipline’s lexicon.

The selection of relevant articles for this scoping review was based on a search conducted in databases related to the human and social sciences, as well as medical research. References were exported to Zotero from ProQuest, PsycINFO, PubMed, and ScienceDirect. The keywords “perceptual,” “auditory,” “assessment,” “evaluation,” “voice,” “voice quality,” “rating scale,” and “rating-scale” were used in the search equations. These terms were selected based on their frequency of occurrence in a preliminary set of 27 articles, deemed relevant to the topic, that had previously been identified by the last author via Google Scholar and addressed the perceptual assessment of voice quality.

Details regarding the final search equations, the date of article extraction for each database, and the number of articles identified before and after duplicate removal are presented in Table 1. Efforts were made to maintain consistency across databases; minor variations were introduced only where required due to differences in accepted search syntax. An additional selection criterion was applied exclusively to the ScienceDirect database, as the initial queries returned many articles unrelated to speech or voice. To address this, only articles published in the *Journal of Voice*, the *International Journal of Pediatric Otorhinolaryngology*, and *Parkinsonism & Related Disorders* were retained from this database.

Given the broad scope of the review, particularly due to the diversity of voice qualities under investigation, the possibility of using a search equation specifically targeting the singing voice was explored. However, this approach was ultimately discarded, as it yielded no results in PubMed and no additional articles in PsycINFO or ProQuest beyond those already retrieved with the original equations. Interestingly, ScienceDirect returned a larger number of results despite the inclusion of a narrowing keyword. Yet, upon sorting the results by relevance, no additional articles meeting the review’s inclusion criteria were found among the first 100 entries. For these reasons, a single, unified search strategy was ultimately adopted for the entire review.

The databases were accessed using the institutional account of a University of Mons (Belgium) member, based on the subscriptions available through the university. The search equation was entered manually into each database, and all articles returned by the query were retained for the initial identification phase. In the first screening step, the first author reviewed the titles of all retrieved articles and excluded those that did not meet the predefined inclusion criteria. A second screening phase followed, involving full-text reading of the remaining articles to assess their eligibility in greater depth. Each phase of the selection process is detailed following Table 1. The initial search was conducted in September 2023 and was updated in February and December 2025[1] to ensure the inclusion of the most recent publications.

Inclusion and exclusion criteria

The different stages of article eligibility prior to final inclusion were determined by a set of inclusion and exclusion criteria, established in accordance with the research questions and detailed below.

Inclusion criteria were defined as follows:

- A perceptual voice quality hetero-assessment tool had to be described in the article.
- The tool had to involve the evaluation of at least one strictly vocal perceptual attribute (excluding, for

[1] All the articles identified in September 2023 were present in the databases and were retrieved, including the excluded and included articles, as well as the duplicates (both within and between the databases). After removing duplicates, eight new articles were identified via ProQuest for selection based on their titles and abstracts. Conversely, two articles from the 2023 selection were deleted from ProQuest. Consultation with a ProQuest expert confirmed that these changes could be made when the databases were updated. These two articles had already been excluded from the final selection in September 2023. Nine new articles were identified for selection by title and abstract in PsycINFO. For PubMed, 12 new articles were identified for selection by title and abstract. Additionally, one article that was unavailable in full text in September 2023 was added to the full-text reading (“eligibility” stage). Finally, changes directly linked to the ScienceDirect database made it difficult to update the equation. After contacting ScienceDirect support, the source of these changes could not be determined. One hypothesis is that the University of Mons’ subscription to ScienceDirect has changed.

TABLE 1.
Equation Used per Database With Number of Results Obtained Before and After Removing Duplicates and Dates of Extraction

Database	Equation Used	Results (n)	Without Duplicates (n)	Date
Proquest	Abstract((perceptual OR auditory) AND (assessment OR evaluation) AND (voice OR “voice quality”) AND (“rating scale” OR rating-scale))	87	50	08/12/2025
PsycINFO	(perceptual OR auditory) AND (assessment OR evaluation) AND (voice OR “voice quality”) AND (“rating scale*” OR rating-scale*)	228	137	17/02/2025
PubMed	(perceptual OR auditory) AND (assessment OR evaluation) AND (voice OR “voice quality”) AND (“rating scale*” OR rating-scale*)	199	102	17/02/2025
ScienceDirect	(perceptual OR auditory) AND (assessment OR evaluation) AND (voice OR “voice quality”) AND (“rating scale” OR rating-scale)	377	345	05/09/2023
Total		891	634	

instance, speech intelligibility, vocal attractiveness, or perceived gender).

- Articles that contravened one or more exclusion criteria could nevertheless be retained, provided they reported a tool not described in any other identified publication.

Exclusion criteria were defined as follows:

- No perceptual voice quality self-assessment tool was included in the article.
- The article was a literature review.
- The article was a case study.
- The study involved the use of synthesized voice samples.
- The article focused on the effects of a treatment, surgery, voice therapy, vocal effort, or similar interventions.
- The article used only a partial version of an established perceptual voice quality hetero-assessment tool.

Screening phase

The article screening phase was initiated with a total of 634 articles. At this stage, the data were exported to Zotero, with one subcollection created for each database. To initiate and facilitate the inter-rater agreement phase of the screening process, a labeling procedure was conducted by the first author during a second round of screening. Each article was annotated to predefine several aspects relevant to the research questions, including the perceptual hetero-assessment tool employed, the number and nature of the evaluated attributes (vocal or nonvocal), the potential association between acoustic correlates and perceptual dimensions, the presence or absence of intra- and inter-rater agreement measures, and whether the study involved the evaluation of treatment or voice therapy effects. From that point onward, all abstracts were independently reviewed by the first author in order to eliminate studies that were clearly outside the scope of the research. Most excluded articles focused on auditory hallucinations in the context of psychotic or schizophrenic disorders. Several other articles were discarded due to their lack of relevance to the concept of “voice.” As a result of this step, this initial screening, 147 articles were retained for further analysis.

As a final step in the screening phase, a random sample comprising 20% of the remaining 147 articles was selected to verify inter-rater agreement prior to the eligibility phase and the full-text review. All titles and abstracts of this subsample were independently reviewed by the second author to assess adherence to the inclusion and exclusion criteria. An inter-rater agreement of 100% was then observed.

Eligibility phase

During the eligibility phase, the remaining 147 articles were distributed among all authors for full-text reading to verify compliance with the predefined inclusion and exclusion criteria. To ensure consistency in data extraction and to prevent the omission of relevant information, all articles were analyzed using a standardized reading grid, developed by the first author and validated by the last author (Appendix 1). Additionally, the readings conducted by each author were reviewed by the first author to confirm adherence to the established grid.

As a result of this phase, 42 articles were excluded, as a detailed examination revealed that they still failed to meet one or more eligibility criteria. Ultimately, 105 articles were retained for inclusion in this scoping review.

Final selection (inclusion phase)

The flowchart displays in Figure 1 summarize all information from source identification to inclusion.

RESULTS

The relevant data charted from each evidence source included are presented in accordance with the review questions and objectives.

Inventory of tools and general properties

In this section, we present some general characteristics of the 105 articles included in this review and the tools they include for the perceptual assessment of voice quality. Figure 2 shows the distribution of the number of articles published per year by year, showing that this number increased overall between February 1982 and December 2025

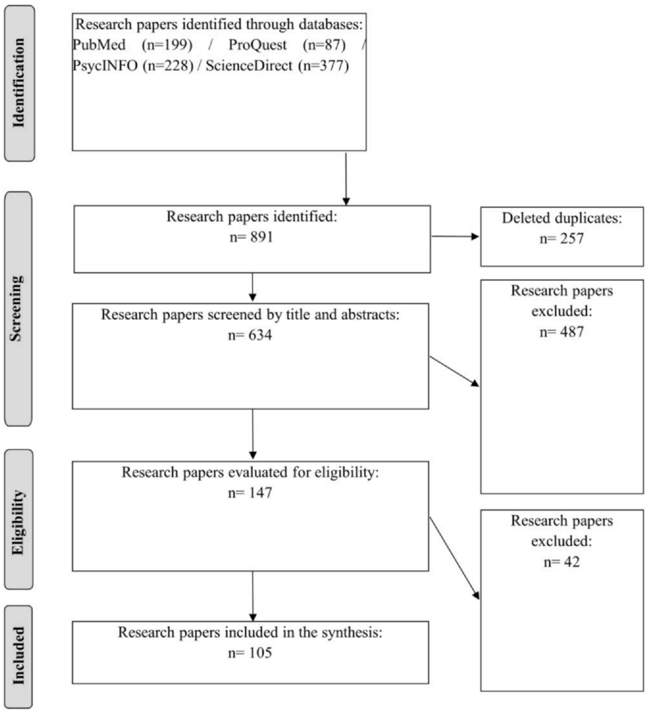

FIGURE 1. PRISMA scoping-review flowchart.

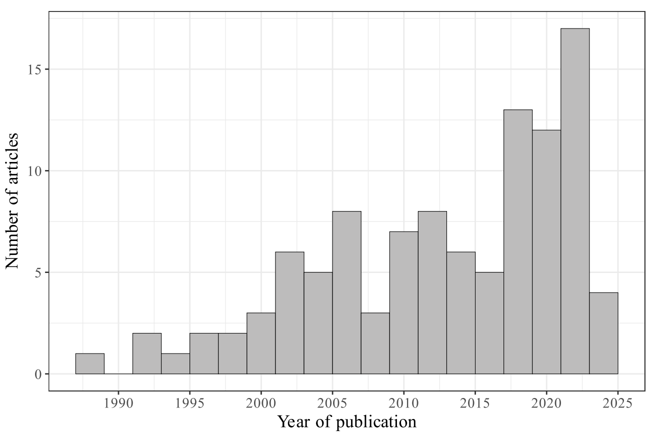

FIGURE 2. Distribution of the number of papers published per year as a function of year (N = 105).

(the February 1982 start date reflects the earliest database record meeting our inclusion criteria, rather than an a priori methodological cutoff, with the first article meeting our inclusion criteria dating from 1988). In particular, the total number of articles doubled between before and after 2015.

Table 2 lists the 64 different tools used in the 105 articles reviewed: acronym (when available) and full name, attributes, scale type, and corresponding reference(s). For clarity and brevity, we refer to tools by their acronym hereafter.

As shown in Figure 3, the most used tools account for one-third of the dataset (36.8%), distributed as follows: CAPE-V (9.4%), GRBAS (22.6%), and GRBASI (4.7%). All the articles concerned, with the exception of one study, focus on pathological voices. Of the remaining 63.2%, the majority of tools do not have a name (ie, 51) and are generally tailored to the needs of a single article, while three publications concern the INFVo protocol, and three others combine speech assessment with voice assessment (VPA, VPCH, and BoDyS). Among these “less used” protocols, one can find most of those used to study normophonic and singing voices (Figure 3).

As indicated in Table 2, more than half of the tools include assessments using ordinal scales (53.8%). One-third use visual analog scales (35.9%), while a minority are nominal (5.1%) or semantic differential (5.1%). Visual analog scales are 100-mm long in 35 of the 37 publications that use them. Ordinal scales comprise 4 to 11 degrees, with the vast majority ranging from 0 to 3 (59.7% of the publications examined), followed by 7 or 11 degrees (11.9% of publications), and 5 degrees (7.5%).

Table 2 and Figure 4 show that the number of attributes used in the tools described in the literature is most often 5. In 30 studies, it ranges from 1 to 4, and in another 30, it ranges from 6 to 12. Finally, a minority of studies (10) use more than 15 and up to 20 different attributes in their perceptual hetero-evaluation of the voice.

Implementation of the tools in the voice studies

In this section, attention is given to how these tools were used for the study of voice quality in the reviewed studies, including the specific types of voice that were evaluated, the gender and age of the speakers, the type of stimuli that were obtained and ultimately submitted to raters, as well as the language in which they were administered. The results are presented comprehensively in Table 3.

As indicated in Table 3, most of the tools concern pathological voices (58.6%), while more than a third of them concern nonpathological voices (38.2%). Finally, a minority of them focus on the perceptual evaluation of singing voices (3.3%). Among the types of pathological voices evaluated, the voice is mainly studied after laryngeal oncosurgery, in cases of voice substitution, muscle tension dysphonia, and other types of dysphonia, in individuals with Friedreich’s ataxia, and in prelingually deaf children, most often in comparison with a control population. Normal voices are those of “control subjects” (undefined), typically children, teenagers and adults, or actors, first-year drama school students, teachers, or heterosexual cross-dressers. The singing voices featured in the selected articles include teenage members of a boys’ choir, solo singers, internationally renowned opera singers in the field of Western lyric singing, carnatic singers, musical theater students, and experts in belting (Table 3).

The age of the speakers is not specified in 17.7% of the studies. About 18.6% the studies include speakers aged between 80 and 90, while 10.7% focus on speakers aged under 18. The average minimum age is 22.2 years (standard deviation: 13.1) and the average maximum age is 62.7 years (standard deviation: 26.3). The vast majority of studies

TABLE 2.
Perceptual Hetero-Evaluation Tools Available in the Literature (Listed in Alphabetical Order by Name): Name, Attributes Used, Type of Scale, and Corresponding Reference

Tool Name	Attributes (s)	Scale Type	Reference (s)
Auditory-Perceptual Ratings Instrument for Operatic Singing Voice	Overall vocal performance, appropriate vibrato, resonance balance (chiaroscuro), ring, pitch accuracy, breath management, evenness throughout the range, strain	1 to 10 ordinal scale	Ahmadi et al, 2022[25]
Bele's protocol	Pitch, pitch range, sonority, variation in loudness, clarity of articulation, overall voice quality, breathiness, hyperfunctional-pressed voice production, hypofunctional-lax voice production, vocal fry-creaky voice production intermittent, vocal fry-creaky voice production at phase endings, roughness-grating, weak phonation at phrase endings, instability of pitch, soft glottal attacks, ringing voice quality	200 mm Visual analog scale for pitch, pitch range, sonority, variation in loudness, clarity of articulation, overall voice quality and 100 mm visual analog scale for all the other attributes	Bele, 2005, 2007[26,27]
BoDyS (Bogenhausen Dysarthria Scales Speech dimensions)	Respiration, pitch/loudness, voice quality, voice stability, articulation, nasal resonance, speaking rate, fluency, prosodic modulation	0 to 4 ordinal scale	Brendel et al, 2013[28]
Buffalo Voice Profile	Laryngeal tone severity rating, High pitch rating, Low pitch rating, Voice too loud, Voice too soft, Hypernasal, Hyponasal, Oral resonance (throatiness), Breath supply, Hypertensive muscles, Hypotensive muscles, Rate too fast, Rate too slow, Speech intelligibility, Overall voice rating	1 to 5 ordinal scale	Webb et al, 2004[29]
CAPE-V (Consensus Auditory-Perceptual Evaluation of Voice)	Severity, Roughness, Breathiness, Strain, Pitch, Loudness	100 mm visual analog scale	Ehrlich et al, 2018;[30] Muss et al, 2010;[31] Joshi et al, 2020;[32] Helou et al, 2010;[33] Duvall et al, 2021;[34] Nagle et al, 2022;[35] Cohen et al, 2019;[36] Walden, 2022;[37] Özcebe et al, 2019;[38] Karnell et al, 2007;[39] Rocha et al, 2024[40]
GRBAS (Grade, Roughness, Breathiness, Asthenia, Strain)	Grade, Roughness, Breathiness, Asthenia, Strain	0 to 3 ordinal scale	Yamaguchi et al, 2003;[41] Gurrado et al, 2018;[42] Ehrlich et al, 2018;[30] Sellars et al, 2009;[43] Moya-Mendez et al, 2020;[44] Webb et al, 2004;[29] Midi et al, 2008;[45] Gurbuzler et al, 2013;[46] Balasubramanium et al, 2019;[47] Maciejewska et al, 2023;[48] Vaz Freitas et al, 2014;[49] Hidaka et al, 2022;[50] Jannetts et al, 2014;[51] Fujimura et al, 2022;[52] Nagy et al, 2020;[53] Fujiki et al, 2023;[54] Wuyts et al, 1999;[55] Pebbili et al, 2021;[56] Rzepakowska et al, 2018;[57] Cohen et al, 2019;[36] Walden, 2022;[37] Özcebe et al, 2019;[38] Karnell et al, 2007;[39] Murry et al, 2004;[58] Benoy et al, 2024;[59] Deborah et al, 2024;[60] Narasimhan et al, 2023[61]
GRBASI (Grade, Roughness, Breathiness, Asthenia, Strain, Instability)	Grade, Roughness, Breathiness, Asthenia, Strain, Instability	0 to 3 ordinal scale	Schindler et al, 2006;[62] Martens et al, 2007;[63] Silva et al, 2012;[64] Baudonck et al, 2011;[65] Moodley et al, 2019[66]
(I)INFVo (Impression of voice quality (I), Impression of intelligibility (I), Unintended additive Noise (N), Fluency (F), and Quality of Voicing (Vo) scale)	Impression, intelligibility, noise, fluency, voicing	0 to 10 ordinal scale	Moerman et al, 2006[67,68]
(I)INFVo-LT (Impression of voice quality, Impression of intelligibility, Unintended additive Noise, Fluency, and Quality of Voicing in Lithuanian (LT))	Impression, intelligibility, noise, fluency, voicing	100 mm visual analog scale	Pribuisis et al, 2022[69]
INFVo (Impression Noise Fluency Voicing)	Impression, intelligibility, noise, fluency, voicing	Visual analog scale from 0 to 10	Mesolella et al, 2023[70]
II EP CAPE-V (II European Portuguese Consensus Auditory-Perceptual Evaluation of Voice)	Severity, Roughness, Breathiness, Strain, Pitch, Loudness	100 mm visual analog scale	de Almeida et al, 2019[71]

TABLE 2 (*Continued*)

Tool Name	Attributes (s)	Scale Type	Reference (s)
PVP (Perceptual Voice Profile)	High pitch, low pitch, monopitch, soft loudness, loud loudness, monoloud, breathy, strained, rough, glottal fry, pitch breaks, phonation breaks, voice arrests, falsetto, tremor, diplophonia	0 to 3 ordinal scale	Tamplin et al, 2014[72]
RASATI (Rouquidão, Aspereza, Soprosidade, Astenia, Tensão, Instabilidade)	Hoarseness (R), level of roughness (A), breathiness (S), asthenia (A), strain (T), instability (I)	0 to 3 ordinal scale	Machado de Machado et al, 2020[73]
RBH (Roughness, Breathiness, Hoarseness)	Roughness, Breathiness, Hoarseness	0 to 10 ordinal scale	Pützer et al, 2017;[74] Schönweiler et al, 2000[75]
SToPS (Sunderland Tracheoesophageal Perceptual Scale)	Overall grade, tonicity: hypertonic-hypotonic OR stenosis, strain, wetness, impairment of volume, impairment of social acceptability, whisper, impairment of intelligibility, stoma noise, impairment of fluency	0 to 3 ordinal scale (Overall grade, strain, wetness, impairment of volume, impairment of social acceptability, whisper, impairment of intelligibility, stoma noise, impairment of fluency)/0 to 5 semantic differential scale/nominal scale (tonicity: hypertonic-hypotonic OR stenosis)	Hurren et al, 2019[76]
SVEA (Swedish Voice Evaluation Approach method)	Intermittent aphonia, breathiness, hyperfunction/strain, hypofunction, vocal fry, hard vocal onset, roughness, vocal scrape, instability, register breaks, diplophonia, pitch	100 mm visual analog scale and 200 mm visual analog scale for ''pitch''	Staffan et al, 2018[77]
VPA (Vocal Profile Analysis Scheme)	Labial (lip rounding, lip spreading, labiodentalization, extensive range, minimized range), Mandibular (close jaw, open jaw, protruded jaw, extensive range, minimized range), Lingual top/blade (advanced, retracted), Lingual body (fronted body, backed body, raised body, lowered body, extensive range, minimized range), Velopharyngeal (nasal, audible nasal escape, denasal), Pharyngeal (constriction), Larynx position (raised, lowered), Phonation type (harshness, whispery, creaky, falsetto, modal voice), Supralaryngeal (tense, lax), Laryngeal (tense, lax), Pitch (high mean, low mean, wide range, narrow range, high variability, low variability), Loudness (high mean, low mean, wide range, narrow range, high variability, low variability), Breath support, Continuity (interrupted), Rate (fast, slow), Rhythmically, Other	Nominal scale	Webb et al, 2004[29]
VPCH (Voice Performance Chart)	Phonation, breathing, articulation, loudness variation, pitch variation, speech rate variation, creation	0 to 4 ordinal scale	Fernandez-Fresard et al, 2021[78]
GSLBI	Grade, Strained, Leaky, Breathy, Irregular	0 to 3 ordinal scale	Saleh et al, 2025[79]
Unnamed#1	Grade, Roughness, Breathiness, Asthenia, Strain, Aphonia, Diplophonia, Staccato, Tremor, Falsetto, Vocal fry	100 mm visual analog scale	Langeveld et al, 2000[80]
Unnamed#2	Breathiness	0 to 10 ordinal scale	Ikuma et al, 2023[81]
Unnamed#3	Breathiness (none-severe), timbre (dark-bright), voice quality (simple/light-hoty/round), overall vocal quality (poor-exceptional), pitch break (yes/no)	145 mm visual analog scale and nominal scale for ''pitch break''	Weinrich et al, 2022[82]
Unnamed#4	Breathiness, hoarseness, glottal fry, glottal attack, diplophonia, intensity, pitch habitual, pitch range, reverse phonation, articulation, stridor	1 to 5 ordinal scale	Smith et al, 1993[83]
Unnamed#5	Breathiness, hypofunction, hyperfunction	100 mm visual analog scale	Södersten et al, 1995[84]
Unnamed#6	Breathiness, roughness	100 mm visual analog scale	Wong et al, 2021;[85] Misono et al, 2012;[86] Eadie et al, 2011;[87] Sauder et al, 2020;[88] Eadie et al, 2010;[89] Nguyen et al, 2023[90]
Unnamed#7	Breathiness, roughness, strain	100 mm visual analog scale	Nguyen et al, 2023[90]
Unnamed#8	Breathiness, roughness, strain, monotone, resonance, hard attack, glottal fry	0 to 5 ordinal scale	Chen et al, 2007[91]
Unnamed#9	Breathiness, roughness, vocal strain, glottal fry, tone onset, tone color, loudness, pitch	100 mm visual analog scale	Madill et al, 2017[92]

TABLE 2 (*Continued*)

Tool Name	Attributes (s)	Scale Type	Reference (s)
Unnamed#10	Completely clear-very hoarse	Semantic differential scale and 100 mm visual analog scale	Kallvik et al, 2015[93]
Unnamed#11	Deviant–normal, unpleasant–pleasant, ugly–beautiful, noise–no noise, monotonous–melodious, expressionless–expressive, weak–powerful, unsteady–steady, jerking–fluent, slow–quick, low–high, deep–shrill, bubbly–not bubbly, breathy–not breathy, rough–not rough, creaky–not creaky, tense–relaxed, dull–clear, hypertonic–not hypertonic, hypotonic–not hypotonic, unintelligible–intelligible	1 to 7 Semantic differential scale	van As et al, 2003[94]
Unnamed#12	Dysphonia severity, intelligibility	1 to 7 ordinal scale	Zhang et al, 2022[95]
Unnamed#13	Global Severity of vocal deviation	100 mm visual analog scale/ 0 to 3 ordinal scale	Contreras-Ruston et al, 2021[96]
Unnamed#14	High pitch-Low pitch, Loud-Soft, Strong-Weak, Smooth-Rough, Pleasant-Unpleasant, Resonant-Shrill, Clear-Hoarse, Unforced-Strained, Soothing-Harsh, Melodious-Raspy, Breathy voice-Full voice, Excessively nasal-Insufficiently nasal, Animated-Monotonous, Steady-Shaky, Young-Old, Slow rate-Rapid rate, I like this voice-I do not like this voice	1 to 9 semantic differential scale	Gelfer, 1993;[97] Gelfer et al, 1988[98]
Unnamed#15	High pitch-Low pitch, Loud-Soft, Strong-Weak, Smooth-Rough, Pleasant-Unpleasant, Resonant-Shrill, Clear-Hoarse, Unforced-Strained, Soothing-Harsh, Melodious-Raspy, Breathy voice-Full voice, Excessively nasal-Insufficiently nasal, Animated-Monotonous, Steady-Shaky, Young-Old, Slow rate-Rapid rate, Masculine-feminine	1 to 7 semantic differential scale	Andrews et al, 1997[99]
Unnamed#16	Hoarseness, speech effort, voice penetration, use of prosody, match of breath, vocal tone, overall intelligibility, overall quality score	1 to 5 ordinal scale and 100 mm visual analog scale for ''overall quality score''	Haderlein et al, 2016[100]
Unnamed#17	Hoarseness/roughness, tremor, breathiness	0 to 3 ordinal scale	Edgar et al, 2001[101]
Unnamed#18	Intelligibility, articulation, intonation, speech rate, speech-breathing coordination, laryngeal resonance, pharyngeal resonance, hyponasality, hypernasality, anterior resonance, posterior resonance, strain, breathiness, roughness, instability, pitch, loudness, overall severity	Visual analog scale	Coelho et al, 2017[102]
Unnamed#19	Intonation accuracy, Freedom throughout vocal range, Appropriate vibrato, Evenness of registration, Resonance/ring, Flexibility, Intensity, Dynamic range, Legato line, Efficient breath management, Color/warmth, Diction	1 to 7 ordinal scale	Wapnick et al, 1997[103]
Unnamed#20	Loudness, pitch stability, pleasant quality	1 to 3 ordinal scale	Raveendran et al, 2022[104]
Unnamed#21	Loudness, vibrato, ring, timbre, focus, nasality, registration breaks (yes or no)	0 to 10 ordinal scale and nominal scale for ''registration breaks''	LeBorgne et al, 2010[105]
Unnamed#22	Musculoskeletal tension, hard glottal attack, glottal fry, restricted tone placement, hoarseness	0 to 3 ordinal scale	Ross et al, 1998[106]
Unnamed#23	Naturalness, overall severity	9 rungs ordinal scale (here called Equal-appearing interval scale [IEA]) vs Direct magnitude estimation scale [DME]	Eadie et al, 2002[107]
Unnamed#24	Onset of vibrato, vibrato rate, vibrato extent	100 mm visual analog scale	Howes et al, 2004[108]
Unnamed#25	Overall severity	100 mm visual analog scale	Nagle et al, 2014[109]
Unnamed#26	Overall severity	1 to 7 ordinal scale (here called Equal-appearing interval scale [IEA]) vs Direct magnitude estimation scale [DME]	Lee et al, 2022[110]
Unnamed#27	Overall severity of voice disorder	0 to 3 ordinal scale/100 mm visual analog scale	Lee et al, 2018[111]
Unnamed#28	Overall severity, breathiness	100 mm visual analog scale/ Visual sort and rate	Kapsner-Smith et al, 2021[112]

TABLE 2 (*Continued*)

Tool Name	Attributes (s)	Scale Type	Reference (s)
Unnamed#29	Overall severity, breathiness, roughness, brokenness, overall fluency	0 to 3 ordinal scale	Chen et al, 2019[113]
Unnamed#30	Overall severity, roughness, breathiness	100 mm visual analog scale	Eadie et al, 2006[114]
Unnamed#31	Overall severity, roughness, breathiness, strain	1 to 10 ordinal scale	Law et al, 2012[115]
Unnamed#32	Overall severity, vocal effort	100 mm visual analog scale	Eadie et al, 2010[116]
Unnamed#33	Overall voice quality	1 to 7 ordinal scale	Barsties et al, 2017[117]
Unnamed#34	Overall voice quality	0 to 10 ordinal scale	Clapham et al, 2015[118]
Unnamed#35	Unnamed: Overall voice quality: voice A is much less dysphonic, less, slightly less, equally, slightly, more dysphonic, much more dysphonic than voice B	Nominal scale	Ghio et al, 2013[119]
Unnamed#36	Overall voice quality, breathiness, roughness, brokenness	100 mm visual analog scale	Cannito et al, 2012[120]
Unnamed#37	Overall voice quality, vocal firmness	0 to 10 ordinal scale	Phadke et al, 2020[121]
Unnamed#38	Projection, breathiness, roughness, strain	1 to 6 ordinal scale	Pinczower et al, 2005[122]
Unnamed#39	Vocal roughness, listening effort	100 mm visual analog scale	Farahani et al, 2023[123]
Unnamed#40	Voice assessment as normal, Voice assessment as hoarse, Voice assessment as breathy, Voice assessment as rough, severity	Nominal scale for ''Voice assessment''/100 mm visual analog scale for ''severity''	Awan et al, 2009[124]
Unnamed#41	Voice assessment as normal, Voice assessment as hoarse, Voice assessment as breathy, Voice assessment as rough, Voice assessment as strained, severity	Nominal scale for ''Voice assessment''/1 to 7 ordinal scale for "severity"	Selby et al, 2003[125]
Unnamed#42	Voice quality, articulation, nasality, prosody, overall intelligibility	0 to 3 ordinal scale	De Bodt et al, 2002[126]
Unnamed#43	Voice severity, voice pleasantness	Ordinal scale (Equal-appearing interval scale [IEA]) vs Direct magnitude estimation scale [DME]	Eadie et al, 2002[127]
Unnamed#44	Voice tremor severity	100 mm visual analog scale	Maryn et al, 2019[128]
Unnamed#45	Voice tremor severity	1 to 7 ordinal scale	Anand et al, 2012[129]

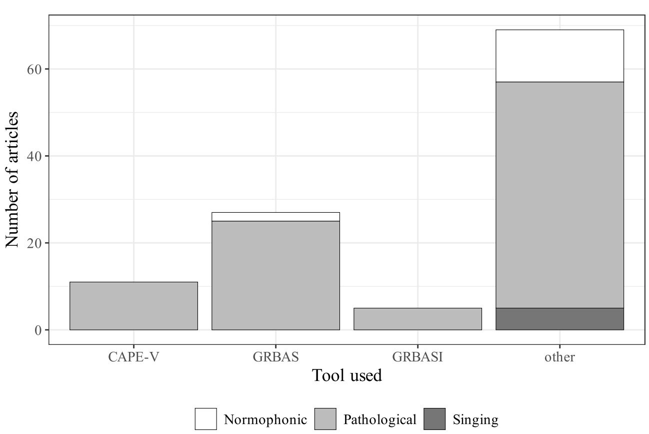

FIGURE 3. Number of papers by tool and voice type (normophonic, pathological, and singing) studied.

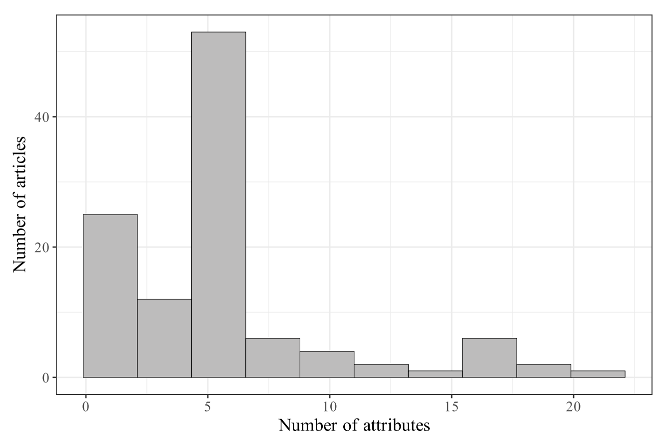

FIGURE 4. Distribution of the number of attributes in the reviewed articles.

therefore concern the evaluation of the voices of a nonelderly adult population. Most studies (79.4%) focus on the perception of male and female voices (8% female only, 8% male only, and the rest unspecified). Table 3 specifies only the total number of subjects, including both men and women.

Table 3 also shows the languages used for voice assessment, indicating that half of them (49%) are carried out in English. The other assessments are conducted in Dutch (9.8%), Chinese (Mandarin or Cantonese, 4.9%), Portuguese (4.9%), Turkish (3.9%), and German (3.9%). Some languages are included but underrepresented: Italian, Hindi, Polish, Japanese, Swedish, Finnish, Spanish (South America), Korean (two papers out of 105), and finally Lithuanian, Iranian, Norwegian, Kannada, Afrikaans, Egyptian Arabic, and French (one paper).

Figure 5 displays the different types of listeners who performed the evaluating tasks and Figure 6 the linguistic material that was used in the tools.

TABLE 3.
Implementation of the Various Tools Ordered by Name: Specific Types of Voice Evaluated, Gender and Age of the Speakers, Raters, Nature of the Stimuli, Language of Administration, and Corresponding References

Tool Name*	Voice Type	Rater (s) (N)	Stimulus (i)	Speakers' Gender (Age Range)	Language (s)	Reference (s)
Auditory-Perceptual Ratings Instrument for Operatic Singing Voice	Pathological voice/singed voice: Traditional Iranian singers with primary muscle tension dysphonia (32)	Singing coach (2)	Sustained vowels, singing a traditional Iranian song "Morh-e Sahar"	Male and female (mean 36.75)	Iranian	Ahmadi et al, 2022[25]
Bele's protocol	Normophonic voice: actors (36) and teachers (35)	SLPs (7) and SLP students (3)	Sustained vowel /a/ at three intensity levels, reading task at two intensity levels	Male (21-73)	Norwegian	Bele, 2005, 2013[26,27]
BoDyS	Pathological voice: Friedreich's ataxia (20)/ Normophonic voice: Controls (10)	SLP used to BoDyS (2)	Answering three questions, repetition task, reading tasks, induced speech	Male and female (15-71)	German	Brendel et al, 2013[28]
Buffalo Voice Profile	Pathological voice: Dysphonics (65)/ Normophonic voice: Controls (5)	SLPs (7)	Reading task, counting from 1 to 10, reciting the days of the week, sustained vowels /a/ and /i/	Male and female (18-87)	English (England)	Webb et al, 2004[29]
CAPE-V	Pathological voice: Dysphonia related to various conditions including Reinke edema, Parkinson's disease, unilateral vocal fold paralysis, and vocal fold lesion (20)/Normophonic voice: Controls (20)	Naive listeners (8)	Sustained vowel /a/, short vowel /a/	Male and female (Unspecified)	English (USA)	Ehrlich et al, 2018[30]
	Pathological voice: Vocal fold nodules (n/a)	SLP familiar with pediatric voices (1)	Unspecified	Male and female (02-18)	English (USA)	Muss et al, 2010[31]
	Pathological voice: Dysphonics (13)/ Normophonic voice: Controls (33)	SLPs with over 20 years' experience (2)	Sustained vowel /a/ and /i/, reading task, spontaneous speech	Male and female (19-78)	Hindi	Joshi et al, 2020[32]
	Pathological voice: Speakers after thyroidectomy (21)	SLPs and ENT specialist (5), naive listeners (5)	Sustained vowel /a/, reading task	Male and female (23-77)	English (USA)	Helou et al, 2010[33]
	Pathological voice: Dysphonics (9)	SLP (1)	Unspecified	Male and female (28-43)	English (USA)	Duvall et al, 2021[34]
	Pathological voice: muscle tension dysphonia, unilateral paralysis, polyp, papilloma, cancer, cyst, amyloidosis, unilateral paresis, nodules, Reinke edema (10)/Normophonic voice: Controls (2)	Voice clinicians (20)	Sustained vowel /a/, reading task	Male and female (21-78)	English (USA)	Nagle et al, 2022[35]
	Pathological voice: Tracheo-esophageal speakers (11)	SLPs (5)	Sustained vowel /a/, reading task, conversational speech	Male and female (5-14)	English (Scotland)	Cohen et al, 2019[36]
	Pathological voice: Unspecified/Normophonic voice: Controls (296 in total)	SLPs used to the GRBAS and the CAPE-V protocols (19)	Sustained vowel /a/ and /i/, reading task	Male and female (14-93)	English (USA)	Walden, 2022[37]
	Pathological voice: Vocal nodules, muscular tension dysphonia, sulcus vocalis, mutational falsetto, polyp, cyst, Reinke edema, unilateral vocal cord paralysis, localized Reinke edema, spasmodic dysphonia (76)/Normophonic voice: Controls (54)	SLPs (2), ENT specialists (2) with over 5 years' experience	Sustained vowel /a/ and /i/, reading task, conversational speech	Male and female (18-69)	Turkish	Özcebe et al, 2019[38]
	Pathological voice: Degenerative disorders (presbylaryngis, bowing), inflammatory conditions (edema, reflux laryngitis), mass changes (nodules, polyps, papilloma), muscle tension dysphonia, neuromotor disorders (paralysis, paresis), and scar (103)	Clinicians (4)	Sustained vowel /a/ and /i/, reading task, answering questions	Male and female (17-90)	English (USA)	Karnell et al, 2007[39]
	Pathological voice: SLA (27)	SLP (1)	Reading six sentences, sustained /a/, /i/, /u/, spontaneous speech	Male and female (mean 60.8)	Portuguese	Rocha et al, 2024[40]

TABLE 3 (*Continued*)

Tool Name*	Voice Type	Rater (s) (N)	Stimulus (i)	Speakers' Gender (Age Range)	Language (s)	Reference (s)
GRBAS	Pathological voice: Unspecified (25)/ Normophonic voice: Controls (4)	Voice experts (26: seven Japanese and 19 Americans)	/u/, /o/, /a/, /e/, /i/	Male and female (16-66)	English (USA)	Yamaguchi et al, 2003[41]
	Pathological voice: Vocal cord paralysis (12)	Phoniatrist, SLP (2)	Unspecified	Male and female (mean 57.1)	Italian	Gurrado et al, 2018[42]
	Pathological voice: Dysphonia related to various conditions including Reinke edema, Parkinson's disease, unilateral vocal fold paralysis, and vocal fold lesion (20)/ Normophonic voice: Controls (20)	Naive listeners (8)	Sustained vowel /a/, short /a/	Male and female (Unspecified)	English (USA)	Ehrlich et al, 2018[30]
	Pathological voice: Asthma (43)	SLPs used to the GRBAS (3)	Spontaneous speech, reading task	Unspecified	English (Scotland)	Sellars et al, 2009[43]
	Pathological voice: Dystonia parkinsonism (32)/ Normophonic voice: Controls (29)	SLPs and ENT specialist (4)	Reading task, sustained vowel /i/	Male and female (Unspecified)	English (USA)	Moya-Mendez et al, 2020[44]
	Pathological voice: Dysphonics (65)/ Normophonic voice: Controls (5)	SLPs (7)	Reading task, counting from 1 to 10, reciting the days of the week, sustained vowels /a/ and /i/	Male and female (18-87)	English (England)	Webb et al, 2004[29]
	Pathological voice: Parkinson's disease (20)/ Normophonic voice: Controls (20)	SLPs and ENT specialist (4)	Unspecified	Male and female (mean 61.6)	Turkish	Midi et al, 2008[45]
	Pathological voice: Fibromyalgia (30)/ Normophonic voice: Controls (30)	Experienced perceptual raters (4)	Reading task	Female (Unspecified)	Turkish	Gurbuzler et al, 2013[46]
	Normophonic voice: Hindu priests considered as voice professionals (22) and controls (22)	SLPs (3)	Spontaneous speech	Male and female (18-30)	Hindi	Balasubramanium et al, 2019[47]
	Pathological voice: Anorexia nervosa (84)/ Normophonic voice: Controls (62)	Phoniatrists with over 10 years' experience (2)	Unspecified	Female (12-19)	Polish	Maciejewska et al, 2023[48]
	Pathological voice: Unspecified (70)/ Normophonic voice: Controls (20)	SLPs with over 6 years' experience and used to the GRBAS (10)	Sustained vowel /a/	Male and female (18 and over)	Portuguese	Vaz Freitas et al, 2014[49]
	Pathological voice: Vocal fold paralysis/paresis, laryngeal cancers, vocal fold polyps, vocal fold nodules, Reinke's edema, muscle tension dysphonia, and spasmodic dysphonia (300)	Experts (8)	Sustained vowel /a/	Male and female (06-96)	Japanese	Hidaka et al, 2022[50]
	Pathological voice: Parkinson's disease and ataxia (53)	Junior researcher (1), senior researcher (1)	Reading task, spontaneous speech	Male and female (46-85)	English (Scotland)	Jannetts et al, 2014[51]
	Pathological voice: Vocal cord polyps, nodules, cysts, laryngitis, atrophy, vocal cord paralysis, vocal cord cancer (945)	Phoniatrists (2), SLP (1)	Sustained vowel /a/	Male and female (Unspecified)	Japanese	Fujimura et al, 2022[52]
	Pathological voice: Chronic hyperfunctional dysphonia (36)	SLP (1)	Sustained vowel /a/, conversational speech	Male and female (5-36)	English (USA)	Nagy et al, 2020[53]
	Pathological voice: Benign vocal fold lesions, nonorganic voice disorders, vocal fold paralysis/paresis, laryngeal cancer/papilloma, neurologic voice disorders or underlying diseases, laryngeal edema or irritation, dysphonia associated with irritable larynx/ laryngospasm (508)	SLP (Unspecified)	Unspecified	Male and female (mean 56.8)	English (USA)	Fujiki et al, 2023[54]
	Pathological voice: Dysphonics (14)	SLPs (25), ENT specialists (4)	Sustained vowel /a/, reading a phonetically balanced text	Male and female (07-65)	Dutch	Wuyts et al, 1999[55]
	Pathological voice: Down's syndrome (17)/ Normophonic voice: Controls (30)	SLPs with over 2 years' experience (3)	Sustained vowel /a/, counting, repetition task	Male and female (4.5-10)	Kannada	Pebbili et al, 2021[56]

TABLE 3 (*Continued*)

Tool Name*	Voice Type	Rater (s) (N)	Stimulus (i)	Speakers' Gender (Age Range)	Language (s)	Reference (s)
	Pathological voice: Benign laryngeal lesions, precancerous conditions of laryngeal mucosa, malignant neoplasms (151)	Laryngologists (2)	Counting from 1 to 20	Male and female (19-81)	Polish	Rzepakowska et al, 2018[57]
	Pathological voice: Patients with complaints of voice disorders, benign vocal cord lesions, muscular tension dysphonia, vocal cord granulomas (50)/Normophonic voice: Controls (45)	SLPs (2)	Conversational speech	Male and female (22-90)	English (USA)	Murry et al, 2004[58]
	Pathological voice: Primary muscle tension dysphonia (MTD-1) (48)	Minimum experience of 5 years after obtaining a post graduate degree in Speech-Language Pathology (3)	Reading of a book in Kannada language, sustained /a/, fricatives /s/ and /z/	Male and female (19-71)	Kannada (India)	Benoy et al, 2024[59]
	Normophonic voice: children and adolescents (457)	SLPs (5 years' experience at least) (3)	Sustained vowel /a/	Boys and girls/5 age groups	Tamil (India)	Deborah et al, 2024[60]
	Pathological voice: dysphonia (15)/ Normophonic voice: Controls (15)	SLP Expert (1)	Sustained vowels /a/, /i/, /u/	Male and female (30-50)	Karnataka (India)	Narasimhan et al, 2023[61]
GRBASI	Pathological voice: Laryngeal surgery (20)	Phoniatrists (2)	Sustained vowel, conversational speech	Male (51-82)	Italian	Schindler et al, 2006[62]
	Pathological voice: Unspecified (70)	SLPs and phoniatrists (6)	Sustained vowel /a/, reading task	Unspecified	Dutch	Martens et al, 2007[63]
	Pathological voice: Parkinson's disease (27) / Normophonic voice: Controls (27)	SLPs (4)	Sustained vowel /a/, counting from 1 to 20, reciting the days of the week	Male (39-82)	Portuguese (Brazil)	Silva et al, 2012[64]
	Pathological voice: Prelingually deaf children using cochlear implant, prelingual severe hearing loss using conventional hearing aids (61)	SLPs (2)	Conversational speech	Male and female (2.6-12)	Dutch	Baudonck et al, 2011[65]
	Pathological voice: Attention deficit hyperactivity disorder (10)/Normophonic voice: Controls (10)	SLPs (Unspecified)	Conversational speech	Male and female (07-09)	Afrikaans	Moodley et al, 2019[66]
(I)INFVo	Pathological voice: Tracheo-esophageal speakers, esophageal speakers, patients with one vocal cord after fronto-lateral laryngectomy (68)	SLP students (24)	Reading task	Unspecified	Dutch	Moerman et al, 2006[67]
		SLPs and phoniatrists (3), SLP students (24)	Reading task	Unspecified	Dutch	Moerman et al, 2006[68]
(I)INFVo-LT	Pathological voice: Laryngeal oncosurgery with substitution voicing (59)/Normophonic voice: Controls (61)	Laryngologists with at least 10 years' experience (4)	Sustained vowel /a/, counting from 1 to 10, reading phonetically balanced Lithuanian sentence	Male (Unspecified)	Lithuanian	Pribuisis et al, 2022[69]
INFVo	Pathological voice: laryngectomees (N = 89) Four groups, depending on the type of substitution voice used, namely, the tracheo-esophageal speakers group (TES, N = 55), the esophageal speakers (ES, N = 13) group, the electrolarynx users (EL, N = 4) group, the no voice patients (NV, N = 17)	Experts in substitution voices with more than 20 years' experience (4)	Sustained vowel /a/, counting from 1 to 10, reading a short text of 100 syllables, word repetition, DDK (eg, pataka)	Male and female (44-81)	Italian	Mesolella et al, 2023[70]
II EP CAPE-V	Pathological voice: Dysphonics (10)/ Normophonic voice: Controls (10)	SLPs with over 5 years' experience (14)	Sustained vowel /a/ and /i/, reading task, spontaneous speech	Male and female (30-61)	Portuguese	de Almeida et al, 2019[71]
PVP	Pathological voice: Tetraplegia (24)	SLP with over 30 years' experience (1)	Sustained vowel, reading task with and without background noise, singing "Happy Birthday"	Male and female (27-70)	English (Australia)	Tamplin et al, 2014[72]
RASATI	Pathological voice: Leishmaniasis (44)	SLPs (4)	Sustained vowel /a/ and /u/	Male and female (mean 48)	Portuguese (Brazil)	de Machado et al, 2020[73]

TABLE 3 (*Continued*)

Tool Name*	Voice Type	Rater (s) (N)	Stimulus (i)	Speakers' Gender (Age Range)	Language (s)	Reference (s)
RBH	Pathological voice: Multiple sclerosis (8)/ Normophonic voice: Controls (16)	Voice quality experts (5), naive listeners (5)	Sustained vowel /i/, /u/ and /a/	Male and female (35-47)	Undetermined	Pützer et al, 2017[74]
	Pathological voice: Hoarse voices from laryngeal mass lesions, incomplete closure, or functional disorders (109)/Normophonic voice: Controls (8)	Clinicians (31)	Sustained vowel /a/	Male and female (11-85)	German	Schönweiler et al, 2000[75]
SToPS	Pathological voice: Tracheo-esophageal speakers (55)	SLPs (12), ENT specialists (10)	Reading task	Male and female (48-80)	English (England)	Hurren et al, 2019[76]
SVEA	Pathological voice: Unilateral cleft lip and palate (73)/Normophonic voice: Controls (63)	SLPs (2)	Reading task, induced speech	Male and female (mean 33.5)	Swedish	Staffan et al, 2018[77]
VPA	Pathological voice: Dysphonics (65)/ Normophonic voice: Controls (6)	SLPs (7)	Reading task, counting from 1 to 10, reciting the days of the week, sustained vowel /a/ and /i/	Male and female (18-87)	English (England)	Webb et al, 2004[29]
GSLBI	Pathological voice: Dysphonics (50)/ Normophonic voice: Controls (80)	Phoniatricians (2)	Sustained /a i u/; Five colloquial Egyptian Arabic sentences; 20 s of spontaneous speech	Male and female (patients: 56% female; controls: 68.87%-female)	Egyptian Arabic	Saleh et al, 2025[79]
VPCH	Normophonic voice: First-year student at Drama School (20)	SLP (1), vocal coaches (4)	Reading task	Male and female (18-27)	Spanish (South America)	Fernandez-Fresard et al, 2021[78]
Unnamed#2	Pathological voice: Vocal cord paralysis, hypofunctional dysphonia, hypotonic dysphonia (123)	SLPs (4)	Sustained vowel /a/, /u/ and /i/ at four pitch profiles	Male and female (Unspecified)	German	Ikuma et al, 2023[81]
Unnamed#3	Normophonic voice: Teenagers in a Boychoir (28)	Singing voice experts (3)	Sung /a/ at several frequencies, singing the song Star Spangled	Male (08-13)	English (USA)	Weinrich et al, 2022[82]
Unnamed#4	Pathological voice: Laryngotracheal stenosis (8)	Voice expert (1)	Sustained vowel /a/, reading task, repetition task	Male and female (5-16)	English (USA)	Smith et al, 1993[83]
Unnamed#5	Normophonic voice: Middle-aged woman (17)	SLPs with 15 to 20 years' experience (3)	/ba:pa:pa:pa:p/ at three intensity levels, conversational speech	Female (45-61)	Swedish	Södersten et al, 1995[84]
Unnamed#6	Pathological voice: Dysphonics (20)	SLPs students (48)	Reading task	Female (18-69)	Chinese	Wong et al, 2021[85]
	Pathological voice: Hyperfunctional dysphonia, nodules, polyps, gastroesophageal reflux, edema, unilateral paralysis (20)/Normophonic voice: Controls (4)	ENT specialists (15), naive listeners (15)	Reading task	Male and female (Unspecified)	English (USA)	Misono et al, 2012[86]
	Pathological voice: Dysphonia with uni- or bilateral lesions, uni- or bilateral paralysis, dysfunctional dysphonia (26)/Normophonic voice: Controls (4)	Experts (8), naive listeners (20)	Reading task	Male and female (25-84)	English (USA)	Eadie et al, 2011;[87] Sauder et al, 2020[88]
	Pathological voice: Dysphonics (38)/ Normophonic voice: Controls (6)	Naive musician listeners (20), naive non-musician listeners (20)	Sustained vowel /a/, reading task	Male and female (26-77)	English (USA)	Eadie et al, 2010[89]
Unnamed#7	Pathological voice: Muscle tension voice disorders (68)/Normophonic voice: Controls trained to manipulate three vocal attributes implicated in functional voice disorders—false vocal fold constriction, vocal fold mass, and larynx height (9)	SLPs and ENT specialists (3)	Sustained vowel /a/, reading task	Male and female (19-84)	English (Australia)	Nguyen et al, 2023[90]
Unnamed#8	Pathological voice: Unspecified (24)	SLPs with over 5 years' experience (3)	Sustained vowel /a/, reading task	Female (26-56)	Chinese (Mandarin)	Chen et al, 2007[91]
Unnamed#9	Normophonic voice: Controls trained to manipulate three vocal attributes implicated in functional voice disorders—false vocal fold constriction, vocal fold mass, and larynx height (9)	SLPs with over 6 years' experience (3)	Reading task	Male and female (19-36)	English (Australia)	Madill et al, 2017[92]

TABLE 3 (*Continued*)

Tool Name*	Voice Type	Rater (s) (N)	Stimulus (i)	Speakers' Gender (Age Range)	Language (s)	Reference (s)
Unnamed#10	Normophonic voice: Children (207)	SLPs students (8), parents (Unspecified), teachers (22)	Induced speech	Male and female (06.4-09.10)	Finnish	Kallvik et al, 2015[93]
Unnamed#11	Pathological voice: Tracheo-esophageal speakers (40)	SLPs (4), naive listeners (20)	Reading task	Male and female (47-82)	Dutch	van As et al, 2003[94]
Unnamed#12	Pathological voice: Tracheo-esophageal speakers (31)	SLP (1), phoniatrists (2) naive listeners (3)	Sustained vowel /a/, /u/ and /i/, reading task	Male and female (50-79)	Chinese	Zhang et al, 2022[95]
Unnamed#13	Pathological voice: Unspecified/Normophonic voice: Controls (211 in total)	SLPs with over 10 years' experience (3)	Counting from 1 to 10	Male and female (19-60)	Spanish (South America)	Contreras-Ruston et al, 2021[96]
Unnamed#1	Pathological voice: Adductor spasmodic dysphonia (77)	SLPs and ENT specialists (3)	Spontaneous speech	Male and female (19-87)	Dutch	Langeveld et al, 2000[80]
Unnamed#15	Normophonic voice: Heterosexual cross-dressers (11)	Students of speech and hearing sciences (88)	Reading task in male voice style and female voice style	Male (35-70)	English (USA)	Andrews et al, 1997[99]
Unnamed#16	Pathological voice: Patients with chronic hoarseness mostly affected by organic dysphonia and functional dysphonia (156)	SLPs (5)	Reading task	Male and female (15-86)	German	Haderlein et al, 2016[100]
Unnamed#17	Pathological voice: Adductor spasmodic dysphonia (18)/Normophonic voice: Controls (10)	Experts (2)	Sustained vowel, reading task	Male and female (33-63)	English (USA)	Edgar et al, 2001[101]
Unnamed#18	Pathological voice: Cochlear implant patients (25)/Normophonic voice: Controls (25)	SLPs (3)	Sustained vowel /a/, counting from 1 to 10, spontaneous speech	Male and female (18-45)	Portuguese (Brazil)	Coelho et al, 2017[102]
Unnamed#20	Normophonic voice: Carnatic singers (16)	SLPs trained in classical singing voice (3)	Sustained vowel /a/, singing the song "aakaram"	Male and female (18-25)	Undetermined	Raveendran et al, 2022[104]
Unnamed#22	Pathological voice: Laryngopharyngeal reflux (49)/Normophonic voice: Controls (20)	SLP with over 15 years' experience (1)	Unspecified	Male and female (Unspecified)	English (USA)	Ross et al, 1998[106]
Unnamed#23	Pathological voice: Tracheo-esophageal speakers (20)	SLPs (20)	Reading task	Male (42-69)	English	Eadie et al, 2002[107]
Unnamed#25	Pathological voice: Unspecified (21)/ Normophonic voice: Controls (4)	Naive listeners (60)	Reading task	Male and female (25-85)	English (USA)	Nagle et al, 2014[109]
Unnamed#26	Pathological voice: Dysphonics (105)	SLP with over 5 years' experience (3)	Reading task	Male and female (mean 57)	Korean	Lee et al, 2022[110]
Unnamed#27	Pathological voice: Patients with vocal complaints (214)	SLPs (3)	Reading task	Male and female (17-79)	Korean	Lee et al, 2018[111]
Unnamed#28	Pathological voice: Unspecified (50)	Naive listeners (22)	Reading task	Male and female (Unspecified)	English (USA)	Kapsner-Smith et al, 2021[112]
Unnamed#29	Pathological voice: Adductor spasmodic dysphonia (20)/Normophonic voice: Controls (20)	SLPs (2)	Sustained vowel /a/, reading task	Male and female (19-59)	Chinese (Mandarin)	Chen et al, 2019[113]
Unnamed#30	Pathological voice: Dysphonics (30)/ Normophonic voice: Controls (6)	SLP students (16)	Sustained vowel /a/, reading task	Male and female (26-81)	English (USA)	Eadie et al, 2006[114]
Unnamed#31	Pathological voice: Unspecified (30)/ Normophonic voice: Controls (10)	SLPs (14)	Sustained vowel /a/, /u/ and /i/, reading task, conversational speech, spontaneous speech	Male and female (24-64)	Chinese (Cantonese)	Law et al, 2012[115]
Unnamed#32	Pathological voice: Dysphonics (20)/ Normophonic voice: Controls (4)	SLPs (10), naive listeners (24)	Reading task	Male and female (25-85)	English (USA)	Eadie et al, 2010[116]
Unnamed#33	Pathological voice: Various organic and nonorganic etiologies and various degrees in voice quality severity without being aphonic (100)	SLPs (5)	Sustained vowel /a/, reading task	Male and female (mean 38)	Dutch	Barsties et al, 2017[117]
Unnamed#34	Pathological voice: Tracheo-esophageal speakers (87)	SLPs (2)	Sustained vowel /a/	Male and female (38-85)	Dutch	Clapham et al, 2015[118]

TABLE 3 (*Continued*)

Tool Name*	Voice Type	Rater (s) (N)	Stimulus (i)	Speakers' Gender (Age Range)	Language (s)	Reference (s)
Unnamed#35	Pathological voice: Dysphonics (53)	SLPs and ENT specialists (7)	Reading task	Male and female (Unspecified)	French	Ghio et al, 2013[119]
Unnamed#37	Normophonic voice: Vocally functional speakers (44)	Voice experts (5)	Sustained vowel /a/, reading task at two intensity levels	Female (mean 42.6)	Finnish	Phadke et al, 2020[121]
Unnamed#38	Normophonic voice: Professional actors (13)	SLP with over 14 years' experience (2)	Interpretation of a passage from William Shakespeare's Julius Caesar: comfortably projected voice vs. maximum projection	Male (26-64)	English (Australia)	Pinczower et al, 2005[122]
Unnamed#39	Pathological voice: Tracheo-esophageal speakers (20)	Naive listeners (20)	Reading task	Male (42-74)	English	Farahani et al, 2023[123]
Unnamed#40	Pathological voice: Unspecified/Normophonic voice: Controls (36 in total)	Naive listeners (40)	Sustained vowel /a/	Female (Unspecified)	English (USA)	Awan et al, 2009[124]
Unnamed#41	Pathological voice: Laryngopharyngeal reflux, laryngeal hyperfunction (13)	SLPs (5)	Sustained vowel /a/, /i/ and /u/, reading task	Male and female (19-68)	English (USA)	Selby et al, 2003[125]
Unnamed#42	Pathological voice: Dysarthria (79)	Voice experts with over 15 years' experience (2)	Reading task, spontaneous speech	Unspecified	Dutch	De Bodt et al, 2002[126]
Unnamed#43	Pathological voice: Dysphonics (24)/ Normophonic voice: Controls (6)	Naive listeners (12)	Reading task	Male and female (23-64)	English	Eadie et al, 2002[127]
Unnamed#44	Pathological voice: Various vocal pathologies with vocal tremor including Parkinson's disease, adductor spasmodic dysphonia, vocal tremor, essential vocal tremor, unilateral vocal cord paralysis (33)	Audiologists (3)	Sustained vowel /a/	Male and female (19-90)	Dutch	Maryn et al, 2019[128]
Unnamed#45	Pathological voice: Essential tremor of the voice (4)	SLPs and ENT specialists with over 20 years' experience (3), naive listeners (3)	Sustained vowel /a/	Male and female (69-76)	English (USA)	Anand et al, 2012[129]
Unnamed#14	Normophonic voice: Women (20)	SLPs (20), naive listeners (20)	Induced speech	Female (19-24)	English (USA)	Gelfer, 1993[97]
	Normophonic voice: Women (20)	SLPs (19), naive listeners (18)	Conversational speech	Female (19-24)	English (USA)	Gelfer et al, 1988[98]
Unnamed#19	Singing voice: Solo singers (17)	Singing teachers (28)	Excerpt from Mozart sung	Unspecified	Undetermined	Wapnick et al, 1997[103]
Unnamed#21	singing voice: Musical theater students, experts in belting (20)	Casting directors (3)	Two vocalizations, six excerpts of Belting	Female (18-25)	English	LeBorgne et al, 2010[105]
Unnamed#24	Singing voice: Internationally renowned opera singers in Western Lyric Singing (11)	Singing teachers (14)	Excerpts from opera songs	Male and female (Unspecified)	English (Australia, USA, England)	Howes et al, 2004[108]
Unnamed#36	Pathological voice: Adductor spasmodic dysphonia (42)	SLPs with over 5 years' experience (6)	Reading task	Male and female (22-79)	English (USA)	Cannito et al, 2012[120]

* For details on the parameters used and evaluated in the tools, see Table 2.

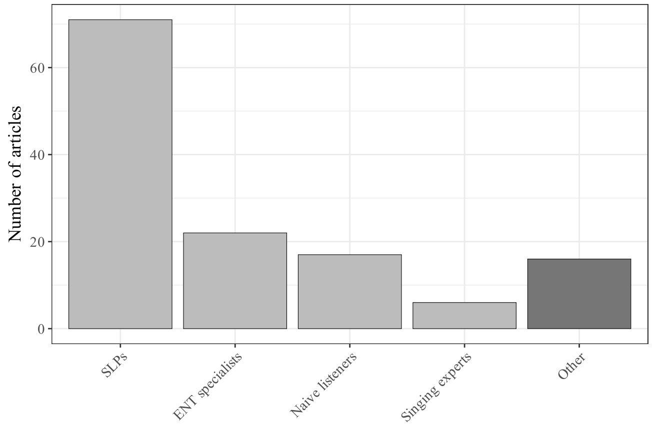

FIGURE 5. Types of listeners/raters.

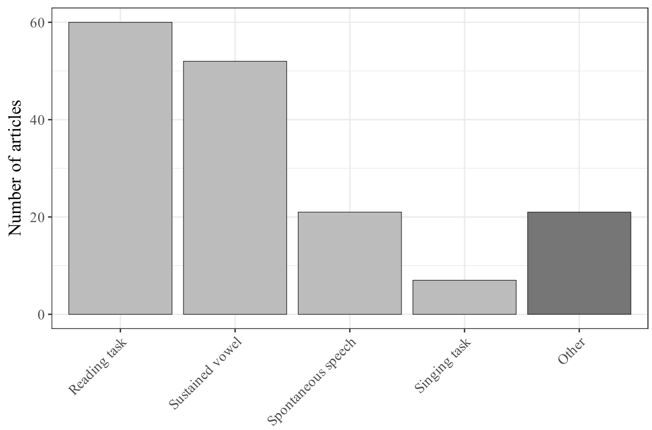

FIGURE 6. Linguistic material used in the tools.

As might be expected, given that nearly 60% of the perceptual scales mentioned in the articles relate to pathological voice, the listeners were mainly speech therapists (in 47.8% of studies) or ENT specialists (15.2%). However, there were also naive listeners (14.5%) and, in articles dealing with the perceptual assessment of the singing voice, singing experts (2.9%). In 19.6% of the studies, other types of listeners were involved, such as students or young researchers, or musicians versus nonmusicians (Table 3 and Figure 5).

The main types of linguistic material used are sustained vowels (35.2% of studies) and read text (33%). These are followed by conversational speech/spontaneous speech (13.6%, both terms being used in the articles) and singing (4.5%). In 13.6% of studies, other tasks are used to perceptually assess the voice, such as counting, repetition, reciting the days of the week, and diadochokinesis (Table 3). The linguistic material was not specified in seven studies. Sustained vowels are associated with reading or conversational speech in half of the articles containing these vowels (Table 3 and Figure 6).

Frequency of attributes appearance

The third objective of this scoping review is to find out the attributes most frequently used in the identified tools.

To better identify these attributes, some were consolidated under the most representative label, as shown in Figure 7. These groupings allowed the identification of 25 distinct categories (Table 4). For example, the label *"instability"* encompasses the attributes *"instability, instability (i), steady–shaky, instability of pitch, pitch break (yes/no), pitch breaks, pitch stability, phonation breaks, continuity: interrupted, unsteady–steady, brokenness, voice arrests, voice stability."* Conversely, "*intermittent aphonia*" was grouped with other attributes such as "*aphonia, phonation, phonation type: modal voice, voicing*" under the broader "*phonation*" label. It should be noted that these groupings inevitably involve a degree of subjectivity, as certain attributes could reasonably fit under multiple categories. The proposed classifications therefore reflect the outcome of extensive discussions and arbitration within the authors' research teams.

Figure 7 and Table 4 also indicate which attributes were used by the literature reviewed in this contribution: Unsurprisingly, attributes from the most frequently used scales such as GRBAS(I) and CAPE-V are mostly used: breathiness for 68 studies, followed by roughness (62), strain, asthenia, and grade. Over 60 studies mention attributes associated with overall quality. Nearly 30 studies use pitch and loudness in the protocols they describe, followed by severity and overall severity. Nearly 10 studies mention three other attributes, such as appropriate vibrato, resonance balance (chiaroscuro), ring, nasal resonance, phonation breaks, tremor, hypofunction, hyperfunction, hard attack, and so on.

Measurement consistency

The fourth objective of the scoping review was to analyze the accuracy and consistency of the measure taken in terms of intra- and inter-agreement. Appendix 2 summarizes 69 studies that conducted intra- and/or inter-rater reliability assessments. It compiles information regarding the reliability of all included perceptual rating scales, specifically: the presence of rater training, the provision of explicit definitions for the attributes being judged, the use of perceptual anchors, and the specific intra- and inter-rater reliability measures reported.

Among the 105 selected studies, 69 reported either intra- or inter-rater reliability: 44 assessed both types of reliability, 19 conducted inter-rater reliability measures only, and seven focused exclusively on intra-rater reliability. Across these studies, 39.1% employed training stimuli, 40.6% provided clear definitions of the perceptual attributes to the raters, but only 15.9% incorporated perceptual anchors.

Regarding intra-rater reliability, the most frequently used statistical procedure was the Intraclass Correlation Coefficient (ICC), reported in 26 studies (52%). This was followed by the Pearson correlation coefficient (10 studies, 20%), the Spearman rank correlation coefficient (five studies, 10%), and percentage of agreement (three studies, 6%). Less commonly used indices included Cronbach's

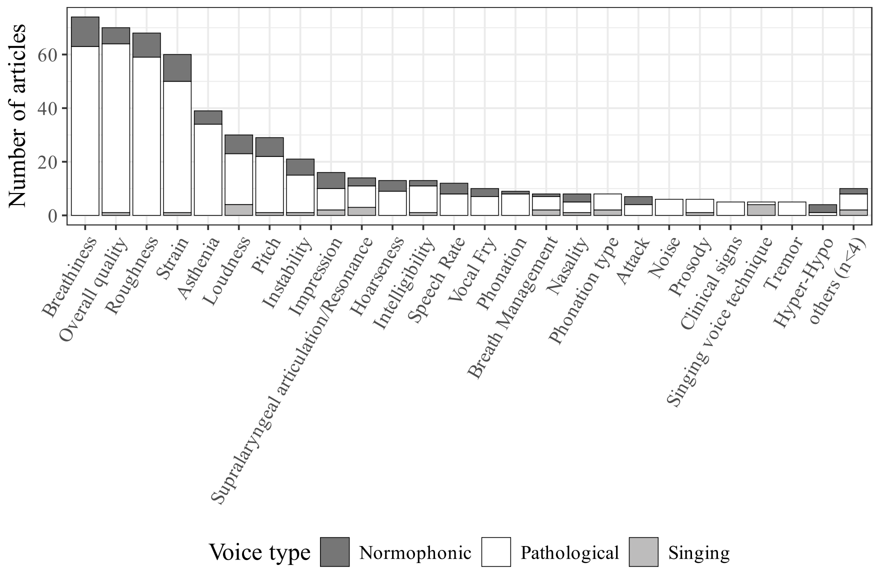

FIGURE 7. Distribution of attributes used in the reviewed studies.

alpha (one study), random-effects ANOVA (1), Conger's kappa (2), and Fleiss' kappa (1).

The reported indices indicate a mean intra-rater reliability of 0.70 (SD = 0.24), with values ranging from 0.05 to 0.99. Notably, 70% of the studies reached an acceptable level of reliability (ie, ≥ 0.70), and 40% achieved the desirable threshold of 0.90 or higher. Among the latter, protocols such as the CAPE-V, GRBAS, the Auditory-Perceptual Ratings Instrument for Operatic Singing Voice, as well as several unnamed tools, were identified.

The intra-judge reliability of the attributes most frequently used in the perceptual scales examined are as follows: 0.80 (0.16) for severity, 0.75 (0.19) for breathiness, 0.74 (0.18) for pitch, 0.72 (0.20) for grade, 0.71 (0.26) for roughness, 0.70 (0.27) for loudness, 0.65 (0.24) for asthenia, and 0.65 (0.23) for strain.

Regarding inter-rater reliability, the most frequently used statistical procedure was the Intraclass Correlation Coefficient (ICC), reported in 34 studies (54.8%). This was followed by the Pearson correlation coefficient (seven studies, 11.3%), the Spearman rank correlation coefficient (five studies, 8.1%), Cronbach's alpha (four studies, 6.5%), and Cohen's kappa (three studies, 4.8%). Less commonly used indices included Kendall's tau (two studies, 3.2%), percentage of agreement (two studies, 3.2%), random-effects ANOVA (1), Conger's kappa (1), Krippendorff's alpha (1), and Fleiss' kappa (1).

The reported indices indicate a mean inter-rater reliability of 0.62 (SD = 0.26), with values ranging from 0.00 to 0.99. Notably, 72.6% of the studies reached the acceptable reliability threshold (ie, ≥ 0.70), and 29% achieved the desirable threshold of 0.90 or higher. Protocols such as the CAPE-V, GRBAS, BoDyS, and the Auditory-Perceptual Ratings Instrument for Operatic Singing Voice, as well as several unnamed tools, were among those meeting this level of reliability.

The inter-judge reliability of the attributes most frequently used in the perceptual scales examined are as follows: 0.81 (0.15) for severity, 0.69 (0.07) for intelligibility, 0.67 (0.23) for breathiness, 0.66 (0.24) for grade, 0.64 (0.27) for strain, 0.62 (0.24) for roughness, and 0.61 (0.30) for asthenia.

Notably, to validate perceptual attributes using objective measures, some studies correlated the dimensions of perceptual rating scales with acoustic features. A total of 14 studies reporting acoustic analyses in relation to perceptual voice assessment were identified. These studies include works by Jannetts et al,[51] Fujiki et al,[54] Nagy et al,[53] Ikuma et al,[81] Nguyen et al,[90] Chen et al,[91] Cannito et al,[120] Phadke et al,[121] Pinczower et al,[122] Selby et al,[125] Anand et al,[129] Rocha et al,[40] Mesolella et al,[70] Deborah et al,[60] and Narasimhan et al.[61] Across these studies, the most frequently used acoustic indices were Cepstral Peak Prominence (CPP or CPPS) and Harmonics-to-Noise Ratio (HNR), each appearing in six studies. Jitter and Shimmer were also widely used (five studies each), followed by Noise-to-Harmonics Ratio (NHR) and various spectral or cepstral indices (eg, CSID, LTAS, and L/H ratio) reported in four studies. Measures of fundamental frequency (F0) and amplitude or energy-based metrics (eg, SPL, mean amplitude, and intensity) were found in three studies each. Maximum phonation time was reported in two studies. Less frequently, specific metrics such as modulation depth/frequency for vocal tremor, Dysphonia Severity Index (DSI), and various perturbation measures (eg, RAP, PPQ) were reported in individual studies.

DISCUSSION

Summary of the results: general patterns and specific observations

This study aimed to establish an extensive and comprehensive inventory of the perceptual evaluation tools available for assessing voice quality and to describe their

TABLE 4.
Label of Attributes, Attributes Considered Equivalent, and Number of Attributes Identified per Label

Attributes	Attributes Considered Equivalents	n
Supralaryngeal articulation/Resonance	lingual body: backed body, lingual body: extensive range, lingual body: fronted body, lingual body: lowered body, lingual body: minimized range, lingual body: raised body, lingual top/blade: advanced, lingual top/blade: retracted, low–high, mandibular: close jaw, mandibular: extensive range, mandibular: minimized range, mandibular: open jaw, mandibular: protruded jaw, oral resonance (throatiness), pharyngeal: constriction, pharyngeal resonance, posterior resonance, resonance, resonance balance (chiaroscuro), resonance/ring, timbre, timbre (dark-bright), tone color, vocal tone, anterior resonance, labial: extensive range, labial: labiodentalization, labial: lip rounding, labial: lip spreading, labial: minimized range, articulation, tense–relaxed, speech effort, supralaryngeal: lax, supralaryngeal: tense	36
Overall quality	voice a is equally dysphonic than voice b, voice a is less dysphonic than voice b, voice a is more dysphonic than voice b, voice a is much more dysphonic than voice b, voice a is slightly less dysphonic than voice b, voice a is slightly more dysphonic than voice b, voice severity, voice assessment as normal, deviant–normal, overall voice quality, overall vocal performance, overall vocal quality (poor-exceptional), overall voice quality: voice a is much less dysphonic than voice b, voice quality (simple/light-hoty/round), overall quality score, voice quality, grade, severity, overall severity, dysphonia severity, global severity of vocal deviation, overall severity of voice disorder, overall voice rating, overall grade	24
Pitch	monopitch, low pitch, monotone, monotonous–melodious, musculoskeletal tension, sonority, pitch: high mean, pitch: high variability, pitch: low mean, pitch: low variability, pitch: narrow range, pitch: wide range, pitch accuracy, pitch variation, pitch/loudness, pitch habitual, restricted tone placement, high pitch, high pitch rating, pitch, high pitch-low pitch, pitch range, low pitch rating, pitch habitual	24
Loudness	loudness, loud-soft, intensity, loud loudness, weak–powerful, loudness: high mean; loudness: high variability; loudness: low mean; loudness: low variability; loudness: narrow range; loudness: wide range; loudness variation, monoloud, variation in loudness, dynamic range, evenness throughout the range, freedom throughout vocal range, soft loudness, impairment of volume, voice too loud, voice too soft, voice penetration, loudness: high mean, loudness: high variability, loudness: low mean, loudness: low variability, loudness: narrow range, loudness: wide range, loudness variation	23
Impression	color/warmth, naturalness, deep–shrill, pleasant quality, expressionless–expressive, impairment of social acceptability, impression, animated-monotonous, melodious-raspy, ugly–beautiful, resonant-shrill, soothing-harsh, young-old, vocal firmness, voice pleasantness, weak–powerful, pleasant-unpleasant, I like this voice-I do not like this voice, unpleasant–pleasant, ringing voice quality, masculine-feminine	21
Instability	instability, instability (i), steady-shaky, instability of pitch, pitch break (yes/no), pitch breaks, pitch stability, phonation breaks, continuity: interrupted, unsteady–steady, brokenness, voice arrests, voice stability, irregular	14
Strain	strain, strain (t), strained, unforced-strained, hyperfunctional-pressed voice production, hyperfunction, hyperfunction/strain, hypertensive muscles, hypertonic–not hypertonic, voice assessment as strained, laryngeal: tense, vocal effort, vocal strain, leaky (incomplete glottal closure with muscle strain leading to airflow leakage under tension, resulting in a hyperfunctional voice)	14
Speech rate	speaking rate, speech rate, slow rate-rapid rate, rate: fast, rate: slow, rate too fast, rate too slow, slow–quick, speech rate variation, jerking–fluent, fluency	11
Breath management	breath management, breathing, match of breath, breath supply, breath support, efficient breath management, respiration, reverse phonation, speech-breathing coordination, stridor	10
Nasality	excessively nasal-insufficiently nasal, nasality, nasal resonance, hypernasality, hyponasality, velopharyngeal: audible nasal escape, velopharyngeal: denasal, velopharyngeal: nasal, hypernasal, hyponasal	10
Phonation type	falsetto, bubbly–not bubbly, phonation type: creaky, phonation type: falsetto, phonation type: harshness, evenness of registration, register breaks, registration breaks (yes or no), whisper, phonation type: whispery	10
Singing voice technique	appropriate vibrato, focus, legato line, onset of vibrato, staccato, vibrato, vibrato extent, vibrato rate, flexibility	9
Asthenia	asthenia, asthenia (a), weak phonation at phrase endings, hypofunction, hypofunctional-lax voice production, hypotensive muscles, hypotonic–not hypotonic, laryngeal: lax	8
Phonation	voicing, phonation, phonation type: modal voice, aphonia, intermittent aphonia, laryngeal resonance, laryngeal tone severity rating, voicing	8
Roughness	roughness, roughness-grating, voice assessment as rough, level of roughness (a), rough–not rough, smooth-rough, vocal roughness, rough	8
Breathiness	breathiness, voice assessment as breathy, breathiness (none-severe), breathiness (s), breathy–not breathy, breathy voice-full voice, breathy	7
Intelligibility	intelligibility, overall intelligibility, clarity of articulation, unintelligible–intelligible, diction, impairment of intelligibility, speech intelligibility	7
Attack	glottal attack, hard attack, hard glottal attack, hard vocal onset, soft glottal attacks, tone onset	6
Hoarseness	hoarseness, clear-hoarse, voice assessment as hoarse, completely clear-very hoarse, hoarseness (r), hoarseness/roughness	6
Prosody	rhythmically, prosodic modulation, prosody, intonation, intonation accuracy, use of prosody	6
Vocal fry	glottal fry, vocal fry, vocal fry-creaky voice production at phase endings, vocal fry-creaky voice production intermittent, creaky–not creaky	5
Clinical signs	wetness, diplophonia, vocal scrape, larynx position (raised/lowered)	4
Noise	noise, noise–no noise, stoma noise	3
Hyper-Hypo	tonicity: hypertonic-hypotonic or stenosis, strong-weak	2
Tremor	tremor, voice tremor severity	2

structure in terms of rating scales and assessed attributes. It also examined how these tools were applied in previous studies, considering the specific voice types evaluated, the speakers' gender and age, the nature of the stimuli presented to raters, and the language in which the assessments were conducted. Furthermore, the study sought to determine which attributes were most frequently used across the identified tools and to evaluate the reliability and consistency of the resulting ratings, based on whether and how intra- and inter-rater agreement were computed and relations with acoustic measures were reported.

A composite profile of the typical study included in this scoping review can be drawn. Such studies typically focus on pathological voices in nonelderly adult populations, comprising both male and female speakers. Participants are generally asked to produce a sustained vowel and/or to read a short text in English. The evaluation tools usually involve the assessment of four to six voice attributes, most commonly overall quality, breathiness, roughness, and strain, occasionally supplemented by asthenia, loudness, and/or pitch, using either a four-point ordinal scale or a 100-mm visual analog scale. Overall, these main characteristics are closely tied to the predominance of the GRBASI and CAPE-V protocols, both primarily designed for the assessment of pathological voices. As for the raters, they are usually experts in the field, primarily speech-language pathologists or ENT specialists. The reliability of their judgments is typically assessed through inter- and/or intra-rater reliability measures, which generally reach acceptable levels.

The extensive inventory conducted in this study is also valuable as it brings to light less common but noteworthy features. As an illustration, beyond read speech and sustained vowels, a wide range of tasks are in fact employed to elicit the audio samples evaluated by raters. Among studies on non-singing voices, roughly one in seven relied on conversational or spontaneous speech, while a smaller proportion employed specific tasks such as sentence or word repetition, counting from 1 to 10 or 20, reciting the days of the week, or performing diadochokinetic sequences. Some tasks were entirely idiosyncratic, tailored to a specific research purpose, for example, reading tasks in male and female voice styles designed to assess the voice quality of heterosexual cross-dressers.[99] There is little doubt that the task itself is a factor influencing evaluation outcomes, and that such diversity may complicate comparisons across studies. Moreover, while many studies favor highly controlled material, most commonly the production of a sustained sound, they vary in the specific sounds used (/a/, but also /i/, /u/ and also /s/, /z/) and often lack detail regarding the instructions given to speakers, which may also introduce variability. Finally, when reading tasks were employed, a wide range of texts were used, further contributing to methodological heterogeneity. With regard to singing voice assessments, the specific corpora consisted of sustained vowels, traditional songs or classical repertoire, or vocal exercises.

Among the gaps identified through this scoping review, one notable limitation concerns the limited number of studies focusing on normophonic elderly speakers and children. Yet, the former often serve as control groups for various pathological populations, while the study of both typical and atypical development of stable voice quality in the latter represents a promising avenue for future research. Similarly, although approximately half of the studies focused on English, 20 other languages were represented across the 105 studies reviewed, leaving hundreds of languages unexamined. In the same vein, surprisingly few studies addressed singing voices or the esthetic dimensions of voice quality, despite our efforts to adapt search strategies and inclusion-exclusion criteria to capture as many as possible. Nonetheless, these few studies provided an interesting extension of the attributes to be considered for a comprehensive evaluation of voice quality, including features such as "*freedom throughout vocal range,*" "*efficient breath management,*" and "*color/warmth*"[103] for singing voices, or "*projection*" for voices of acting professionals.[122] It is worth noting that certain esthetic aspects of voice quality are also assessed in studies on pathological voices, as illustrated by the work of van As and colleagues on tracheoesophageal speech, which evaluates dimensions such as "*monotonous–melodious,*" "*deep–shrill,*" and "*bubbly–not bubbly.*"[94] Henrich et al[23] developed a common terminology for describing voice quality in vocal pedagogy, voice therapy, and musical acoustics: the results of this study indicate that not only can sound be perceived (articulation of vowels and consonants, intelligibility, accentuation and phrasing, timbre, pitch, intensity, and dynamic possibilities), but also vocal gestures or techniques (inspiratory and expiratory dynamics, eg, hearing noisy breathing, vibrato, attack and decay transients, laryngeal mechanism used, pitch accuracy, melodic articulation, forward or backward placement, etc). In this publication, the listening sheets provided facilitate the perceptual and verbal description of voice quality in singing. They allow the listener to focus on a given aspect of voice quality and provide voice professionals with a consensus terminology to express their perception of singing voice quality. They can also be used as a tool for vocal pedagogy and ear training.

The considerable diversity of attributes and their combinations, resulting in a wide range of tools (19 named and 45 unnamed out of 105 studies), warrants discussion. Over 250 distinct attribute names were identified and subsequently consolidated into 25 overarching labels, representing considered equivalences. This grouping process was carried out following extensive discussions and arbitration within the authors' research teams, which included numerous voice experts, and should still be viewed as somewhat tentative. Inevitably, it involves a certain degree of subjectivity and ad hoc decisions regarding the definition of categories and the assignment of attributes, as some could reasonably fall under multiple labels. Altogether, these observations highlight the fact that voice quality is a rich, multilayered, and complex construct that can be operationalized in multiple ways depending on the subfield and research objectives, some dimensions being regarded as peripheral by certain scholars but as central by others. Overall, further harmonization efforts are needed within the research community to refine both the conceptualization and the measurement of voice quality.

While the authors of this scoping review do not aim to promote any single auditory-perceptual assessment method, all these observations highlight nevertheless the relevance of offering guidance for clinical and pedagogical practice, particularly by outlining key criteria to consider when selecting an appropriate evaluation tool. As stated by Oates,[12] regardless of the user, choosing an appropriate perceptual evaluation tool requires a thoughtful analysis of its conceptual and methodological foundations. Indeed, clinicians and teachers should examine the theoretical model supporting the tool, the voice quality dimensions it assesses, and the clarity of the operational definitions used to describe those dimensions. Attention should also be given to practical aspects, such as the format of the rating scale, the procedures for voice sampling and recording, the availability of standardized training materials, and the evidence of reliability and validity supporting its use. In this perspective, we recommend the consultation of key publications such as those by Carding et al,[22] Paz et al,[130] Wang et al,[131] and Yousef et al,[132] which provide valuable insights into qualities of perceptual voice assessment tools.

Alternative perceptual assessment models may also help clinicians make informed and context-appropriate decisions. Several "unnamed tools" listed in Table 2, initially tailored to the requirements of particular research or clinical settings, may prove relevant for subsequent investigations.

Quality of the evaluation tools: limitations and directions for improvement

The findings of the present study allowed for the identification of general trends and the emergence of an overall profile. While this approach is informative, it may overemphasize invariance at the expense of capturing specificities, or the heterogeneity observed in certain dimensions of the collected tools. As an illustration, one may consider the assessment of the reliability of the ratings produced through these evaluation protocols. Although most studies report some measure of judgment reliability, only 42 out of 105 systematically provide both intra- and inter-rater reliability indices. Of these, two-thirds reach only "acceptable" levels of agreement, an outcome that can be considered reasonable given the nature of the task and the complex, multidimensional character of the construct.

It should be noted, however that although our analyses indicated that intra- and inter-rater reliability scores generally reached acceptable thresholds, more fine-grained investigations reveal substantial variability. For instance, the recent study by Pommée et al,[133] which aimed to validate the French version of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-Vf), reported good intra-rater reliability for attributes such as overall severity, strain, and pitch, whereas roughness, breathiness, and loudness showed only moderate reliability. By contrast, inter-rater reliability was generally low across most dimensions, with the exception of overall severity. This work merely reinforces previously established conclusions in the field, showing that the reliability of auditory-perceptual voice evaluations is not uniform: it varies depending on the specific dimensions assessed (some being judged more consistently than others) and the nature of the judgment considered (intra-rater vs inter-rater). The exploratory study by Wang et al[131] revealed the existence of the following factors likely to influence the reliability of assessors in the auditory-perceptual assessment of the voice: nature of voice samples, vocal parameters (the authors use this term correctly), characteristics of assessors, nature of assessment scales, nature of listener training, and assessment tasks.

From a reliability standpoint, all existing perceptual assessment systems present certain limitations. As Kent[134] points out, "*virtually any application of perceptual assessment for any communication disorder runs some risk of false or unreliable information*" (p. 19). This variability arises from multiple, interrelated factors, and numerous explanations have been proposed in the literature to account for it.[126] Sources of measurement error include aspects specific to the raters: their experience with patients presenting voice disorders, or with trainee actors or singers, their prior experience in auditory-perceptual voice evaluation, and their type and amount of training received in auditory-perceptual assessment.

Overall, these findings call for a cautious and critical use of such tools, as well as thorough evaluator training. Indeed, reliability could likely be improved by implementing certain methodological precautions, such as the use of training stimuli, the provision of clear definitions of perceptual attributes to raters, and the inclusion of perceptual anchors. Such methods have been shown to significantly enhance the reliability of perceptual voice evaluation, even among naive listeners.[135,136]

Generally speaking, evaluator training constitutes a critical dimension that warrants close attention. In the context of evidence-based practice (EBP), speech-language pathologists, but also other voice professionals are encouraged to actively engage in and reflect upon the EBP decision-making process as part of their routine clinical care. This engagement supports the delivery of high-quality, evidence-informed services to both individuals and communities.

In this regard, Fissel Brannick et al,[137] in their scoping review on clinical evidence in speech-language pathology, discuss the relevance of Practice-Based Evidence and highlight the role of training in reinforcing EBP. They argue that both preservice training programs and continuing education should aim to fulfill three educational objectives: (1) teaching how to align various data collection methods with client goals and intervention strategies, (2) offering opportunities to critically evaluate diverse forms of practice-based evidence, and (3) providing instruction in descriptive, quantitative, and qualitative approaches to aggregate and analyze data from multiple clients.

A term to be questioned: the notion of a "gold standard" in perceptual evaluation

Auditory-perceptual evaluation is frequently referred to as the "gold standard" in the assessment of voice disorders, a designation that has been widely adopted in the clinical literature, notably because perceptual evaluation remains the most commonly used method for documenting voice disorders and aligns with the perceptual nature of voice itself.[11] While this designation suggests both widespread recognition and methodological superiority, it deserves critical examination. On the one hand, this label is largely based on the frequency of use of this method in clinical practice. Its ease of implementation, low cost, and accessibility for clinicians and teachers explain its predominance. However, high prevalence alone cannot justify its status as a methodological benchmark. On the other hand, the designation of "gold standard" implies an exemplary procedural quality, in terms of validity, reliability, and sensitivity. Yet, numerous studies have highlighted the inherent limitations of perceptual evaluation, including subjectivity, inter- and intra-rater variability, and its sensitivity to contextual factors such as evaluator training, task type, or the characteristics of the voice sample being analyzed. The present review further highlights the considerable diversity of tools and methodological choices observed across studies, which contributes to this variability.

In this context, the use of the term “gold standard” appears ambiguous, as it tends to confuse prevalence of use with methodological excellence. It seems more appropriate to consider auditory-perceptual evaluation as a central yet imperfect tool, whose robustness depends heavily on the conditions of its implementation, and whose validity can be strengthened when combined with complementary objective measures, such as acoustic analysis.

Another essential reflection must also be addressed: how can we explain the enduring popularity of this type of evaluation, despite its known limitations in terms of reliability? Several factors contribute to this: the perceptual nature of voice itself,[126] the intuitive understanding of perceptual terms across different audiences (clinicians, patients, and families but also other voice professionals), the simplicity and low cost of implementation, the ease of administration, minimal equipment requirements, its noninvasive nature, and eventually the lack of accessible or robust instrumental alternatives in many clinical[138] and pedagogical contexts. These pragmatic and conceptual advantages help explain why auditory-perceptual evaluation continues to occupy a central place in voice assessment practices.[12]

Moreover, since it often constitutes the first indicator of a vocal disorder, an alteration in voice quality frequently prompts patients to seek professional help. Furthermore, patients, but also apprentices (in the field of spoken or sung vocal performance), often assess the success of treatment based on the perceived improvement in their voice. From this perspective, several authors emphasize that auditory-perceptual measures provide essential information for both clinicians, teachers, patients, and students. For this reason, speech and voice professionals tend to give a privileged place to perceptual assessment, often regarded as more representative of the vocal experience than the more technical instrumental measures.

Toward a multidimensional approach of pathological voice assessment

The pathological voice remained very prevalent in most of the articles examined. Beyond the simple fact that these tools are widely used, a core clinical consideration concerns the clinical utility and relevance of auditory-perceptual measures in voice assessment. The systematic review conducted by Roy et al[6] aimed to identify the body of research evidence supporting the use of voice assessment measures in the clinical assessment of patients with voice disorders. Specifically, the review addressed three key clinical questions: the extent to which a given assessment procedure can determine (1) the presence or absence of a voice disorder, (2) its underlying etiology, and (3) its severity. Taken together, the findings support the potential clinical utility of selected acoustic, laryngeal imaging-based, auditory-perceptual, and aerodynamic measures as effective components of voice assessment tools. Most of the available evidence highlights their ability to detect the presence of voice disorders.

However, the authors draw attention to the fact that integrating many of the identified measures into routine clinical practice will necessitate further refinement and more rigorous validation, particularly to establish their effectiveness not only in detecting the presence of voice disorders but also in accurately characterizing their nature and severity. Moreover, the authors highlight that many articles lacked sufficient data to directly report diagnostic accuracy metrics and consider therefore that future research should clearly present diagnostic indicators such as sensitivity, specificity, and likelihood ratios to strengthen the evidence base.

Although this paper focuses on auditory-perceptual evaluation, the discussed limitations highlight the inadequacy of relying on clinician ratings alone. According to current clinical guidelines, the evaluation of voice quality is inherently multidimensional and requires the assessment of both subjective impressions and objective acoustic or physiological outcomes.

Empirical studies[139] have indeed demonstrated that combining auditory-perceptual evaluation with acoustic analysis leads to higher accuracy in detecting the presence, absence, and severity of voice disorders (compared with using either method in isolation). In contrast, when employed alone, auditory-perceptual evaluation has shown limited accuracy, particularly in reliably identifying the specific features of a voice disorder. Best practice now supports multimodal approaches that integrate subjective and instrumental measures, provided the latter are implemented with methodological rigor.[6,126,140]

It is worth noting that objective measures are also subject to significant inter- and intra-individual variability, making it difficult to establish normative values. They are also influenced by numerous confounding factors, including recording conditions, the hardware and software used, testing protocols, individual characteristics, and the severity of the voice disorder.[141]

Understudied populations

Our analysis revealed a lower proportion of studies focusing on the pediatric population compared with those addressing adults, a disparity that likely reflects the relatively recent emergence of specialized pediatric voice clinics. As emphasized by Cohen et al[142] and Rickert et al,[143] ENT and speech-language pathology (SLP) services have only recently begun to implement such clinics, to align pediatric assessment tools with established adult standards. In line with this trend, Cohen et al[144] proposed a clinical assessment protocol specifying the key measures to be collected: they recommend auditory-perceptual evaluation. Additionally, they reviewed the evidence supporting the four ELS (European Laryngological Society) attributes, highlighting how these can be adapted for use in pediatric populations. These four attributes are perceptual evaluation of voice (as previously mentioned), videostroboscopic examination, evaluation of aerodynamic performance in voice, and acoustic analysis.

Despite these recent developments, pediatric voice disorders remain insufficiently addressed in both research and clinical practice. A study by Fujiki et al[145] reported that 6.7% of children aged 4 to 12 currently present with a voice disorder, and 12% have experienced one at some point. Given this relatively high prevalence, the limited availability of validated assessment tools is particularly concerning. Vocal impairments in childhood and adolescence can significantly disrupt communication and contribute to adverse psychosocial outcomes. For example, research by Connor et al[146] indicates that children with dysphonia may be perceived more negatively by peers and adults alike, further exacerbating the social burden associated with these disorders.

Pediatric voice disorders raise important methodological challenges. The scarcity of normative auditory-perceptual data for children's voices substantially restricts the interpretability of dysphonia assessments in pediatric populations. Indeed, research on developmental norms for typical pediatric voice quality remains in its early stages. Krasnodębska et al[147] conducted a prospective cohort study of 93 dysphonic children (including 29 girls) aged 4.5 to 14 years using a Pediatric Voice Handicap Index (pVHI) questionnaire completed by the adult accompanying the child to the appointment. The voice was assessed by the doctor using the GRBAS scale and the Yanagihara scale, but only the overall scores (all attributes combined) were presented. These perceptual scores were correlated with the pVHI results. Munjal et al[148] perceptively assessed the vocal characteristics of 13 boys and seven girls of school age who are actively involved in a basketball team using GRBASI and the Buffalo III profile. A VHI (Voice Handicap Index) completed the analyses. The results revealed that almost half of these children had mild hoarseness (confirmed by the Buffalo III voice screening profile), 20% had asthenia, and 26.7% had tightness. While this research is the first of its kind to focus on voice disorders in school-aged children who are members of a basketball team, highlighting the need to advise these children appropriately in order to take preventive measures and resort to voice therapy if necessary, neither the listeners nor the respondents to this questionnaire are described in this study, nor are the degrees of alteration of the various GRBASI attributes (from 0 to 3), nor the details of the reported alterations of the Buffalo III Profile parameters.

A follow-up publication by Fujiki et al[145] pointed to the limited investigation into the psychometric properties of auditory-perceptual evaluation tools in children, whether they exhibit dysphonia. While instruments such as the CAPE-V have demonstrated robust reliability in adult populations, evidence regarding their reliability and clinical relevance in pediatric settings remains scarce. This highlights the urgent need for targeted validation studies and for the adaptation of existing tools to the specific characteristics of children's voices.

In this context, the recent study by Anand et al[149] represents a noteworthy advance, as it constitutes the first large-scale investigation of perceptual evaluation methods applied to a diverse cohort of children with dysphonia.

CONCLUSION

The aim of this exploratory literature review was to provide an analytical inventory of perceptual evaluation tools of voice quality, whether spoken, sung, or pathological. It provided details on these tools in terms of the gender and age of the speakers, the nature of the stimuli presented, the language in which the evaluation tools were conducted, and the most frequently used attributes.

It highlights the existence of standardized scales, most often ordinal with 4 or 6 degrees, existing mainly for pathological adult voices, and in English for half of them; some were excluded from this research, such as the numerous studies comparing voice quality before and after "treatment": for our part, we focused on voice quality itself. However, there are many other unnamed tools exploring other vocal paradigms and attributes.

While GRBASI and CAPE-V protocols predominate, this research also identified a wide variety of tasks, attributes, and combinations thereof, resulting in a broad range of evaluation tools. All these results confirm the multidimensional aspect of voice quality, with some attributes that are rarely cited but deserve to be highlighted, as we have attempted to do in this scoping review. These results also confirm the need for careful and critical use of these tools, as well as in-depth training for evaluators.

This study also highlights the existence of several understudied populations in perceptual voice quality assessment tools: We therefore recommend producing such tools for childlike spoken and sung voices, but also for voice evaluation in the field of autism spectrum (in children as well as in adults), for which it has been suggested that voice quality attributes could constitute promising diagnostic tools.[150] He et al,[151] as to them, have proposed similar perspectives for Alzheimer's disease. The sociolinguistic aspect of voice quality could also benefit from such evaluation tools, whether to define a community of belonging[19] in terms of gender (or nongender), language and/or associated culture, or a particular esthetic, outside the beaten track of vocal pathology.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix 1. Common reading grid created by the first author and validated by the last author

- Tool used (acronym if available OR full name if available without acronym OR indicate "No name: and list parameters"
- Type of scale
- Parameters (n)
- List of parameters evaluated in each tool (in English)
- Language
- Type of voice (pathological voice AND/OR singing voice AND/OR normal voice)
- Speakers (type)
- Gender of speakers
- Pathology of speakers (0 if not relevant)
- Singing technique of speakers (0 if not relevant)
- Speakers (n)
- Age range (or mean)
- Listeners (type)
- Listeners (n)
- Stimulus(i) used
- Listener training (1 if relevant, 0 if not relevant)
- The parameters indicated are clearly defined to the listeners (1 if relevant, 0 if not relevant)
- Anchor stimuli (1 if relevant, 0 if not relevant)
- Comparison with acoustics (0 if not relevant)
- Inter-rater agreement (0 if not relevant)
- Intra-rater agreement (0 if not relevant)

Appendix 2. Intra-rater reliability, rater training, parameter definition, and perceptual anchoring for the perceptual scales included in this scoping review

If only one coefficient is reported, it refers to the full tool. Italicized text indicates specific experimental conditions; values separated by slashes (/) refer to distinct subconditions. Perceptual parameters: G = Grade, R = Roughness, B = Breathiness, A = Asthenia, S = Strain, I = Instability. Statistical indices: ICC = Intraclass Correlation Coefficient, Pearson r = Pearson correlation coefficient, Spearman ρ = Spearman rank correlation coefficient, Kendall τ = Kendall rank correlation coefficient, Cronbach α = Cronbach's alpha, Cohen's κ = Cohen's kappa, Fleiss' κ = Fleiss' kappa, Conger's κ = Conger's kappa, Krippendorff's α = Krippendorff's alpha, % agreement = percentage of agreement, ANOVA = random-effects analysis of variance.

Tool	Training	Definition	Anchor	Intra-Judge Reliability	Inter-Judge Reliability	Reference
Auditory-Perceptual Ratings Instrument for Operatic Singing Voice	No	No	No	ICC = 0.93	ICC = 0.96	Ahmadi et al, 2022[25]
Bele's protocol	Yes	Yes	No	Cronbach α for the reading task *with high/normal intensity*: Pitch: 0.73/0.49, pitch range: 0.75/0.47, sonority: 0.59/0.45, variation in loudness: 0.86/0.44, clarity of articulation: 0.82/0.41, breathiness: 0.53/, hyperfunctional-pressed voice production: 0.61/, hypofunctional-lax voice production: 0.39/0.4, vocal fry-creaky voice production intermittent: 0.37/0.65, vocal fry-creaky voice production at phase endings: 0.32/, roughness-grating: 0.47/0.41, weak phonation at phrase endings: 0.05, instability of pitch: 0.35, soft glottal attacks: 0.28/0.3, ringing voice quality: 0.79/0.55, overall voice quality: 0.80/0.46	Cronbach α for the reading task *with high/normal intensity*: Pitch: 0.73/0.84, pitch range: 0.75/0.84, sonority: 0.59/0.87, variation in loudness: 0.86/0.88, clarity of articulation: 0.82/0.84, B: 0.53/0.71, hyperfunctional-pressed voice production: 0.61/0.82, hypofunctional-lax voice production: 0.39/0.59, vocal fry-creaky voice production intermittent: 0.37/0.72, vocal fry-creaky voice production at phase endings: 0.32/0.55, roughness-grating: 0.47/0.71, weak phonation at phrase endings: 0.05/0.43, instability of pitch: 0.35/0.24, soft glottal attacks: 0.28/0.61, ringing voice quality: 0.79/0.82, overall voice quality: 0.80/0.88	Bele, 2005, 2013[26,27]
BoDyS	No	No	No	No	Pearson r between 2 raters: 0.91	Brendel et al, 2013[28]

Tool	Training	Definition	Anchor	Intra-Judge Reliability	Inter-Judge Reliability	Reference
Buffalo Voice Profile	Yes	No	No	ICC = Laryngeal tone severity rating: 0.83, High pitch rating: 0.72, Low pitch rating: 0.58, Voice too loud: 0.6, Voice too soft: 0.71, Oral resonance (throatiness): 0.55, Breath supply: 0.69, Hypertensive muscles: 0.63, Hypotensive muscles: 0.31, Rate too fast: 0.15, Rate too slow: 0, Speech intelligibility: 0, Overall voice rating: 0.80	ICC = Laryngeal tone severity rating: 0.77, High pitch rating: 0.52, Low pitch rating: 0.54, Voice too loud: 0.01, Voice too soft: 0.56, Hypernasal: n/a, Hyponasal: n/a, Oral resonance (throatiness): 0.18, Breath supply: 0.57, Hypertensive muscles: 0.43, Hypotensive muscles: 0.37, Rate too fast: 0.29, Rate too slow: 0.25, Speech intelligibility: 0.69, Overall voice rating: 0.79	Webb et al, 2004[29]
CAPE-V	Yes	Yes	Yes	ICC = Severity: 0.67, R: 0.57, B: 0.57, S: 0.51, Pitch: 0.1, Loudness: 0.13	No	Ehrlich et al, 2018[30]
	Yes	No	Yes	ICC *for naive listeners/SLPs and ENT specialists*: Severity: 0.84/0.91, R: 0.8/0.87, B: 0.79/0.87, S: 0.58/0.79, Pitch: 0.61/0.82, Loudness: 0.38/0.72	ICC for *naive listeners/SLPs and ENT specialists*: Severity: 0.53/0.72, R: 0.57/0.64, B: 0.41/0.62, S: 0.27/0.31, Pitch: 0.18/0.36, Loudness: 0.11/00.43	Helou et al, 2010[33]
	No	No	No	No	Pearson r (two raters) for the *control/patient group*: Severity: 0.83/0.93, R: −0.37/0.7, B: 0.92/0.82	Joshi et al, 2020[32]
	No	No	Yes	Spearman ρ = Severity: 0.88 to 0.93	Spearman ρ = Severity: 0.86 to 0.93	Karnell et al, 2007[39]
	No	Yes	No	No	ICC = Overall Severity: 0.83, B: 0.76, R: 0.64, S: 0.7	Nagle et al, 2022[35]
	No	No	No	ICC = All items: 0.913, Severity: 0.943, R: 0.896, B: 0.911, S: 0.908, Pitch: 0.878, Loudness: 0.905	ICC = All items: 0.86, Severity: 0.918, R: 0.789, B: 0.827, S: 0.829, Pitch: 0.856, Loudness: 0.870	Walden, 2022[37]
	Yes	No	Yes	ICC = Overall severity: 0.93 to 0.98, R: 0.89 to 0.98, B: 0.89 to 0.98, S: 0.76 to 0.86, Loudness: 0.88 to 0.99, Pitch: 0.86 to 0.97	ICC = Overall severity: 0.90, R: 0.81, B: 0.84, S: 0.80, Loudness: 0.81, Pitch: 0.88	Özcebe et al, 2019[38]
II EP CAPE-V	No	No	No	ICC = Overall severity: 0.87, R: 0.61, B: 0.87, S: 0.73, Pitch: 0.92, Loudness: 0.69	ICC = Overall severity: 0.96, R: 0.92, B: 0.95, S: 0.84, Pitch: 0.86, Loudness: 0.90	de Almeida et al, 2019[71]
GRBAS	Yes	Yes	Yes	ICC = G: 0.92, R: 0.91, B: 0.8, A: 0.89, S: 0.87	ICC = G: 0.94, R: 0.94, B: 0.8, A: 0.91, S: 0.87	Cohen et al, 2019[36]
	No	No	No	No	ICC = 0.83 (auditory perceptual and acoustic analysis)	Deborah et al, 2024[60]
	Yes	Yes	Yes	ICC = G: 0.67, R: 0.589, B: 0.58, A: 0.503, S: 0.424	No	Ehrlich et al, 2018[30]
	Non	Yes	No	No	Krippendorff's α = G: 0.6, R: 0.48, B: 0.43, A: 0.36, S: −0.01	Fujimura et al, 2022[52]
	No	No	No	Conger's κ = G: 0.8 to 0.09, R: 0.73 to 0.11, B: 0.78 to 0.05, A: 0.53 to 0.2, S: 0.69 to 0.08	Conger's κ = G: 0.67, R: 0.4, B: 0.6, A: 0.34, S: 0.48	Hidaka et al, 2022[50]
	No	Yes	No	Spearman ρ on 50% of the stimuli: G: 0.86; R: 0.82; B: 0.81; A: 0.75	Spearman ρ on 10% of the stimuli: G: 0.9; R: 0.68; B: 0.69; A: 0.78	Jannetts et al, 2014[51]
	No	No	Yes	Spearman ρ: G: 0.83 to 0.91	Spearman ρ: G: 0.80 to 0.89	Karnell et al, 2007[39]
	No	No	No	No	Procedure not specified: *Control/patient* group: All times: 0.99/0.98, G: 0.99/0.91, R: 0.99/0.91, B: 0.99/0.91, A: 0.99/0.93, S: 0.99/0.88	Murry et al, 2004[58]
	No	No	No	Random effects ANOVA models: All items:81.8%, G: 69.6%, R: 56.3%, B: 62.4%, A: 49.6%, S: 68.7%	Random effects ANOVA models: All items: 78.1%, G: 64.7%, R: 45.3%, B: 52.8%, A: 43.4%, S: 54.4%	Sellars et al, 2009[43]
	No	No	No	ICC = G: 0.95, R: 0.94, B: 0.94, A: 0.89, S: 0.88	No	Vaz Freitas et al, 2014[49]
	No	No	No	ICC = All items: 0.89, G: 0.9, R: 0.85, B: 0.88, A: 0.89, S: 0.86	ICC = All items: 0.86, G: 0.91, R: 0.79, B: 0.84, A: 0.84, S: 0.84	Walden, 2022[37]
	Yes	No	No	ICC = G: 0.81, R: 0.79, B: 0.73, A: 0.69, S: 0.73	ICC = G: 0.78, R: 0.68, B: 0.70, A: 0.69, S: 0.48	Webb et al, 2004[29]
	No	No	No	No	Cohen's κ for the *visual analog scale/ordinal scale*: G: 0.22/0.38, R: 0.16/0.23, B: 0.16/0.22, A: 0.23/0.27, S: 0.11/0.13	Wuyts et al, 1999[55]
	Yes	Yes	No	No	ICC = *American/Japanese experts*: G: 0.98/0.96, R: 0.98/0.96, B: 0.98/0.92, A: 0.96/0.82, S: 0.95/0.88	Yamaguchi et al, 2003[41]
(I)INFVo	Yes	Yes	No	No	Pearson r for *SLP students/ SLPs and phoniatrists*: Impression: 0.68/0.85, Intelligibility: 0.68/0.75, Noise: 0.57/0.8, Fluency: 0.67/0.78, Voicing: 0.58/0.76	Moerman et al, 2006[68]

Tool	Training	Definition	Anchor	Intra-Judge Reliability	Inter-Judge Reliability	Reference
(I)INFVo	Yes	Yes	No	No	Kendall τ = Impression: 0.79, Intelligibility: 0.75, Noise: 0.80, Fluency: 0.78, Voicing: 0.76	Moerman et al, 2006[67]
(I)INFVo-LT	Yes	Yes	No	ICC = 0.976	ICC = 0.825	Pribuisis et al, 2022[69]
RASATI	No	No	No	No	Cohen's κ = 0.78	de Machado et al, 2020[73]
SToPS	Yes	No	No	Cohen's κ = Overall grade: 0.78, tonicity: 0.64, strain: 0.74, wetness: 0.67, impairment of volume: 0.72, impairment of social acceptability: 0.75, whisper: 0.69, impairment of intelligibility: 0.68, stoma noise: 0.64, impairment of fluency: 0.68	Cohen's κ = Overall G: 0.70, Tonicity: 0.40, S: 0.61, Wetness: 0.49, Impairment of Volume: 0.56, Impairment of Social Acceptability: 0.68, Whisper: 0.58, Impairment of Intelligibility: 0.57, Stoma Noise: 0.51, Impairment of Fluency: 0.59	Hurren et al, 2019[76]
VPA	Yes	No	No	ICC = Lip rounding/spreading: 0.13, Labial range (extensive/minimised): 0.17, Mandibular jaw (close/open): 0.25, Mandibular range (extensive/ minimised): 0.26, Lingual tip (advanced/retracted): 0.17, Lingual body (fronted/ backed): 0.16, Lingual body (raised/lowered): 0.23, Lingual body range (extensive/ minimised): 0.06, Pharyngeal constriction: 0.48, Larynx position (raised/lowered): 0.59, Phonation type harshness: 0.65, Phonation type whispery: 0.69, Phonation type creaky: 0.62, Supralarangeal (tense/lax): 0.43, Laryngeal (tense/lax): 0.48	ICC = Lip rounding/spreading: 0.02, Labial range (extensive/minimised): 0.21, Mandibular jaw (close/open): 0.22, Mandibular range (extensive/ minimised): 0.22, Lingual tip (advanced/retracted): 0.00, Lingual body (fronted/ backed): 0.11, Lingual body (raised/lowered): 0.29, Lingual body range (extensive/ minimised): 0.12, Pharyngeal constriction: 0.23, Larynx position (raised/lowered): 0.44, Phonation type harshness: 0.47, Phonation type whispery: 0.62, Phonation type creaky: 0.49, Supralarangeal (tense/lax): 0.17, Laryngeal (tense/lax): 0.32	Webb et al, 2004[29]
Unnamed#11	No	No	No	Pearson r for *SLPs/naive listeners*: 0.81/0.6	Cronbach α for *SLPs/naive listeners*: 0.57 to 0.93/0.8 to 0.94	van As et al, 2003[94]
Unnamed#2	No	Yes	No	No	ICC = 0.84	Ikuma et al, 2023[81]
Unnamed#3	No	No	No	No	ICC = B: 0.7227, Timbre: 0.6357, Voice Quality: 0.8723, Overall Vocal Quality: 0.6453	Weinrich et al, 2022[82]
Unnamed#6	No	Yes	No	Percentage of agreement (repetition of 20% of the samples) for *nonmusicians/ musicians*: B: 63.5/73.5, roughness: 64.5/66.5	Percentage of agreement for *nonmusicians/musicians*: B: 44.7/49.1, roughness: 40.9/41.4	Eadie et al, 2010[89]
	Yes, before the second session	Yes	No	Percentage of agreement of the *first/ second session* for naive listeners: R: 53.33/57.58, B: 57.33/60.61 – of the *first/ second session* for ENT specialists: R: 58.89/63.64, B: 61.11/60.61	Percentage of agreement of the *first/ second session* for naive listeners: R: 59.17/57.58, B: 50.83/60.61 – of the *first/ second session* for ENT specialists: R: 55.83/52.65, B: 59.17/62.26	Misono et al, 2012[86]
	No	Yes	No	ICC = B: 0.979, R: 0.525	ICC = B: 0.72, R: 0.6	Sauder et al, 2020[88]
Unnamed#7	No	No	No	ICC = B: 0.948 to 0.738, R: 0.889 to 0.812, S: 0.829 to 0.896	ICC = B: 0.659, R: 0.696, S: 0.691	Nguyen et al, 2023[90]
Unnamed#8	Yes	No	No	No	ICC = 0.85	Chen et al, 2007[91]
Unnamed#9	Yes	Yes	Yes	ICC = Between 0.852 and 0.968	ICC = Breathy: 0.622, Glottal fry: 0.674, Loudness: 0.754, Pitch: 0.483, R: 0.345, Strained: 0.597, Tone color: 0.549, Tone onset: 0.607	Madill et al, 2017[92]
Unnamed#10	Yes	No	No	Spearman ρ = SLPs students: 0.554 to 0.821	Spearman ρ = SLPs students: 0.606 to 0.799	Kallvik et al, 2015[93]
Unnamed#12	No	No	No	ICC for *SLPs and phoniatrists/naive listeners*: 0.75 to 0.77/0.27 to 0.9	ICC for *SLPs and phoniatrists/naive listeners*: 0.63/0.7	Zhang et al, 2022[95]
Unnamed#13	Yes	No	Yes	ICC for the *visual analog scale/ordinal scale*: 0.95 to 0.97/0.89 to 0.94	ICC for the *visual analog scale/ordinal scale*: 0.95/0.92	Contreras-Ruston et al, 2021[96]
Unnamed#1	No	No	No	ICC = G: 0.77, R: 0.36, B: 0.52, A: 0.59, S: 0.72, Aphonia: 0.76, Diplophonia: 0.07, Staccato: 0.72, Tremor: 0.64, Falsetto: 0.43, Vocal fry: 0.69	ICC = G: 0.58, R: 0.30, B: 0.41, A: 0.41, S: 0.52, Aphonia: 0.72, Diplophonia: 0.07, Staccato: 0.56, Tremor: 0.41, Falsetto: 0.30, Vocal fry: 0.65	Langeveld et al, 2000[80]

Tool	Training	Definition	Anchor	Intra-Judge Reliability	Inter-Judge Reliability	Reference
Unnamed#14	No	No	No	No	Kendall τ for SLPs/naïve listeners: High pitch-Low pitch: 0. 69/0.57, Loud-Soft: 0.4/0.5, Strong-Weak: 0.47/0.5, Smooth-Rough: 0.5/0.36, Pleasant-Unpleasant: 0.21/0.17, Resonant-Shrill: 0.36/0.45, Clear-Hoarse: 0.17/0.14, Unforced-Strained: 0.39/0.35, Soothing-Harsh: 0.35/0.26, Melodious-Raspy: 0.29/0.26, Breathy voice-Full voice: 0.28/0.23, Excessively nasal-Insufficiently nasal: 0.25/0.22, Animated-Monotonous: 0.21/0.27, Steady-Shaky: 0.3/0.19, Young-Old: 0.28/0.22, Slow rate-Rapid rate: 0.27/0.26, I like this voice-I do not like this voice: 0.34/0.31	Gelfer et al, 1988[98]
Unnamed#16	No	No	No	No	Pearson r = Overall voice quality = 0.87	Haderlein et al, 2016[100]
Unnamed#18	Yes	Yes	No	ICC for *sustained vowel/counting*: Laryngeal resonance: 0.323 to 0.926/0.55 to 0.74, pharyngeal resonance: 0.14 to 0.99/0.36 to 0.78, hyponasality: /0.46 to 0.95, hypernasality: 0.8 to 0.99/0.81 to 0.92, anterior resonance: 0.34 to 0.97/0.42 to 0.95, posterior resonance: 0.287 to 0.948/0.79 to 0.94, strain: 0.38 to 0.9/0.12 to 0.74, breathiness: 0.78 to 0.91/0.24 to 0.64, roughness: 0.65 to 0.82/0.17 to 0.99, instability: 0.65 to 0.93/0.22 to 0.93, pitch: 0.75 to 0.98/0.92 to 0.96, loudness: 0.89 to 0.98/0.65 to 0.82, Overall severity: 0.39 to 0.98/0.62 to 0.98 and for spontaneous speech: Intelligibility: 0.918 to 0.992, articulation: 0.750 to 0.980, intonation: 0.860 to 0.937, speech rate: 0.872 to 0.939, speech-breathing coordination: 0.392 to 0.994, laryngeal resonance: 0.343 to 0.799, pharyngeal resonance: 0.736 to 1.0, hyponasality: 0.278 to 0.834, hypernasality: 0.869 to 0.999, anterior resonance: 0.521 to 1.0, posterior resonance: 0.867 to 0.997, strain: 0.855 to 0.991, breathiness: 0.241 to 0.972, roughness: 0.496 to 0.954, instability: 0.787 to 0.998, pitch: 0.867 to 0.995, loudness: 0.849 to 1.0, overall severity: 0.870 to 0.986	ICC for *sustained vowel/counting*: Laryngeal Resonance: 0.22/0.3, Pharyngeal Resonance: 0.28/0.61, Hyponasality: n.a./0.84 Hypernasality: 0.53/0.61, Anterior Resonance: 0.23/0.46, Posterior Resonance: 0.62/0.7, S: 0.26/0.69, B: 0.59/0.21, R: 0.24/0.37, I: 0.5/0.44, Pitch: 0.79/0.8, Loudness: 0.82/0.87, Overall Severity: 0.49/0.85 and for spontaneous speech: Intelligibility: 0.752, Articulation: 0.851, Intonation: 0.879, Speech Rate: 0.785, Speech-Breathing Coordination: 0.392, Laryngeal Resonance: 0.300, Pharyngeal Resonance: 0.464, Hyponasality: 0.790, Hypernasality: 0.621, Anterior Resonance: 0.286, Posterior Resonance: 0.747, S: 0.781, B: 0.185, R: 0.246, I: 0.626, Pitch: 0.758, Loudness: 0.402, Overall Severity: 0. 847	Coelho et al, 2017[102]
Unnamed#19	No	No	No	Pearson r (test/retest): Appropriate vibrato: 0.76, Resonance/ring: 0.75, Intonation accuracy: 0.74, Legato line: 0.74, Freedom throughout vocal range: 0.73, Efficient breath management: 0.73, Evenness of registration: 0.72, Intensity: 0.70, Dynamic range: 0.69, Diction: 0.69, Flexibility 0.65, Color/warmth: 0.65	Pearson r between raters: Intonation accuracy: 0.64, Freedom throughout vocal range: 0.55, Appropriate vibrato: 0.52, Evenness of registration: 0.52, Resonance/ring: 0.51, Flexibility: 0.50, Intensity: 0.48, Dynamic range: 0.48, Legato line: 0.47, Efficient breath management: 0.46, Color/warmth: 0.43, Diction: 0.34	Wapnick et al, 1997[103]
Unnamed#23	No	Yes	Yes, for the DME scale	Pearson r/percentage of agreement (repetition of 20% of the sample) for the *Direct Magnitude Estimation scale/ordinal scale*: Naturalness: 0.5 to 0.99/84%, Overall severity: 0.62 to 0.98/96%	Cronbach α for the *Direct Magnitude Estimation scale/ordinal scale*: Naturalness: 0.95/0.96, Overall severity: 0.87/0.97	Eadie et al, 2002[107]
Unnamed#24	Yes	No	No	Percentage of agreement: Onset of vibrato: 36 to 100%, vibrato rate: 45 to 100%, vibrato extent: 55 to 91%	No	Howes et al, 2004[108]
Unnamed#26	Yes	Yes	Yes, for the DME scale	Procedure not specified: For the Direct Magnitude Estimation scale/ordinal scale: 0.91 to 0.93/0.93 to 0.95	No	Lee et al, 2022[110]
Unnamed#25	Yes	Yes	No	ICC: 0.870	No	Nagle et al, 2014[109]
Unnamed#27	Yes	No	No	Pearson r (repetition of 25% of the sample) for the *visual analog scale/ordinal scale*: 0.98 to 0.99/0.92 to 0.95	No	Lee et al, 2018[111]

Tool	Training	Definition	Anchor	Intra-Judge Reliability	Inter-Judge Reliability	Reference
Unnamed#28	Yes	No	No	Pearson r (repetition of the sample) for the *visual analog scale/visual sort and rate*: Overall severity: 0.92/0.88, B: 0.94/0.9	ICC for the *visual analog scale/visual sort and rate*: Overall severity: 0.8/0.71, B:84/0.8	Kapsner-Smith et al, 2021[112]
Unnamed#29	No	No	No	Spearman ρ: Overall severity: 0.93, B: 0.83, R: 0.83, Brokenness: 0.95, Overall fluency: 0.97	Spearman ρ: Overall severity: 0.9, B: 0.93, R: 0.79, Brokenness: 0.91, Overall fluency: 0.94	Chen et al, 2019[113]
Unnamed#30	Yes, before the second session	Yes, before the second session	Yes, before the second session	Pearson r (repetition of 40% of the sample) of the *first/second session* for the sustained vowel: Overall severity: 0.9/0.94, R: 0.62/0.8, B: 0.7/0.79 - of the *first/second session* for the reading task: Overall severity: 0.92/0.96, R: 0.79/0.83, B: 0.7/0.8	ICC of the *first/second session* for the sustained vowel: Overall severity: 0.77/0.77, R: 0.41/0.71, B: 0.4/0.61 - of the *first/second session* for the reading task: Overall severity: 0.85/0.92, R: 0.6/0.8, B: 0.68/0.68	Eadie et al, 2006[114]
Unnamed#31	Yes	Yes	Yes	ICC for the *sustained vowels/reading task/conversational speech*: Overall Severity: 0.8/0.87/0.92, R: 0.74/0.82/0.87, B: 0.61/0.79/0.82, S: 0.77/0.83/0.87, Mean: 0.73/0.83/0.88	ICC for the *sustained vowels/reading task/conversational speech*: Overall Severity: 0.66/0.78/0.77, R: 0.6/0.78/0.65, B: 0.54/0.57/0.56, S: 0.54/0.74/0.61, Mean: 0.58/0.71/0.65	Law et al, 2012[115]
Unnamed#32	No	Yes	No	ICC for *naive listeners/SLPs: Overall severity*: 0.87/0.91, Vocal effort: 0.84/0.94	ICC *for naive listeners/SLPs*: Overall severity: 0.99/0.98, Vocal effort: 0.97/0.97	Eadie et al, 2010[116]
Unnamed#34	No	Yes	No	No	ICC = 0.73	Clapham et al, 2015[118]
Unnamed#36	No	Yes	Yes	No	ICC = Overall quality: 0.89, B: 0.61, R: 0.73, Brokenness: 0.73	Cannito et al, 2012[120]
Unnamed#37	No	Yes	No	No	Cronbach α for the *sustained vowel/reading task with normal intensity/reading task with high intensity*: Voice quality: 0.83/0.82/0.6, Vocal firmness: 0.8/0.83/0.82	Phadke et al, 2020[121]
Unnamed#33	No	No	No	Fleiss's κ = 0.470	Fleiss's κ = 0.340	Barsties et al, 2017[117]
Unnamed#39	No	Yes	Yes, for half of the evaluators	Pearson r (test/retest) for the group *with/without* perceptual anchor: Vocal roughness: 0.44 to 0.87/0.31 to 0.73, Listening effort: 0.27 to 0.86/0.09 to 0.81	ICC for the group *with/without* perceptual anchor: Vocal roughness: 0.69/0.57, Listening effort: 0.49/0.3	Farahani et al, 2023[123]
Unnamed#41	No	Yes	No	Pearson r: Voice assessment: 0.82 to 1, Severity: 0.70 to 0.91	Pearson r: Voice assessment: 0.75 to 1, Severity: 0.71 to 0.93	Selby et al, 2003[125]
Unnamed#40	Yes	No	Yes (4 groups)	Pearson r for the group *without anchor/with textual anchor/with perceptual anchor/with combined anchor*: 0.81/0.79/0.84/0.85	Pearson r for the group *without anchor/with textual anchor/with perceptual anchor/with combined anchor*: 0.7/0.74/0.75/0.78	Awan et al, 2009[124]
Unnamed#43	No	Yes	Yes, for the DME scale	Pearson r for the *Direct Magnitude Estimation scale/ordinal scale*: Voice pleasantness: 0.72/0.69, Voice severity: 0.74/0.84	ICC for the *Direct Magnitude Estimation scale/ordinal scale*: Voice pleasantness: 0.97/0.96, Voice severity: 0.97/0.98	Eadie et al, 2002[127]
Unnamed#45	No	No	No	ICC = 0.90	ICC = 0.70	Anand et al, 2012[129]
Unnamed#44	Yes	No	No	ICC = 0.83	ICC = 0.72	Maryn et al, 2019[128]

References

1. American Speech-Language-Hearing Association. (2004). Preferred practice patterns for the profession of speech-language pathology [Preferred Practice Patterns]. Accessed on: 1st September 2023. Available at: www.asha.org/policy/.
2. Choi SH. Speech-language pathologists' voice assessment and voice therapy practices: a survey for standard clinical guideline and evidence-based practice. *Commun Sci Disord.* 2013;18:473–485.
3. 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 Oto-rhino-laryngol.* 2001;258:77–82.
4. Kempster, G.B., Gerratt, B.R., Abbott, K.V., et al. (2009). Consensus auditory-perceptual evaluation of voice: development of a standardized clinical protocol.
5. 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;27:887–905.
6. Roy N, Barkmeier-Kraemer J, Eadie T, et al. Evidence-based clinical voice assessment: a systematic review. *Am J Speech Lang Pathol.* 2013;22:212–226.
7. Lechien JR, Geneid A, Bohlender JE, et al. Consensus for voice quality assessment in clinical practice: guidelines of the European Laryngological Society and Union of the European Phoniatricians. *Eur Arch Oto-rhino-laryngol.* 2023;280:5459–5473.

8. Crocco L, Madill CJ, McCabe P. Evidence-based frameworks for teaching and learning in classical singing training: a systematic review. *J Voice.* 2017;31:130–e7.
9. Ragan K. Defining evidence-based voice pedagogy: a new framework. *J Sing.* 2018;75:157.
10. Kreiman J, Gerratt BR, Kempster GB, et al. Perceptual evaluation of voice quality: review, tutorial and a framework for future research. *J Speech Hear Res.* 1993;36:21–40.
11. Behrman A. Common practices of voice therapists in the evaluation of patients. *J Voice.* 2005;l9:454–469.
12. Oates J. Auditory-perceptual evaluation of disordered voice quality: pros, cons and future directions. *Folia Phoniatr Logop.* 2009;61:49–56.
13. Harmegnies, B. (2015) Conférence collège Belgique.
14. Abercrombie D. *Elements of General Phonetics.* Edinburgh, UK: Edinburgh University Press; 1967.
15. Laver J. *The Phonetic Description of Voice Quality.* Cambridge, UK: Cambridge University Press; 1980.
16. Crystal D. *A Dictionary of Linguistics and Phonetics*. 3rd ed Oxford, UK: Blackwell Publishers; 1991.
17. Nolan F. *The Phonetic Bases of Speaker Recognition. Cambridge Studies in Speech Science and Communication.* Cambridge, UK: Cambridge University Press; 1983.
18. Esling JH, Moisik SR, Benner A, et al. *Voice Quality: The Laryngeal Articulator Model.* Cambridge, UK: Cambridge University Press; 2019.
19. Podesva RJ, Callie P. Voice quality and identity. *Annu Rev Appl Linguist.* 2015;35:173–194. https://doi.org/10.1017/S0267190514000270.
20. Allen MJ, Yen WM. *Introduction to Measurement Theory.* Pacific Grove, CA: Brooks/Cole, Wadsworth Inc; 1979.
21. Suhail IS, Kazi RA, Jagade M. Perceptual evaluation of tracheoesophageal speech: is it a reliable tool? *Indian J Cancer.* 2016;53:127–131.
22. Carding P, Carlson E, Epstein R, et al. Formal perceptual evaluation of voice quality in the United Kingdom. *Logop Phoniatr Vocol.* 2000;25:133–138.
23. Henrich N, Bézard P, Expert R, et al. Towards a common terminology to describe voice quality in western lyrical singing: contribution of a multidisciplinary research group. *J Interdiscip Music Stud.* 2008;2:71–93.
24. Tricco AC, Lillie E, Zarin W, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. *Ann Intern Med.* 2018;169:467–473.
25. Ahmadi N, Abbott KV, Rajati F, et al. Effects of laryngeal Manual Therapy on Primary Muscle Tension Dysphonia (MTD-1): implications for MTD-1 type. *J Voice.* 2022 Nov;38:1377–1385. https://doi.org/10.1016/j.jvoice.2022.04.002.
26. Bele IV. Reliability in perceptual analysis of voice quality. *J Voice.* 2005;19:555–573. https://doi.org/10.1016/j.jvoice.2004.08.008.
27. Bele IV. Dimensionality in voice quality. *J Voice.* 2007;21:257–272.
28. Brendel B, Ackermann H, Berg D, et al. Friedreich ataxia: dysarthria profile and clinical data. *Cerebellum.* 2013;12:475–484. https://doi.org/10.1007/s12311-012-0440-0.
29. Webb AL, Carding PN, Deary IJ, et al. The reliability of three perceptual evaluation scales for dysphonia. *Eur Arch Oto-rhino-laryngol Off J Eur Fed Oto-rhino-laryngol Soc (EUFOS): Affil German Soc Oto-rhino-laryngol Head Neck Surg.* 2004;261:429–434. https://doi.org/10.1007/s00405-003-0707-7.
30. Ehrlich B, Lin L, Jiang J. Concatenation of the moving window technique for auditory-perceptual analysis of voice quality. *Am J Speech Lang Pathol.* 2018;27:1426–1433. https://doi.org/10.1044/2018_AJSLP-17-0103.
31. Muss RC, Ward J, Huang L, et al. Correlation of vocal fold nodule size in children and perceptual assessment of voice quality. *Ann Otol Rhinol Laryngol.* 2010;119:651–655.
32. Joshi A, Baheti I, Angadi V. Cultural and linguistic adaptation of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) into Hindi. *J Speech Lang Hear Res.* 2020;63:3974–3981. https://doi.org/10.1044/2020_JSLHR-20-00348.
33. Helou LB, Solomon NP, Henry LR, et al. The role of listener experience on Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) ratings of postthyroidectomy voice. *Am J Speech Lang Pathol.* 2010;19:248–258. https://doi.org/10.1044/1058-0360(2010/09-0012).
34. Duvall A, Dion GR. Characterization of vocal fold pathology in military drill instructors. *J Voice Off J Voice Found.* 2021 Nov;37:907–912. https://doi.org/10.1016/j.jvoice.2021.05.013.
35. Nagle KF. Clinical use of the CAPE-V scales: agreement, reliability and notes on voice quality. *J Voice.* 2022 May;39:685–698. https://doi.org/10.1016/j.jvoice.2022.11.014.
36. Cohen W, Lloyd S, Wynne DM, et al. Perceptual evaluation of voice disorder in children who have had laryngotracheal reconstruction surgery and the relationship between clinician perceptual ratingof voice quality and parent proxy/child self-report of voice-related quality of life. *J Voice.* 2019;33:945.e27–945.e35. https://doi.org/10.1016/j.jvoice.2018.07.009.
37. Walden PR. Perceptual Voice Qualities Database (PVQD): database characteristics. *J Voice.* 2022;36:875.e15–875.e23. https://doi.org/10.1016/j.jvoice.2020.10.001.
38. Özcebe E, Aydinli FE, Tiğrak TK, et al. Reliability and validity of the Turkish Version of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V). *J Voice.* 2019;33:382.e1–382.e10. https://doi.org/10.1016/j.jvoice.2017.11.013.
39. Karnell MP, Melton SD, Childes JM, et al. Reliability of clinician-based (GRBAS and CAPE-V) and patient-based (V-RQOL and IPVI) documentation of voice disorders. *J Voice.* 2007;21:576–590. https://doi.org/10.1016/j.jvoice.2006.05.001.
40. Rocha PS, Bento N, Svärd H, et al. Voice assessment in patients with amyotrophic lateral sclerosis: an exploratory study on associations with bulbar and respiratory function. *Brain Sci.* 2024;14:1082.
41. Yamaguchi H, Shrivastav R, Andrews ML, et al. A comparison of voice quality ratings made by Japanese and American listeners using the GRBAS scale. *Folia Phoniatr Logop.* 2003;55:147–157.
42. Gurrado A, Pasculli A, Pezzolla A, et al. A method to repair the recurrent laryngeal nerve during thyroidectomy. *Can J Surg.* 2018;61:278–282. https://doi.org/10.1503/cjs.010317.
43. Sellars C, Stanton AE, McConnachie A, et al. Reliability of perceptions of voice quality: evidence from a problem asthma clinic population. *J Laryngol Otol.* 2009;123:755–763. https://doi.org/10.1017/S0022215109004605.
44. Moya-Mendez ME, Madden LL, Ruckart KW, et al. Auditory-perceptual voice and speech evaluation in ATP1A3 positive patients. *J Clin Neurosci Off J Neurosurg Soc Australas.* 2020;81:133–138. https://doi.org/10.1016/j.jocn.2020.09.007.
45. Midi I, Dogan M, Koseoglu M, et al. Voice abnormalities and their relation with motor dysfunction in Parkinson's disease. *Acta Neurol Scand.* 2008;117:26–34. https://doi.org/10.1111/j.1600-0404.2007.00965.x.
46. Gurbuzler L, Inanir A, Yelken K, et al. Voice disorder in patients with Fibromyalgia. *Auris Nasus Larynx.* 2013;40:554–557. https://doi.org/10.1016/j.anl.2013.04.002.
47. Balasubramanium RK, Karuppali S, Bajaj G, et al. Acoustic-perceptual correlates of voice in Indian Hindu purohits. *J Voice.* 2019;33:804.e1–804.e4. https://doi.org/10.1016/j.jvoice.2018.03.006.
48. Maciejewska B, Maciejewska-Szaniec Z, Małaczyńska B, et al. Acoustics features of voice in adolescent females with anorexia nervosa. *J Voice.* 2023. Sep;39:1410.e9–1410.e17. https://doi.org/10.1016/j.jvoice.2023.04.012.
49. Vaz Freitas S, Pestana PM, Almeida V, et al. Audio-perceptual evaluation of Portuguese voice disorders—an inter- and intrajudge reliability study. *J Voice.* 2014;28:210–215. https://doi.org/10.1016/j.jvoice.2013.08.001.
50. Hidaka S, Lee Y, Nakanishi M, et al. Automatic GRBAS scoring of pathological voices using deep learning and a small set of labeled voice data. *J Voice.* 2022. May;39:846.e1–846.e23. https://doi.org/10.1016/j.jvoice.2022.10.020.
51. Jannetts S, Lowit A. Cepstral analysis of hypokinetic and ataxic voices: correlations with perceptual and other acoustic measures. *J Voice.* 2014;28:673–680. https://doi.org/10.1016/j.jvoice.2014.01.013.

52. Fujimura S, Kojima T, Okanoue Y, et al. Classification of voice disorders using a one-dimensional convolutional neural network. *J Voice.* 2022;36:15–20. https://doi.org/10.1016/j.jvoice.2020.02.009.
53. Nagy A, Elshafei R, Mahmoud S. Correlating undiagnosed hearing impairment with hyperfunctional dysphonia. *J Voice.* 2020;34:616–621. https://doi.org/10.1016/j.jvoice.2019.02.002.
54. Fujiki RB, Thibeault SL. Examining relationships between GRBAS ratings and acoustic, aerodynamic and patient-reported voice measures in adults with voice disorders. *J Voice.* 2023;37:390–397. https://doi.org/10.1016/j.jvoice.2021.02.007.
55. Wuyts FL, De Bodt MS, Van de Heyning PH. Is the reliability of a visual analog scale higher than an ordinal scale? An experiment with the GRBAS scale for the perceptual evaluation of dysphonia. *J Voice.* 1999;13:508–517. https://doi.org/10.1016/S0892-1997(99)80006-X.
56. Pebbili GK, Kashyap R, Karike A, et al. Laryngeal aerodynamic analysis of glottal valving in children with down syndrome. *J Voice.* 2021;35:156.e15–156.e21. https://doi.org/10.1016/j.jvoice.2019.05.011.
57. Rzepakowska A, Sielska-Badurek E, Osuch-Wójcikiewicz E, et al. Multiparametric assessment of voice quality and quality of life in patients undergoing microlaryngeal surgery—correlation between subjective and objective methods. *J Voice.* 2018;32:257.e21–257.e30. https://doi.org/10.1016/j.jvoice.2017.04.016.
58. Murry T, Medrado R, Hogikyan ND, et al. The relationship between ratings of voice quality and quality of life measures. *J Voice.* 2004;18:183–192. https://doi.org/10.1016/j.jvoice.2003.11.003.
59. Benoy J.J., Jayakumar T. Effect of anchor voices and listener expertise on auditory-perceptual judgments of voice quality using the GRBAS scale. *J Voice*. 2024. Jan 9:S0892-1997(23)00397-1.
60. Deborah R., Samayan K. Cepstral analysis of voice in school-aged children. J Voice. 2024. Apr 26:S0892-1997(24)00090-0.
61. Narasimhan SV, Sahana P, Sahana K, et al. Adaptation and validation of the voice symptom scale into Kannada (VoiSS-K). *J Voice.* 2023 Nov;39:1701.e1–1701.e6.
62. Schindler A, Favero E, Nudo S, et al. Long-term voice and swallowing modifications after supracricoid laryngectomy: objective, subjective, and self-assessment data. *Am J Otolaryngol.* 2006;27:378–383. https://doi.org/10.1016/j.amjoto.2006.01.010.
63. Martens JW, Versnel H, Dejonckere PH. The effect of visible speech in the perceptual rating of pathological voices. *Arch Otolaryngol Head Neck Surg.* 2007;133:178–185.
64. Silva LFE, Gama ACC, Cardoso FEC, et al. Idiopathic Parkinson's disease: vocal and quality of life analysis. *Arq Neuro-Psiquiatr.* 2012;70:674–679. https://doi.org/10.1590/S0004-282X2012000900005.
65. Baudonck N, D'haeseleer E, Dhooge I, et al. Objective vocal quality in children using cochlear implants: a multiparameter approach. *J Voice.* 2011;25:683–691. https://doi.org/10.1016/j.jvoice.2010.05.005.
66. Moodley D-T, Swanepoel C, van Lierde K, et al. Vocal characteristics of school-aged children with and without attention deficit hyperactivity disorder. *J Voice.* 2019;33:945.e37–945.e45. https://doi.org/10.1016/j.jvoice.2018.06.008.
67. Moerman MBJ, Martens JP, Van der Borgt MJ, et al. Perceptual evaluation of substitution voices: development and evaluation of the (I)INFVo rating scale. *Eur Arch Oto-rhino-laryngol Off J Eur Fed Oto-rhino-laryngol Soc (EUFOS) Affil German Soc Oto-rhino-laryngol Head Neck Surg.* 2006;263:183–187. https://doi.org/10.1007/s00405-005-0960-z.
68. Moerman M, Martens J-P, Crevier-Buchman L, et al. The INFVo perceptual rating scale for substitution voicing: development and reliability. *Eur Arch Oto-rhino-laryngol Off J Eur Fed Oto-rhino-laryngol Soc (EUFOS) Affil German Soc Oto-rhino-laryngol Head Neck Surg.* 2006;263:435–439. https://doi.org/10.1007/s00405-005-1033-z.
69. Pribuisis K, Pribuisiene R, Liutkevicius V, et al. Investigation of relationship between auditory-perceptual methods and self-reported speech handicap index in the assessment of substitution voicing. *J Voice.* 2022;36:435.e23–435.e31. https://doi.org/10.1016/j.jvoice.2020.06.011.
70. Mesolella M, Allosso S, D'aniello R, et al. Subjective perception and psychoacoustic aspects of the laryngectomee voice: the impact on quality of life. *J Pers Med.* 2023;13:570. https://doi.org/10.3390/jpm13030570.
71. de Almeida SC, Mendes AP, Kempster GB. The Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) psychometric characteristics: II European Portuguese Version (II EP CAPE-V). *J Voice.* 2019;33:582.e5–582.e13. https://doi.org/10.1016/j.jvoice.2018.02.013.
72. Tamplin J, Baker FA, Buttifant M, et al. The effect of singing training on voice quality for people with quadriplegia. *J Voice.* 2014;28:128.e19–128.e26. https://doi.org/10.1016/j.jvoice.2013.08.017.
73. Machado de Machado FC, Lessa MM, Cielo CA, et al. Phonotherapeutic intervention in patients with mucosal leishmaniasis sequelae. *J Voice.* 2020;34:720–731. https://doi.org/10.1016/j.jvoice.2018.12.015.
74. Pützer M, Wokurek W, Moringlane JR. Evaluation of phonatory behavior and voice quality in patients with multiple sclerosis treated with deep brain stimulation. *J Voice.* 2017;31:483–489. https://doi.org/10.1016/j.jvoice.2016.10.022.
75. Schönweiler R, Hess M, Wübbelt P, et al. Novel approach to acoustical voice analysis using artificial neural networks. *J Assoc Res Otolaryngol JARO.* 2000;1:270–282. https://doi.org/10.1007/s101620010020.
76. Hurren A, Miller N, Carding P. Perceptual assessment of tracheoesophageal voice quality with the SToPS: the development of a reliable and valid tool. *J Voice.* 2019;33:465–472. https://doi.org/10.1016/j.jvoice.2017.12.006.
77. Staffan M, Lindestad PÅ, Mats H, et al. Voice quality in adults treated for unilateral cleft lip and palate. *Cleft Palate Craniofac J.* 2018;55:1103–1114. https://doi.org/10.1177/1055665618764521.
78. Fernandez-Fresard G, Acevedo K. Voice performance chart: a pedagogical tool to enhance vocal expressive ability in acting students. *J Voice.* 2021: May;38(3):797.e1–797.e10. https://doi.org/10.1016/j.jvoice.2021.10.020.
79. Saleh M, Ayman D. Validation of GSLBI (Grade, Strained, Leaky, Breathy, Irregular): an Egyptian instrument for auditory perceptual assessment of voice. *Egypt J Otolaryngol.* 2025;41:59.
80. Langeveld TPM, Drost HA, Frijns JHM, et al. Perceptual characteristics of adductor spasmodic dysphonia. *Ann Otol Rhinol Laryngol.* 2000;109:741–748.
81. Ikuma T, McWhorter AJ, Oral E, et al. Formant-aware spectral analysis of sustained vowels of pathological breathy voice. *J Voice.* 2023: Sep;39(5):1413.e1–1413.e16. https://doi.org/10.1016/j.jvoice.2023.05.002.
82. Weinrich B, Brehm SB, LeBorgne W, et al. Perceptual measures of boychoir voices during the phases of pubertal voice mutation. *J Voice.* 2022;36:142.e1–142.e8. https://doi.org/10.1016/j.jvoice.2020.04.002.
83. Smith ME, Marsh JH, Cotton RT, et al. Voice problems after pediatric laryngotracheal reconstruction: videolaryngostroboscopic, acoustic, and perceptual assessment. *Int J Pediatr Otorhinolaryngol.* 1993;25:173–181. https://doi.org/10.1016/0165-5876(93)90051-4.
84. Södersten M, Hertegård S, Hammarberg B. Glottal closure, transglottal airflow, and voice quality in healthy middle-aged women. *J Voice.* 1995;9:182–197. https://doi.org/10.1016/S0892-1997(05)80252-8.
85. Wong DW-M, Chan RW, Wu C-H. Effect of training with anchors on auditory- perceptual evaluation of dysphonia in speech-language pathology students. *J Speech Lang Hear Res.* 2021;64:1136–1156. https://doi.org/10.1044/2020_JSLHR-20-00214.
86. Misono S, Merati AL, Eadie TL. Developing auditory-perceptual judgment reliability in otolaryngology residents. *J Voice.* 2012;26:358–364. https://doi.org/10.1016/j.jvoice.2011.07.006.
87. Eadie T, Sroka A, Wright DR, et al. Does knowledge of medical diagnosis bias auditory-perceptual judgments of dysphonia? *J Voice.* 2011;25:420–429. https://doi.org/10.1016/j.jvoice.2009.12.009.
88. Sauder C, Eadie T. Does the accuracy of medical diagnoses affect novice listeners' auditory-perceptual judgments of dysphonia severity? *J Voice.* 2020;34:197–207. https://doi.org/10.1016/j.jvoice.2018.08.001.
89. Eadie TL, Van Boven L, Stubbs K, et al. The effect of musical background on judgments of dysphonia. *J Voice.* 2010;24:93–101. https://doi.org/10.1016/j.jvoice.2008.04.008.

90. Nguyen DD, Madill C. Auditory-perceptual parameters as predictors of voice acoustic measures. *J Voice.* 2023;39:1225–1235. https://doi.org/10.1016/j.jvoice.2023.02.030.
91. Chen SH, Hsiao T-Y, Hsiao L-C, et al. Outcome of resonant voice therapy for female teachers with voice disorders: perceptual, physiological, acoustic, aerodynamic, and functional measurements. *J Voice.* 2007;21:415–425. https://doi.org/10.1016/j.jvoice.2006.02.001.
92. Madill CJ, Sheard C, Heard R. Are instructions to manipulate specific parameters of laryngeal function associated with auditory-perceptual ratings of voice quality in nondisordered speakers? *J Voice.* 2017;31:504.e21–504.e33. https://doi.org/10.1016/j.jvoice.2016.10.008.
93. Kallvik E, Lindström E, Holmqvist S, et al. Prevalence of hoarseness in school-aged children. *J Voice.* 2015;29:260.e1–260.e19. https://doi.org/10.1016/j.jvoice.2013.08.019.
94. van As CJ, Koopmans-van Beinum FJ, Pols LCW, et al. Perceptual evaluation of tracheoesophageal speech by naive and experienced judges through the use of semantic differential scales. *J Speech Lang Hear Res.* 2003;46:947–959.
95. Zhang F, Nguyen DD, Madill C, et al. Do the nonlinear dynamic acoustic measurements, nonlinear energy difference ratio and spectrum convergence ratio, correlate with perceptual evaluation of esophageal voice speakers? *J Voice.* 2022 Nov;38:1278–1287. https://doi.org/10.1016/j.jvoice.2022.06.004.
96. Contreras-Ruston F, Guzman M, Castillo-Allendes A, et al. Auditory-perceptual assessment of healthy and disordered voices using the voice deviation scale. *J Voice Off J Voice Found.* 2021. May;38:654–659. https://doi.org/10.1016/j.jvoice.2021.10.017.
97. Gelfer M. A multidimensional scaling study of voice quality in females. *Phonetica.* 1993;50:15–27.
98. Gelfer MP. Perceptual attributes of voice: development and use of rating scales. *J Voice.* 1988;2:320–326. https://doi.org/10.1016/S0892-1997(88)80024-9.
99. Andrews ML, Schmidt CP. Gender presentation: perceptual and acoustical analyses of voice. *J Voice.* 1997;11:307–313.
100. Haderlein T, Döllinger M, Matousek V, et al. Objective voice and speech analysis of persons with chronic hoarseness by prosodic analysis of speech samples. *Logop Phoniatr Vocol.* 2016;41:106.
101. Edgar JD, Sapienza CM, Bidus K, et al. Acoustic measures of symptoms in abductor spasmodic dysphonia. *J Voice.* 2001;15:362–372. https://doi.org/10.1016/S0892-1997(01)00038-8.
102. Coelho AC, Brasolotto AG, Fernandes ACN, et al. Auditory-perceptual evaluation of voice quality of cochlear-implanted and normal-hearing individuals: a reliability study. *J Voice.* 2017;31:774.e1–774.e8. https://doi.org/10.1016/j.jvoice.2017.02.012.
103. Wapnick J, Ekholm E. Expert consensus in solo voice performance evaluation. *J Voice.* 1997;11:429–436.
104. Raveendran R, Yeshoda K. Effects of resonant voice therapy on perceptual and acoustic source and tract parameters – a preliminary study on Indian carnatic classical singers. *J Voice.* 2022. Mar;39(2):560.e1–560.e9. https://doi.org/10.1016/j.jvoice.2022.09.023.
105. LeBorgne W, Lee L, Stemple JC, et al. Perceptual findings on the broadway belt voice. *J Voice.* 2010;24:678–689. https://doi.org/10.1016/j.jvoice.2009.02.004.
106. Ross J-A, Noordzji JP, Woo P. Voice disorders in patients with suspected laryngo-pharyngeal reflux disease. *J Voice.* 1998;12:84–88. https://doi.org/10.1016/S0892-1997(98)80078-7.
107. Eadie TL, Doyle PC. Direct magnitude estimation and interval scaling of naturalness and severity in tracheoesophageal (TE) speakers. *J Speech Lang Hear Res.* 2002;45:1088–1096.
108. Howes P, Callaghan J, Davis P, et al. The relationship between measured vibrato characteristics and perception in Western operatic singing. *J Voice.* 2004;18:216–230. https://doi.org/10.1016/j.jvoice.2003.09.003.
109. Nagle K, Helou LB, Solomon NP, et al. Does the presence or location of graphic markers affect untrained listeners' ratings of severity of dysphonia? *J Voice.* 2014;28:469–475.
110. Lee Y, Park H, Lim D, et al. Usefulness of Direct Magnitude Estimation (DME) in auditory perceptual assessments measuring dysphonia severity. *J Voice.* 2022 Aug 22. https://doi.org/10.1016/j.jvoice.2022.09.001.S0892-1997(24)00225-X.
111. Lee YW, Kim GH, Bae IH, et al. The cut-off analysis using visual analogue scale and cepstral assessments on severity of voice disorder. *Logop Phoniatr Vocol.* 2018;43:175–180. https://doi.org/10.1080/14015439.2018.1461925.
112. Kapsner-Smith MR, Opuszynski A, Stepp CE, et al. The effect of visual sort and rate versus visual analog scales on the reliability of judgments of dysphonia. *J Speech Lang Hear Res.* 2021;64:1571–1580. https://doi.org/10.1044/2021_JSLHR-20-00623.
113. Chen Z, Li J, Ren Q, et al. Acoustic and perceptual analyses of adductor spasmodic dysphonia in Mandarin-speaking Chinese. *J Voice.* 2019;33:333–339. https://doi.org/10.1016/j.jvoice.2017.12.007.
114. Eadie TL, Baylor CR. The effect of perceptual training on inexperienced listeners judgements of dysphonic voice. *J Voice.* 2006;20:527–544.
115. Law T, Kim JH, Lee KY, et al. Comparison of rater's reliability on perceptual evaluation of different types of voice sample. *J Voice.* 2012;26:666.e13–666.e21. https://doi.org/10.1016/j.jvoice.2011.08.003.
116. Eadie TL, Kapsner M, Rosenzweig J, et al. The role of experience on judgments of dysphonia. *J Voice.* 2010;24:564–573. https://doi.org/10.1016/j.jvoice.2008.12.005.
117. Barsties v. Latoszek B, Maryn Y, Gerrits E, et al. The Acoustic Breathiness Index (ABI): a multivariate acoustic model for breathiness. *J Voice.* 2017;31:511.e11–511.e27. https://doi.org/10.1016/j.jvoice.2016.11.017.
118. Clapham RP, van As-Brooks CJ, van Son RJJH, et al. The relationship between acoustic signal typing and perceptual evaluation of tracheoesophageal voice quality for sustained vowels. *J Voice.* 2015;29:517.e23–517.e29. https://doi.org/10.1016/j.jvoice.2014.10.002.
119. Ghio A, Révis J, Merienne S, et al. Top-down mechanisms in dysphonia perception: the need for blind tests. *J Voice.* 2013;27:481–485. https://doi.org/10.1016/j.jvoice.2013.03.015.
120. Cannito MP, Doiuchi M, Murry T, et al. Perceptual structure of adductor spasmodic dysphonia and its acoustic correlates. *J Voice.* 2012;26:818.e5–818.e13. https://doi.org/10.1016/j.jvoice.2012.05.005.
121. Phadke KV, Laukkanen A-M, Ilomäki I, et al. Cepstral and perceptual investigations in female teachers with functionally healthy voice. *J Voice.* 2020;34:485.e33–485.e43. https://doi.org/10.1016/j.jvoice.2018.09.010.
122. Pinczower R, Oates J. Vocal projection in actors: the long-term average spectral features that distinguish comfortable acting voice from voicing with maximal projection in male actors. *J Voice.* 2005;19:440–453.
123. Farahani M, Parsa V, Doyle PC. Auditory-perceptual and pupillometric evaluation of vocal roughness and listening effort in tracheoesophageal speech. *J Voice.* 2023. Nov;39(6):1702.e11–1702.e26. https://doi.org/10.1016/j.jvoice.2023.04.021.
124. Awan SN, Lawson LL. The effect of anchor modality on the reliability of vocal severity ratings. *J Voice.* 2009;23:341–352. https://doi.org/10.1016/j.jvoice.2007.10.006.
125. Selby JC, Gilbert HR, Lerman JW. Perceptual and acoustic evaluation of individuals with laryngopharyngeal reflux pre- and post-treatment. *J Voice.* 2003;17:557–570. https://doi.org/10.1067/S0892-1997(03)00017-1.
126. De Bodt MS, Hernández-Díaz HME, Van De Heyning PH. Intelligibility as a linear combination of dimensions in dysarthric speech. *J Commun Disord.* 2002;35:283–292. https://doi.org/10.1016/s0021-9924(02)00065-5.
127. Eadie TL, Doyle PC. Direct magnitude estimation and interval scaling of pleasantness and severity in dysphonic and normal speakers. *J Acoust Soc Am.* 2002;112:3014–3021. https://doi.org/10.1121/1.1518983.

128. Maryn Y, Leblans M, Zarowski A, et al. Objective acoustic quantification of perceived voice tremor severity. *J Speech Lang Hear Res.* 2019;62:3689–3705. https://doi.org/10.1044/2019_JSLHR-S-19-0024.
129. Anand S, Shrivastav R, Wingate JM, et al. An acoustic-perceptual study of vocal tremor. *J Voice.* 2012;26:811.e1–811.e7. https://doi.org/10.1016/j.jvoice.2012.02.007.
130. Paz KEDS, Paiva MAAD, Lima DO, et al. Training for perceptive-auditory voice analysis: scope review. *Audiol Commun Res.* 2023;28:e2768.
131. Wang Z., Wei Y., Yu W.S., et al. Factors influencing the reliability of auditory-perceptual evaluation of voice: a scoping review. *J Voice*. 2025. Jun 14:S0892-1997(25)00206-1.
132. Yousef AM, Cantor-Cutiva LC, Hunter EJ. Mapping 74 years in acoustic analysis of voice disorders: a bibliometric review and future research directions. *J Commun Disord.* 2025;117:106555.
133. Pommée T., Shanks M., Morsomme D., et al. Validation of the European French Version of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-Vf). *J Voice*. 2024. Nov 6:S0892-1997(24)00364-3.
134. Kent RD. Hearing and believing: some limits to the auditory-perceptual assessment of speech and voice disorders. *Am J Speech Lang Pathol.* 1996;5:7–23.
135. Eadie TL, Baylor CR. The effect of perceptual training on inexperienced listeners' judgments of dysphonic voice. *J Voice.* 2006;20:527–544. https://doi.org/10.1016/j.jvoice.2005.08.007.
136. Chan KM, Yiu EM. The effect of anchors and training on the reliability of perceptual voice evaluation. *J Speech Lang Hear Res.* 2002;45:111–126.
137. Fissel Brannick S, Wolford GW, Wolford LL, et al. What is clinical evidence in speech-language pathology? A scoping review. *Am J Speech Lang Pathol.* 2022;31:2943–2958.
138. McAlister S, Yanushevskaya I. Voice assessment practices of speech and language therapists in Ireland. *Clin Linguist Phon.* 2019;34:29–53.
139. Lee Y, Kim G, Kwon S. The usefulness of auditory perceptual assessment and acoustic analysis for classifying the voice severity. *J Voice.* 2020;34:884–893.
140. Hillman RE, Montgomery WW, Zeitels SM. Current diagnostics and office practice: appropriate use of objective measures of vocal function in the multidisciplinary management of voice disorders. *Curr Opin Otolaryngol Head Neck Surg.* 1997;5:172–175.
141. Maryn Y, Roy N, De Bodt M, et al. Acoustic measurement of overall voice quality: a meta-analysis. *J Acoust Soc Am.* 2009;126:2619–2634.
142. Cohen W, Wynne DM. Paediatric voice disorder: who to refer and how to assess? A summary of recent literature. *Curr Opin Otolaryngol Head Neck Surg.* 2024;32.156-16.
143. Rickert SM, O'Cathain E. Pediatric voice. *Pediatr Clin N Am.* 2022;69:329–347.
144. Cohen W, Wynne DM, Kubba H, et al. Development of a minimum protocol for assessment in the paediatric voice clinic. Part 1: evaluating vocal function. *Logop Phoniatr Vocol.* 2012;37:33–38.
145. Fujiki RB, Venkatraman A, Murray ESH. The pediatric vocal mechanism: structure and function. *J Voice.* 2025. https://doi.org/10.1016/j.jvoice.2025.03.025.Published online April 4.
146. Connor NP, Cohen SB, Theis SM, et al. Attitudes of children with dysphonia. *J Voice.* 2008;22:197–209. https://doi.org/10.1016/j.jvoice.2006.09.005.
147. Krasnodębska P, Piluch P, Szkiełkowska A. Use of the Polish Adaptation of the Pediatric Voice Handicap Index in predicting severe voice disorders in children. *J Voice.* 2025 Aug 7.S0892-1997(25)00250-4.
148. Munjal S, Alam MN, Panda NK. Subjective evaluation of voice characteristics of school aged children in a basket ball team. *Indian J Otolaryngol Head Neck Surg.* 2019;71(suppl 1):465–468.
149. Anand S, Awan SN, Park Y, et al. Voice quality in pediatric dysphonia: assessing expert rater reliability across three perceptual methods. *J Voice.* 2025 Aug 14.S0892-1997(25)00297-8.
150. Fusaroli R, Lambrechts A, Bang D, et al. Is voice a marker for Autism spectrum disorder? A systematic review and meta-analysis. *Autism Res.* 2017;10:384–407.
151. He R, Chapin K, Al-Tamimi J, et al. Automated classification of cognitive decline and probable Alzheimer's dementia across multiple speech and language domains. *Am J Speech Lang Pathol.* 2023;32:2075–2086.